IEEE
Advancing Technology for Humanity

IEEE ComSoc
IEEE Communications Society

The 26th International Conference on Advanced Communications Technology

“ Toward secure and comfortable life in emergingAI and data-driven era! ”

http://www.icact.org

icact 2024

Phoenix Park, Pyeongchang
Korea (south)
Feb. 4 ~ 7, 2024

PROCEEDINGS & JOURNAL

IEEE Catalog Number: CFP24561-ART
ISBN: 979-11-88428-12-0
ISSN: 1738-9445

Organizers

GIRI
Global IT Research Institute

Sponsors

IEEE ComSoc IEEE Communications Society

NIA NATIONAL INFORMATION SOCIETY AGENCY

ETRI

GWCVB Gangwon Convention & Visitors Bureau

ITI Information Technology Institute of Vietnam National University

KICS

IEEK ComSoc

Korean Institute of Information Scientists and Engineers

OSIA Open Standards and Internet Association

Korea Institute of Information Security & Cryptology

The 26th
International Conference on Advanced Communications Technology

"Toward secure and comfortable life in emerging AI and data-driven era!!"

Phoenix Pyeongchang

Korea (south)

Feb. 4 ~ 7, 2024

Proceedings & Journal

IEEE Catalog: CFP24561-ART
ISBN 979-11 -88428-12-0
ISSN: 1738-9445

Organizer

Global IT Research Institute (GIRI)

Sponsors

IEEE Communications Society (IEEE ComSoc)

National Information Society Agency (NIA)

Electronics and Telecommunications Research Institute (ETRI)

Gangwon Convention & Visitors Bureau (GWCVB)

Information Technology Institute-Vietnam National University (ITI-VNU)

Korean Institute of Communications Sciences (KICS)

IEEK Communication Society (IEEK)

Korean Institute of Communication Scientist and Engineers (KIISE)

Open Standards and Internet Association (OSIA)

Korea Institute of Information Security & Cryptology (KIISC)

Copyright notice

Articles in this publication may be cited in other publications. In order to facilitate access to the original publication source, the following form for the citation is suggested:

Name of Author(s), "Title of Paper," in the 26th International Conference on Advanced Communications Technology, Technical Proceedings, 2024, page numbers.

IEEE Catalog Number: CFP24561-ART

ISSN 1738-9445

ISBN 979-11-88428-12-0

Copyright ©2024; Global IT Research Institute

Copy right and reprint permissions: Individual readers of these technical proceedings acting for them are permitted to make fair use of the material in it, such as to copy an article for use in teaching and research, without payment of fee, providing that such copies are not sold. For other copying, reprint of republication permissions, write to IEEE, 445 Hoes Lane, Piscataway, NJ 08854, USA.

Permission is granted to quote excerpts from articles in these technical proceedings in scientific works with the customary acknowledgement of the source, including the author's name and the name of the technical proceedings, page, year, and name of the Institute. Reproduction of figures and tables is likewise permitted in other articles and books provided that the same information is printed with them and notification is given to the Global IT Research Institute. Is addition, Global IT Research Institute may require that permission also be obtained from one of the authors.

Additional copies can be ordered from:

Conference Publication Operations

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERERS, INC.

445 HOES LANE, PISCATAWAY, NJ 08854-4141, USA

+1 732 981 0060 (phone)

+1 732 981 1769 (fax)

confpubs@ieee.org (e-mail)

Printed in Korea

Message from the ICACT2024 General Chair

Prof. Andrey Koucheryavy, PhD
Dean of department "Communication Networks and Data Transmission, Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Russia, and former Chairman of Study Group 11 ITU-T. He worked in the Telecommunication Research Institute (LONIIS), from 1974 up to October 2003 (from 1986 up to 2003 as the First Deputy Director).

On behalf of all the ICACT2024 international conference committees, I would like to welcome all of you to ICACT 2024. It is a great honor for me to host the 26th IEEE International Conference on Advanced Communications Technology (ICACT) at Phoenix PyeongChang, Gangwon-do, Korea.
The ICACT is an annual international conference in the communications field, where all experts from home and abroad come together to present their work, and share new ideas and visions for achieving the future telecommunications age.
The ICACT has held annually since 1999, and also the last conference, ICACT2023 , was successfully completed with selected 81 papers among 276 submissions by 3 peer reviewers , presented in full virtual sessions , because of the COVID -19 pandemic . There were 66 foreign and 15 domestic paper presenters from 23 countries.
All the accepted and presented papers will be published in the conference proceedings and submitted to IEEE Xplore as well as other Abstracting and Indexing (A&I) databases , such as SCOPUS , EI Compendex , INSPEC , and Conference Proceedings Citation Index (CPCI), etc. Recent 3 years Impact factor H index is 21 score by the SCImago Lab, using Scopus Data Source, which means the rank of this international conference is the 63rd among 754 international conferences.

.

Message from the Organization and Operation Committee Chair

With technically co-sponsored by IEEE ComSoc (Communications Society); IEEE ComSoc CISTC (Communications & Information Security Technical Community), and IEEE ComSoc ONTC (Optical Networking Technical Community), the ICACT (International Conference on Advanced Communications Technology) Conference has been providing an open forum for scholars, researchers, and engineers to the extensive exchange of information on newly emerging technologies, standards, services, and applications in the area of the advanced communications technology. The conference official language is English. All the presented papers have been published in the Conference Proceedings, and posted on the ICACT Website and IEEE Xplore Digital Library since 2004. The honorable ICACT Out-Standing Paper Award list has been posted on the IEEE Xplore Digital Library also, and all the Out-Standing papers are subjected to the invited paper of the "ICACT Transactions on the Advanced Communications Technology" Journal issued by GiRI(Global IT Research Institute) accredited by Korea Government(Reg. No. 220-82-07506) hosting the ICACT Conference. All the accepted and presented papers will be published in the conference proceedings and submitted to IEEE Xplore as well as other Abstracting and Indexing (A&I) databases such as SCOPUS, INSPEC, Engineering Index (EI), Conference Proceedings Citation Index (CPCI), etc. IEEE Conference Status

Prof. Thomas Byeong-Nam Yoon, PhD.
Organizer, International Conference on Advanced Communications Technology (ICACT)
Editor-in-Chief, ICACT Transactions on the Advanced Communications Technology (TACT) Journal
President, Global IT Research Institute (GiRI)

Objectives

International Conference on Advanced Communications Technology (ICACT) provides an open forum for researchers, engineers, policy makers, network planners, and service providers in telecommunications. Extensive exchange of information will be provided on newly emerging systems, standards, services, and variety of applications on the area of telecommunications.

ICACT2024 Committee

International Steering Committee

Prof. Andrey Koucheryavy, PhD, (ISC Chair) ITU-T SG11 Chair (2017~)
Central Science Research Telecommunication Institute, Russia

Prof. Yi Qian, PhD, IEEE ComSoc CISTC (2014-15 Chair), University of Nebraska-Lincoln, USA

Dr. Peter Mueller, IEEE ComSoc CISTC (2012-13 Chair), Zurich IBM, Switzland

Prof. Hsiao-Hwa chen, PhD, IEEE ComSoc CISTC (2010-11 Chair),
National Cheng Kung University, Taiwan

Prof. Min-Ho Kang, PhD, ICACT(2006-7, General Chair), KAIST University, Korea

Prof. Mun-Kee Choi, PhD, ICACT(2009, General Chair), KAIST University, Korea

Prof. Chu-Hwan YIM, PhD, ICACT(1999, General Chair), Korea University, Korea

Prof. Byrav Ramamurthy, PhD, IEEE ComSoc ON-TC(2010-11 Chair),
University of Nebraska-Lincoln, USA

Prof. JUN-KYUN CHOI, PhD, IEEE ComSoc ON-TC, KAIST, Korea

Prof. HYEONG-HO LEE, IEEE ComSoc ON-TC,
Seoul National University of Science and Technology, Korea

Prof. Byeong-Nam Yoon, PhD, IEEE ComSoc CISTC (2010~ Active Member),
Global IT Research Institute, Korea

Prof. DAE-YOUNG KIM, PhD, IEEE ComSoc ON-TC, CNU Univ., Korea

Dr. Myung-Won Song, IEEE ComSoc ON-TC, National Information Society Agency, Korea

Prof. WHAN-WOO KIM, PhD, IEEE ComSoc ON-TC, CNU Univ., Korea

Prof. Do Nang Toan, PhD, ITI Vietnam National University, Vietnam

Prof. Jianping WU, PhD, Computer Science, Tsinghua University, China

Dr. Ahmet Kaplan, VP&CTO TurkSat, Turkey

Prof. Lakshmi Prasad Saikia, PhD., GIMT-Guwahati (ASTU), India

Mr. Mohammed Qadeer, Aligarh Muslim University, India

Prof. Biswanath Mukherjee, PhD, University of California, Davis, USA

Prof. Moo Wan Kim, PhD., Tokyo University of Information Sciences, Japan

Dr. Hong Son Nguyen, PTIT, Vietnam

Dr. Nicolai Kuntze, Fraunhofer Institute for Secure Information Technology, Germany

Prof. Nguyen Kim Lan, PhD, VNPT, Vietnam

Dr. Pasi Ojala, University of Oulu, Finland

Prof. Sugata Sanyal, PhD, Tata Consultancy Services, Mumbai,India, India

Prof. David Hutchison, Lancaster University, UK

Dr. Yuji INOUE, TTC, Japan

Dr. Chen-Shie Ho, Oriental Institute of Technology, Taiwan

Dr. Yukinobu Fukushimam, Okayama University, Japan

Prof. Hiroshi ESAKI, PhD, University of Tokyo, Japan

Prof. Bin Sun, PhD, Beijing University of Posts and Telecommunications, China

Prof. Harekrishna Misra, Institute of Rural Management Anand, India

Prof. Khairuddin Ab. Hamid, PhD, Universiti Malaysia Sarawak, Malaysia

Organization & Operation Committee

Prof. Byeong-Nam Yoon, PhD, (OOC Chair) Global IT Research Institute, Korea

Prof. Minjae Park, PhD, Computer Software Dept. Daelim University, Korea

Dr. Chunming Zhao, Fujitsu Microelectronics America, USA

Pro. Young-Il Kwon, Hoseo University, Korea

Dr. Cong Hung Tran, PTIT, Vietnam

Prof. Mounir Hamdi, PhD, Hong Kong Univ. of S & T, Hong Kong

Prof. Hsu-Chen Cheng, PhD, Chinese Culture University, Taiwan

Prof. Ho-Kyung Lee, PhD, IEEK ComSoc(2011 Chair), Hongik University, Korea

Prof. Heung Youl Youm, PhD, KIISC(2011 Chair), Soon Chun Hyang University, Korea

Prof. Jin-Pyo Hong, PhD, KIISE(2010 Chair), HUFS (University), Korea

Prof. Jose Piquer, PhD, DCC, University of Chile, Chile

Prof. Hyunkook Kahang, PhD, IEEE Deajeon Section, Korea

Prof. Kwang-Hoon Kim, PhD, Kyonggi University, Korea

Dr. Kwang-Teak Ryu, PhD, National Information Society Agency, Korea

Dr. Chin-Feng Lin, National Taiwan Ocean University, Taiwan

Prof. Jiann-Liang Chen, PhD, NTUST, Taiwan

Prof. Man-Seop Lee, PhD, KAIST (Univ.), Korea

Prof. Min-Xiou Chen, PhD, National Dong Hwa University, Taiwan

Dr. Mohd Nazrul Hanif Nordin, Telekom Research & Development Sdn Bhd, Malaysia

Dr. Razi Arshad, NUST, Pakistan

Prof. Altair Santin, PhD, PUCPR, Brazil

Prof. Selim YAZICI, PhD, Faculty of Political Sciences, Istanbul University, Turkey

Prof. Seong-Ho Jeong, PhD, HUFS (University), Korea

Dr. Sanghong Lee, KT, Korea

Prof. Sang-Jo Yoo, PhD, KICS(2011 Chair), Inha University, Korea

Prof. Sergio F. Ochoa, PhD, DCC, University of Chile, Chile

Dr. Sang-Chul Shin, NIPA, Korea

Dr. Seung-Won Shon, OSIA(2011 Chair), ETRI, Korea

Prof. Vu Truong Thanh, Waseda University, Japan

Prof. Hua Wang, PhD, Shandong University, China

Dr. Xu Huang, University of Canberra, Australia

Prof. Chi-Jung Hwang, PhD, CNU(University), Korea

Prof. Yoonjoon Lee, PhD, KIISE(2011 Chair), Korea

Dr. Young-Ro Lee, National Standards Coordinator Office, Korea

Prof. Charles Kim, PhD, Inje University, Korea

Technical Program Committee

Prof. Kwang-Hoon Kim, PhD, (TPC Chair) Kyonggi University, Korea

Prof. Hyeong-Ho Lee, (TPC Co-Chair: Program)
Seoul National University of Science and Technology, Korea

Prof. Andrey Koucheryavy, PhD, (TPC Co-Chair: Standard)
State University of Telecommunications, Russia

Prof. Hee-Chang Chung, PhD, (TPC Co-Chair: Standard) Dongeui University, Korea

Dr. Myung-Won Song, (TPC Co-Chair: Review) National Information Society Agency, Korea

Prof. Sun-Moo Kang, PhD, (TPC Co-Chair: Journal) Kyunghee University, Korea

Prof. Yeong Il Kwon, (TPC Co-Chair: Journal) Hoseo University, Korea

Prof. Hee-Cheol Kim, PhD, (TPC Co-Chair, Tutorial) Inje University, Korea

Prof. Do Nang Toan, PhD, (TPC Co-Chair, Workshop)ITI Vietnam National University, Vietnam

Prof. Minjae Park, PhD, (TPC Co-Chair, Workshop) Computer Software Dept.
Daelim University, Korea

Dr. Chua Kee Chaing, National University of Singapore, Singapore

Dr. Jeong-Ju Yoo, ETRI, Korea

Prof. Cheon-Shik Kim, PhD, Anyang University, Korea

Prof. Dongsoo Stephen Kim, PhD, Sungkyunkwan University, USA

Prof. Rudra Datta, PhD, North Carolina State University, USA

Prof. Eun-Joon Yoon, PhD, Daegu Polytechnic College University, Korea

Prof. Wei-Tsung Su, PhD, Aletheia University, Taiwan

Dr. Enes Koytak, UN-SPIDER(Germany), Turkey

Prof. Eun-young Lee, PhD, Dongduk Womans University, Korea

Prof. Aysegul Yayimli, PhD, Istanbul Technical University, Turkey

Dr. GYU MYOUNG LEE, Institut TELECOM SudParis, France

Prof. You-Sik Hong, PhD, Sangji University, Korea

Prof. Hee-Cheol Lee, PhD, Huree University, Mongolia

Prof. Arash Dana, PhD, Central Tehran branch Islamic Azad University, Iran

Prof. Ra ilkyeun, PhD, University of Colorado Denver, USA

Dr. Derek Pao, City University, Hong Kong

Dr. Cuibo Yu, Beijing University of Posts and Telecommunications, China

Dr. Jose Gutierrez, Aalborg University, Denmark

Prof. junhui zhao, PhD, Beijing Jiaotong University, China

Prof. Jin-Seek Choi, PhD, Hanyang University, Korea

Prof. Jun-Kyun Choi, PhD, KAIST (Univ.), Korea

Prof. Khaled R. Ahmed, PhD, Southern Illinois University, USA

Dr. Byoung Whi Kim, ETRI, Korea

Prof. Kyung-Heon Koo, PhD, Incheon National University, Korea

Dr. Hsin-Kun Lai, Asia-Pacific Institute of Creativity, Taiwan

Dr. Mohamed M. A. Moustafa, Egyptian Russian University, Egypt

Prof. Takahiro Matsumoto, Kagoshima University, Japan

Dr. Ming-Shen Jian, Shu-Te University, Taiwan

Dr. Mitch Haspel, Stochastikos Solutions R&D, Jerusalem College of Technology, Israel

Prof. Mitsuji Matsumoto, PhD, Waseda University, Japan

Prof. Minseok Oh, PhD, Kyonggi University, Korea

Dr. Oscar Martinez Bonastre, Miguel Hernandez University of Elche, Spain

Dr. Prabu Do, EMC Corp, India

Prof. Sungchang Lee, PhD, Korea Aerospace University, Korea

Prof. S. Srinivasan, PhD, University of Louisville, USA

Prof. Lee Tony T., PhD, Chinese University, Hong Kong

Prof. Wen Guangjun, PhD, Univ. of Electronic Science and Technology of China, China

Prof. Whan-Woo Kim, PhD, CNU(University), Korea

Prof. Yao-Chung Chang, PhD, National Taitung University, Taiwan

Prof. Yong-Hee Jeon, PhD, Catholic Univ. of Daegu, Korea

Prof. Yong-Ik Yoon, PhD, Sookmyung University, Korea

Prof. Yeong Min Jang, PhD, Kookmin University, Korea

Dr. Yulei Wu, University of Bradford, UK

Dr. Abhishek D. Joshi, Seoul National University, Korea

Dr. Chun-Hsin Wang, Chung Hua University, Taiwan

Prof. Ying-Ren Chien, PhD, I-lan University, Taiwan

Prof. Dae-Ki Kang, PhD, Dongseo University, Korea

Dr. Dongkyu Kim, Yonsei Univ., Korea

Prof. Heung-Gyoon Ryu, PhD, Chungbuk National University, Korea

Dr. Eun-Hee Shin, Hanyang University, Korea

Prof. Francis Lau, PhD, Hong Kong Polytechnic University, Hong Kong

Dr. Yukinobu Fukushimam, Okayama University, Japan

Prof. Hoon Jae Lee, PhD, Dongseo University, Korea

Dr. Ming-Shen Jian, National Formosa University, Taiwan

Prof. Jongsub Moon, PhD, CIST, Korea University, Korea

Dr. Kyuchang Kang, ETRI, Korea

Prof. Lin You, PhD, Hangzhou Dianzi Univ, China

Dr. Mostafa Zaman Chowdhury, Kookmin University, Korea

Dr. Okgee Min, ETRI, Korea

Dr. Pasquale Pace, University of Calabria - DEIS - Italy, Italy

Dr. Razi Arshad, NUST(National University of Sciences and Technology), Pakistan

Dr. Che-Sheng Chiu, Chunghwa Telecom Laboratories, Taiwan

Prof. Altair O. Santin, PhD, PUCPR(Pontifical Catholic University of Parana), Brazil

Dr. Muhammad Shoaib Siddiqui, Kyung Hee University, Korea

Dr. Sang-Hwan Ryu, Korea Railroad Research Institute, Korea

Dr. Se-Jin Oh, Korea Astronomy & Space Science Institute, Korea

Dr. Tam Van Nguyen, National University of Singapore, Singapore

Dr. Xingzhong Xiong, Sichuan University of Science & Engineering, China

Dr. Youssef SAID, National Engineering School of Tunis/ Tunisie Telecom, Tunis

Prof. Namgi Kim. PhD, Kyonggi University, Korea

Prof. Byoung-Dai Lee, PhD, Kyonggi University, Korea

Prof. Daeyoung Kim. PhD, KAIST University, Korea

Prof. Tae Oh, PhD, Rochester Institute of Technology, USA

Prof. Bing-Yuh Lu, PhD, Guangdong University of Petrochemical Technology, China

Prof. Ashraf A. M. Khalaf, PhD, Minia University, Egypt

Prof. Harekrishna Misra, PhD, Institute of Rural Management Anand, Gujarat, India

Prof. Sunmoo KANG, PhD, Kyung Hee University, Korea

Prof. Fire Tomohisa Wada, University of the Ryukyus, JAPAN

Dr. Yue Wang, George Mason University, USA

Dr. Jin-Woong Cho, KETI, Korea

Prof. Ruslan Kirichek, PhD, St.Petersburg State University of Telecommunications, Russia

Prof. Kim Hyun Ah, PhD, Computer Science Dept. Kyonggi University, Korea

Prof. Dinh-Lam Pham, Vietnam National University, Vietnam

Prof. Seung Hyong Rhee, PhD, Kwangwoon University, Korea

Dr. Anhong Dang, Peking University, China

Prof. Syed Muhammad Owais, COMSTAS university Islamabad, Pakistan

Prof. Yong Jin, Tokyo Institute of Technology, Japan

Prof. NGUYEN Ha Nam, Vietnam National University, Vietnam

Prof. Ahn Hyun, PhD, Hanshin University , Korea

Prof. Viet-Vu Vu, PhD, Vietnam National University, Vietnam

Dr. Yoon-seok Ko, National Information Society Agency, Korea

Prof. Taku Yamazaki, Shibaura Institute of Technology, Japan

Dr. Hyunho Park, ETRI, Korea

Dr. Hyung-soon Kim, National Information Society Agency, Korea

Prof. Jiann-Liang CHEN, National Taiwan University of Science and Technology, Taiwan

Prof. Mangal Sain, PhD, Dongseo University, India

Dr. Chang-shin Chung, TTA, Korea

Prof. Otgonbayar Bataa, Mongolian University of Science and Technology, Mongolia

Prof. Ammar Muthanna, State University of Telecommunications, Russia

Dr. Artem Volkov, State University of Telecommunications, Russia

Active Reviewer List

Dr. Mengduo Zhou, Nanjing University of Posts and Telecommunications, China

Mr. Syed Muhammad Owais, COMSATS Institute of Information Technology (CIIT), Pakistan

Dr. Sangwoon Kwak, ETRI, Korea

Dr. Joungil Yun, ETRI, Korea

Dr. Hyoungsoo Lim, ETRI, Korea

Dr. Namho Hur, ETRI, Korea

Dr. Chandana Gamage, University of Moratuwa, Sri Lanka

Dr. Jeong-Woo Son, Electronics and Telecommucations Research Institut, Korea

Dr. Kamarulzaman Ab. Aziz, Multimedia University, Malaysia

Dr. Nagesh K.N, Nagarjuna College of Engineering & Technology, India

Dr. Seungsoo Yoo, Konkuk Univ., Korea

Dr. Sheikh Tahir Bakhsh, Computer Skills Unit, Faculty of Computing, KAAU, Saudi Arabia

Dr. Viktor Zaharov, Saint Petersburg State University, Rusia

Dr. Chun Yeow Yeoh, Telekom Research & Development Sdn. Bhd., Malaysia

Dr. ZengGuang Liu, Alcatel-Lucent, China

Prof. Gyu Myoung Lee, Liverpool John Moores University, United Kingdom

Prof. Saugata Bose, University of Liberal Arts Bangladesh, Bangladesh

Dr. Siwaruk siwamogsatham, NECTEC, Thailand

Dr. Yancui Shi, Tianjin University of Science& Technology, China

Dr. Ji-In Kim, Kyungpook National University, korea

Dr. Jay Kumar Jain, MANIT, Bhopal, India

Dr. Yong-Ju Lee, ETRI, Korea

Dr. Jaegon Kim, Korea Aerospace University, Korea

Dr. YongSoo Choi, SungKyul University, Korea

Prof. Sajjad Haider, NUML, Pakistan

Dr. Nagesh POOJARY, Middle East College, Knowledge Oasis, Muscat, Oman

Dr. Jongkuk Lee, ETRI, Korea

Dr. Javed Ferzund, COMSATS Institute of Information Technology (CIIT), Pakistan

Dr. Hyun Ahn, Kyonggi University, Korea

Prof. Borhanuddin Bin Mohd Ali, UPM, Malaysia

Dr. Ahmed Khairy, Alexandria University, Egypt

Dr. Yoshiyasu Takahashi, Hitachi ltd., Japan

Dr. Ji-in Kim, Silla Systems, Korea

Dr. Ashok Sapkal, College of Engineering Pune, India

Dr. Marsono, Muhammad Nadzir, Universiti Teknologi Malaysia, Malaysia

Dr. Hela Mliki, National School of Engineering Sfax University, Tunisia

Dr. Mohd Naz'ri Mahrin, Universiti Teknologi Malaysia, Malaysia

Dr. Soong-hee Lee, Inje University, Korea

Dr. Yongjun Ren, Nanjing University of Information Science & Technology, China

Prof. Liu Liu, Beijing Jiatong Uniersity, China

Prof. Yan Gao, Northeastern Univesity, China

Prof. Sulochana Sooriyaarachchi, PhD, University of Moratuwa, Sri Lanka

Dr. Irfan Zafar, Institute of Communication Technologies, Pakistan

Dr. Pardeep Kumar, University of Oxford, United Kingdom

Dr. Moon Kyeong-Deok, ETRI, Korea

Dr. Sang-Yun Lee, ETRI, Korea

Dr. Hsu-Chen CHENG, Wisdom Garden Research Center, Taiwan

Dr. Dr. Manu Pratap Singh, Dr. Bhim rao Ambedkar University, India

Prof. Myoungbeom Chung, Sungkyul University, Korea

Dr. Mamdoh Goda, Misr University, Egypt

Dr. Dr. Ritambhra Korpal, University of Pune, India

Dr. Sangil Choi, Korea Institute of Civil Engineering and Building Technology, Korea

Dr. Fahim Khan, The University of Tokyo, Japan

Dr. Juyul Lee, ETRI, Korea

Dr. Ting Peng, Changan University, China

Dr. Yonghoon Cho, Comesta Inc., Korea

Prof. Chu-Sing Yang, NCKU, Taiwan

Dr. Seohyun Jeon, ETRI, Korea

Prof. Zhong Chen, Peking University, China

Dr. Jae Hong MIN, ETRI, Korea

Prof. JUN ZHENG, Hankuk Academy of Foreign Studies, Korea

Dr. Ehab Adel, Alexandria University, Egypt

Dr. Charoenchai Wongwatkit, Mae Fah Luang University, Thailand

Prof. Kyuchang Kang, Kunsan University, Korea

Prof. Byeong-Nam Yoon, Kyonggi University, Korea

Prof. Hyeong-Ho Lee, Seoul National University of Science and Technology, Korea

Dr. Yung-Chien Shih, MediaTek Inc., Taiwan

Dr. H K Lau, The Open University of Hong Kong, Honh Kong

Dr. Sunghun KIM, ETRI, Korea

Dr. Mostafa Zaman Chowdhury, Kookmin University, Korea

Dr. Che-Sheng Chiu, Mobile Business Group, Chunghwa Telecom, Taiwan

Dr. Sung Moon Shin, ETRI, Korea

Dr. Fateme Khalili, K.N.Toosi. University of Technology, Ohio University, USA

Prof. Minseok Oh, Kyonggi University, Korea

Dr. Noor Zaman, King Faisal University, Al Ahsa Hofuf, Saudi Arabia

Prof. Young Woong Ko, PhD., Hallym University, Korea

Prof. Jun-Chul Chun, PhD., Kyonggi University, Korea

Dr. Jeong-Ju YOO, ETRI, Korea

Dr. Myungwon Song, National Information Society Agency, Korea

Prof. Harekrishna Misra, Institute of Rural Management Anand, India

Dr. Ming-Shen Jian, National Formosa University, Taiwan

Dr. Se-Jin Oh, Korea Astronomy & Space Science Institute, Korea

Prof. Man Soo Han, PhD., Mokpo National Univ., Korea

Prof. Hyo-Hoon Park, KAIST, Korea

Prof. Davar Pishva, Ritsumeikan Asia Pacific niversity (APU), Japan

Dr. Jens Myrup Pedersen, Aalborg University, Denmark

Dr. Tam Van Nguyen, University of Dayton, USA

Prof. Eun-young Lee, PhD., Dongduk Woman s University, Korea

Prof. Takahiro Matsumoto, Kagoshima University, Japan

Prof. Ying-Ren Chien, National Ilan University, Taiwan

Prof. Yao-Chung Chang, PhD., National Taitung University, Taiwan

Dr. Dongkyun Kim, KISTI(Korea Institute of Science and Technology Information), Korea

Dr. hyeokchan kwon, ETRI, korea

Dr. Tae-Gyu Lee, Korea Institue of Industrial Technology(KITECH), korea

Dr. Mangal Sain, Dongseo University, India

Dr. Cong Hung Tran, In Charge of Postgraduate and International Cooperation, Vietnam

Dr. Porkumaran K, NGP institute of technology India, India

Prof. CheonShik Kim, PhD., Sejong University, Korea

Prof. Jesuk Ko, PhD., Gwangju University, Korea

Prof. Plamena Zlateva, PhD., BAS(Bulgarian Academy of Sciences), Bulgaria

Dr. Kamrul Hasan Talukder, Hiroshima University, Khulna University, Bangladesh, Bangladesh

Dr. Saba Mahmood, Air University Islamabad Pakistan, Pakistan

Prof. Dae-Ki Kang, PhD., Dongseo University, Korea

Prof. Chen-Shie Ho, PhD., Oriental Institute of Technology, Taiwan

Prof. Andrey KOUCHERYAVY, St.Petersburg State Univerrsuty of Telecommunication, Russia

Prof. Hanxin WANG, South-Central University for Nationalities, China

Dr. Rajendra Prasad Mahajan, RGPV Bhopal, India

Dr. Kim, Do-Young, ETRI, Korea

Dr. ChuHwan Yim, KICI, Korea

Prof. Vishal Bharti, PhD., Dronacharya College of Engineering, India

Dr. Kyuchang Kang, ETRI, Korea

Prof. Sungchang Lee, Korea Aerospace University, Korea

Prof. Chang-Sheng Chen, National Chiao Tung University, Taiwan, Taiwan

Prof. Lakshmi Prasad Saikia, PhD., GIMT-Guwahati (ASTU), India

Dr. Razi Arshad, National University of Sciences and Technology, Pakistan

Prof. Jongsub Moon, Korea University, Korea

Prof. Dong-Her Shih, National Yunlin University of Science & Technology, Taiwan

Dr. Alireza Ghobadi, University Technology Malaysia/SOHA Sdn. Bhd., Malaysia

Dr. Hitoshi Okada, National Institute of Informatics, Japan

Dr. Chawalit Benjangkaprasert, King Mongkut's Institute of Technology Ladkrabang,, Thailand

Prof. Mai Yi-Ting, PhD., Hsiuping University of Science and Technology, Taiwan

Dr. Vasaka Visoottiviseth, Mahidol University, Thailand

Prof. Seok-Joo Koh, Kyungpook National University, Korea

Prof. Francis C.M. Lau, Hong Kong Polytechnic University, Hong Kong

Prof. Jaeshin Jang, Inje University, Korea

Dr. Woo-Jin Byun, ETRI, Korea

Dr. Chirawat Kotchasarn, Rajamangala University of Technology Thanyaburi, Thailand

Dr. Srinivas Mantha, SASTRA University, Thanjavur, India

Prof. Sherif Welsen Shaker, PhD., Kuang-Chi Institute of Advanced Technology, China

Dr. Chin-Feng Lin, National Taiwan Ocean University, Taiwan

Prof. Shintaro Uno, PhD., Aichi University of Technology, Japan

Prof. Chi-Chung Tao, Tamkang University, Taiwan

Prof. Seong Gon Choi, PhD., Chungbuk National University, Korea

Dr. Jing Li, State Key Laboratory of Astronautic Dynamics, China

Dr. Gyanendra Prasad Joshi, Yeungnam University, korea

Dr. Pasquale Pace, University of Calabria - DEIS - Italy, Italy

Prof. Sang-Sun Lee, Hangyang universit, Korea

Dr. Kim, Seong-Hwan, ETRI, Korea

Dr. Jung ho EOM, Daejeon University, Korea

Prof. Hwang Soo Lee, KAIST, Korea

Dr. Mohammed M. Kadhum, Queens University, Canada, Canada

Prof. Jitae Shin, Sungkyunkwan Univ., Korea

Prof. Hoon Jae Lee, Dongseo University, Korea

Dr. Dae Won Kim, ETRI, Korea

Dr. Anhong Dang, Peking University, China

Dr. Minsu Kim, SAMSUNG ELECTRONICS, Korea

Dr. Augustine Ikechi Ukaegbu, KAIST, Korea

Dr. S. Mehta, VIT, Mumbai, India, India

Dr. Yu-Doo Kim, LG Electronics, Korea

Dr. Alok AGGARWAL, AK Technical University (AKTU), Lucknow, India, India

Prof. Xiaofeng Qiu, Beijing University of Posts and Telecommunications, China

Dr. Tawfig Eltaif, Multimedia University, Malaysia

Prof. Minoru Okada, Nara Institute of Science and Technology, japan

Dr. Yoon-Seop Chang, ETRI, korea

Dr. Dong Kyoo Kim, ETRI, Korea

Dr. Zelalem Shibeshi, University of Fort Hare, South Africa

Dr. Yukinobu Fukushimam, Okayama University, Japan

Dr. Chun-Hsin Wang, Chung Hua University, Taiwan

Prof. Sang Uk Shin, Pukyong National University, Korea

Dr. Hyunho Park, ETRI, Korea

Prof. Bing-Yuh Lu, PhD, Guangdong University of Petrochemical Technology
Taipei University of Business, Taiwan

Prof. Yu-Shan Lin, National Taitung University, Taiwan

Prof. Kwang Hoon Kim, Kyonggi University, Korea

Dr. Ashraf Khalaf, Minia University, Egypt

Dr. Tadasuke Minagawa, Meiji University, Japan

Dr. Alexandru Murgu, Alexandru Murgu, University of Cape Town, South Africa

Dr. Changwoo YOON, ETRI, Korea

Dr. Okgee Min, ETRI, Korea

Dr. Abdallah Handoura, Engineering school of Gabes - Tunisia, Tunisia

Dr. Ahmad Rabiah, Universiti Teknikal Malaysia Melaka, Malaysia

Dr. Wimol San-um, Thai-Nichi Institute Technology, Thailand

Dr. Rafidah Md Noor, University of Malaya, Malaysia

Dr. Jin REN, North china university of technology, China

Dr. Hyung-soon Kim, National Information Society Agency, Korea

Prof. Chi-ho LIN, CHungbuk Univ., korea

Dr. Soo-Cheol Oh, ETRI, Korea

Prof. You-Sik Hong, Sangji University, Korea

Dr. Seong Joon Lee, Korea Electrotechnology Research Institute, korea

Prof. Lin You, PhD., Hangzhou Dianzi Univ, China

Dr. Twittie Senivongse, Chulalongkorn University, Thailand

Prof. Anisha Lal, PhD., VIT university, India

Dr. JongHyun Park, ETRI, Korea

Dr. Peng Gong, Beijing Institute of Technology, China

Dr. Meixiang Zhang, Yangzhou University, China

Dr. Pisit Boonsrimuang, King Mongkut‘s Institute of Technology Ladkrabang, Thailand

Dr. . Ho-Kyung Son, ETRI, Korea

Dr. Jun Liu, China Academy of Railway Sciences, China

Dr. aamir shahzad, Chon Buk National University, Korea

Dr. Rawya Rizk, Port Said University., Egypt

Prof. Byoung-Dai Lee, Kyonggi University, Korea

Dr. Noppamas Pukkhem, Faculty of Science, Thaksin University, Thailand

Dr. Muhamad Shahbani Abu Bakar, UUM Sintok, Malaysia

Dr. Sani Muhamad Isa, Bina Nusantara University, Indonesia

Dr. Pethur Raj CHELLIAH, IBM Global Cloud Center of Excellence, India

Prof. Dhananjay Singh, Hankuk University of Foreign studies (HUFS), korea

Dr. Hai Hoang, Post and Telecommunication Institute of Technology, Vietnam

Dr. Zhen Luo, Shenyang Aerospace University, China

Dr. Buseung Cho, KISTI, Korea

Prof. Bong Kyo MOON, Dongguk University, Korea

Prof. Jiann-Liang Chen, PhD., National Taiwan University of Science and Technology, Taiwan

Dr. Vladimir Mochalov, IKIR FEB RAS, Russia

Dr. Roger FAYE M., Universiti Cheikh Anta Diop Daka(UCAD/ESP), Senegal

Dr. Sudsanguan Ngamsuriyaroj, Mahidol University, Thailand

Prof. Vitaly Klyuev, Fukushima, Japan

Dr. Robiah Yusof, Universiti Teknikal Malaysia Melaka, Malaysia

Dr. manjur kolhar, Prince Sattam Bin Abdulaziz university, Kingdom of Saudi Arabia, Saudi Arabia

Dr. Vanvisa Chutchavong, King Mongkut’s Institute of Technology Ladkrabang, Thailand

Dr. Julian Webber, Advanced Telecommunications Research Institute International, Japan

Prof. Namgi Kim, Kyonggi University, Korea

Dr. Wen-Kuei Hsieh, De Lin Institute of Technology, China

Prof. Byung Gil LEE, ETRI, Korea

Dr. Olga Simonina, The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Russian Federation

Prof. Wei-Tsung Su, Aletheia University, Taiwan

Prof. Dimiter G. Velev, UNWE(University of National and World Economy), Bulgaria

Prof. Rongke LIU, Beihang University, China

Prof. Yuzo |ano, UNICAMP, Brazil

Dr. Bhavana Jharia, Ujjain Engineering College, Ujjain (M.P.), India

Prof. Kiyoung Kim, Seoil University, Korea

Dr. Waleed Saad, Menoufiya University, Egypt

Dr. CHOON SUNG NAM, Yonsei University, Korea

Dr. Heeyoung Jung, ETRI, Korea

Dr. Nemesio Jr. Macabale, Central Luzon State University, Philippines

Prof. WenWei Liao, University of Colorado Boulder, USA

Prof. Panos Fitsilis, University of Applied Sciences, Greece

Dr. Pauline Kongsuwan, Rajamangala University of Technology Thanyaburi, Thailand

Dr. Taeseon Yoon, Hankuk Academy of Foreign Studies, Korea

Prof. Marcelo Santos, UNICAMP, SENAC, Brazil

Prof. Kuo-Yu Tsai, Chinese Culture University, Taiwan

Prof. Dinh-Lam Pham, Vietnam National University, Vietnam

Dr. Ling-Yuan Hsu, St. Marys Junior College of Medicine, Nursing and Management, Taiwan

Dr. Jae Woo Kim, ETRI, Korea

Prof. Christelle Aupetit-Berthelemot, Xlim-University of Limoges, France

Prof. Shrishail MULE, SINHGAD COLLEGE OF ENGINEERING, India

Dr. Low Jung, Universiti Teknologi Petronas, Malaysia

Prof. Jun Tian, Fujitsu R&D Center, Co. LTD., China

Prof. Wha Sook Jeon, Seoul National University, Korea

Prof. Hany Harb, MUST University, IT College, Egypt

Dr. Jihyun Lee, ETRI, Korea

Dr. Sookjin LEE, ETRI, Korea

Dr. Nguyen-Son VO, Duy Tan University, Vietnam

Dr. Hoon Lee, ETRI, Korea

Dr. Fan Yang, Fujitsu Research and Development Center Co., Ltd, China

Dr. Hyenyoung YOON, Seoul National University, Korea

Prof. Yao-Liang Chung, National Taiwan Ocean University, Taiwan

Dr. Patrick Hosein, The University of the West Indies(UWI), Trinidad

Dr. Jaeho Lee, ETRI, Korea

Dr. Sushanth Babu Maganti, Jayamikhi Institute of Science & Technology, India

Dr. farhan ullah, COMSATS Institute of Information Technology (CIIT), Pakistan

Dr. Yanming CHENG, Beihua Universtiy, China

Prof. Lianfen HUANG, Xiamen University, China

Prof. Wei-Lung Mao, National Yunlin University of Science and Technol, Taiwan

Dr. Xianjun Yang, Fujitsu Research and Development Center, China

Prof. Praveen Kumar Devulapalli, Malla Reddy Engineering College, Thailand

Dr. Grigorii Fokin, The Bonch-Bruevich St. Petersburg State University of Telecommunications, Rusia

Dr. Ali A.Mohammed, Huazhong University of Science and technology, China

Dr. Ma Yuehong, Beihang University, China

Prof. Abdel-Aziz HASSANIN, Faculty of Electronic Engineering, Menufia Univers, Egypt

Dr. Noik Park, ETRI, Korea

Dr. Xianming Gao, National University of Defense Technology, China

Dr. HanSeok Kim, Samsung Electronics, Korea

Prof. Anh Vu Dinh Duc, Vietnam National University - Ho-Chi-Minh city, Vietnam

Dr. Yi-Huai Hsu, Industrial Technology Research Institute, Taiwan

Dr. Jongsuk Ruth Lee, KISTI, Korea

Dr. soyeon Lee, ETRI, Korea

Dr. Yong-Moo Kwon, Korea Institute of Science and Technology, Korea

Prof. Haksung Kim, Dongnam Health College, Korea

Dr. Ingeol Chun, ETRI, Korea

Dr. syed muhammad owais, COMSATS Institute of Information Technology (CIIT), Pakistan

Prof. Junkyun Choi, KAIST, Korea

Prof. Meng-Lun Hsueh, Hwa Hsia University of Technology, Taiwan

Dr. Kwihoon Kim, ETRI, Korea

Dr. Ming-Sheg Jian, National Formosa University, Taiwan

Dr. Zhaoqiang Xia, Northwestern Polytechnical University, China

Dr. An Wu, University of Science and Technology of China, China

Prof. Sumalatha Bandari, Dr. Daulatrao Aher College of Engineering, Sumalatha

Prof. Loc Ho Dac, Ho Chi Minh City University of Technology(HUTECH), Vietnam

Dr. TaeYeon Kim, ETRI, Korea

Prof. Jibin Yang, PLA university of science and technology, China

Dr. Hyungekeuk Lee, ETRI, Korea

Dr. Ahmed AbdElAziz, Arab Academy for Science and Technology, Egypt

Prof. Mohamed RIZK, Alexandria University, Egypt

Prof. Jong RHEE, Myongji University, Korea

Dr. Ruslan Kirichek, St.Petersburg State University of Telecommunications, Russia

Dr. Nae-Soo Kim, ETRI, Korea

Dr. Bo Gu, Kogakuin University, Japan

Dr. Beihua Ying, Ningbo Institute of Technology, Zhejiang Universit, China

Prof. Alan Marshall, University of Liverpool, United Kingdom

Dr. Hazrat Ali, COMSATS Institute of Information Technology Abbott, Pakistan

Dr. Kyeong-Hoon Jung, Kookmin University, Korea

Prof. Hyo Hoon Park, KAIST, Korea

Dr. Rustam Pirmagomedov,
The Bonch-Bruevich St. Petersburg State University of Telecommunications, Rusia

Dr. Aida Shafiabady, Universiti Teknologi Malaysia, Malaysia

Dr. Seng-Kyoun Jo, ETRI, Korea

Prof. Baosheng Wang, College of Computer, NUDT, China

Dr. James TAMGNO KOUAWA, ESMT, Senegal

Dr. Chih-Chien Hu, National Chiao Tung University, Taiwan

Dr. Sang Gi Hong, ETRI, Korea

Prof. Euihyun Jung, Anyang University, Korea

Dr. Chih-Lin HU, National Central University, Taiwan

Dr. Andrei Vladyko, The Bonch-Bruevich Saint-Petersburg State University of Telecommunications (SPbSUT), Russia

Dr. Trong Thua Huynh, PTITPosts and Telecommunications Institute of Technology), Vietnam

Prof. Hsien-Chou Liao, Chaoyang University of Technology, Taiwan

Dr. Konstantin Izrailov,
The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Rusia

Dr. Raihana Syahirah Abdullah, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia

Prof. Megha Ainapurkar, Goa University, India

Dr. Hamed Shawky Zied, Alexandria University, Egypt

Dr. Jingpei Wang, China CEPREI Laboratory, China

Dr. SHEKH FAISAL ABDUL-LATIP, Universiti Teknikal Malaysia Melaka, Malaysia

Dr. Rongchen Sun, Beijing Jiaotong University, China

Prof. Xiaoming Hu, Shanghai Polytechnic University, China

Dr. Siti Rahayu Selamat, Universiti Teknikal Malaysia Melaka, Malaysia

Dr. mehwish mukhtar, UOG Sialkot, Pakistan

Prof. Jie Liu, China CEPREI Laboratory, China

Dr. Soo-Hyung Kim, ETRI, Korea

Dr. Naeem Khan, UET Peshawar, Pakistan

Dr. HYUN WOOK, ETRI, korea

Prof. Chae-Woo Lee, Ajou University, Korea

Prof. Li-Der Chou, PhD., National Central University, Taiwan

Dr. Seok Ho Won, ETRI, Korea

Dr. Maxim Zaharov, Saint-Petersburg State University of Telecommunications, Russia

Dr. Pornpimon CHAYRATSAMI, King Mongkut Institute of Technology Ladkrabang, Thailand

Dr. Brij Gupta, National Institute of Technology Kurukshetra, India

Dr. Konstantin Izrailov, The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Rusia

Prof. Hee Chang JUNG, Dongei University, Korea

Prof. Chwen-Yea Lin, PhD., Tatung Institute of Commerce and Technology, Taiwan

Dr. Yue Wang, George Mason University, USA

Dr. Somkiat Kitjongthawonkul, Australian Catholic University, St Patricks Campus, Australia

Dr. Brownson ObaridoaObele, Hyundai Mobis Multimedia R&D Lab, Korea

Dr. Chien-Chung Tu, School of Information Mana, Taiwan

Prof. Hsu-Chen Cheng, Chinese Culture University, Taiwan

Dr. Sharifah Kamilah Yusof, Universiti Teknologi Malaysia, Malaysia

Dr. Yong Sun, Beijing University of Posts and Telecommunications, China

Prof. Seongjoo Lee, Sejong University, Korea

Dr. Sawky Shaaban, Alexandria University, Egypt

Dr. Hoang Dang Hai, Post and Telecommunication Institute of Technology, Vietnam

Dr. Svetlana KIM, Sookmyung Womens Univ., Korea

Prof. Chao-Lieh Chen, NKFUST(National Kaohsiung First Univ. of Sci-Tech), Taiwan

Dr. Mitch Haspel, Stochastikos Solutions R&D, Israel

Prof. Hosung Jo, Hanyang University, Korea (s)

Dr. JongWon Kim, GIST (Gwangju Institute of Science & Technology), Korea

Dr. Xi Chen, State Grid Corparation of China, China

Prof. Arash Dana, PhD., Islamic Azad university , Central Tehran Branch, Iran

Dr. Tony Tsang, Hong Kong Polytechnic UNiversity, Hong Kong

Prof. Kwang-Hoon Kim, PhD., Kyonggi University, Korea

Prof. Rosilah Hassan, PhD., Universiti Kebangsaan Malaysia(UKM), Malaysia

Dr. Christian Esteve Rothenberg, CPqD - R&D Center for. Telecommunications, Brazil

Prof. Moo Wan Kim, PhD., Tokyo University of Information Sciences, Japan

Prof. Yong-Hee Jeon, PhD., Catholic Univ. of Daegu, Korea

Dr. E.A.Mary Anita, Prathyusha Institute of Technology and Management, India

Prof. Wilaiporn Lee, PhD., King Mongkut's University of Technology North, Thailand

Dr. Zhi-Qiang Yao, XiangTan University, China

Prof. Bin Shen, PhD., Chongqing Univ. of Posts and Telecommunications (CQUPT), China

Mr. Muhammad Yasir Malik, Samsung Electronics, Korea

Prof. Yeonseung Ryu, PhD., Myongji University, Korea

Dr. Pasi Ojala, University of Oulu, Finland

Dr. Anna bruno, University of Salento, Italy

Prof. Zhiming Cai, PhD., Macao University of Science and Technology, Macau

Mr. Jose Gutierrez, Aalborg University, Denmark

Dr. Youssef SAID, Tunisie Telecom, Tunisia

Dr. Shahriar Mohammadi, KNTU University, Iran

Prof. Beonsku An, PhD., Hongik University, korea

Dr. Guanbo Zheng, University of Houston, USA

Prof. Sangho Choe, PhD., The Catholic University of Korea, korea

Prof. Ilkyeun Ra, PhD., University of Colorado Denver, USA

Dr. Yulei Wu, Chinese Academy of Sciences, China

Mr. Anup Thapa, Chosun University, korea

Dr. Vo Nguyen Quoc Bao, Posts and Telecommunications Institute of Technology, Vietnam

Dr. Harish Kumar, Bhagwant institute of technology, India

Dr. Joseph Kandath, Electronics & Commn Engg, India

Dr. Mohamed M. A. Moustafa, Arab Information Union (AIU), Egypt

Prof. Francis C.M. Lau, PhD., Hong Kong Polytechnic University, Hong Kong

Prof. Ju Bin Song, PhD., Kyung Hee University, korea

Prof. KyungHi Chang, PhD., Inha University, Korea

Prof. Seung-Hoon Hwang, PhD., Dongguk University, Korea

Prof. Dal-Hwan Yoon, PhD., Semyung University, korea

Prof. Chongyang ZHANG, PhD., Shanghai Jiao Tong University, China

Prof. Ying-Ren Chien, PhD., Department of Electrical Engineering, National Ilan University, Taiwan

Dr. Sang-Hwan Ryu, Korea Railroad Research Institute, Korea

Dr. Kuan Hoong Poo, Multimedia University, Malaysia

Dr. Michael Leung, CEng MIET SMIEEE, Hong Kong

Dr. Abu sahman Bin mohd Supaat, Universiti Teknologi Malaysia, Malaysia

Prof. Amit Kumar Garg, PhD., Deenbandhu Chhotu Ram University of Science & Technology, India

Dr. Jamshid Sangirov, KAIST, Korea

Prof. Ahmed Dooguy KORA, PhD., Ecole Sup. Multinationale des Telecommunications, Senegal

Dr. Mohammed M. Kadhum, School of Computing, Goodwin Hall, Queens University, Canada

Dr. Gopal Chandra Manna, BSNL, India

Dr. Il Kwon Cho, National Information Society Agency, Korea

Prof. Ruay-Shiung Chang, PhD., National Dong Hwa University, Taiwan

Dr. Yong-Sik Choi, Research Institute, IDLE co., ltd, Korea

Dr. Xuena Peng, Northeastern University, China

Dr. Soobin Lee, KAIST Institute for IT Convergence, Korea

Prof. Yongpan Liu, PhD., Tsinghua University, China

Prof. Chih-Lin HU, PhD., National Central University, Taiwan

Dr. Hyoung-Jun Kim, ETRI, Korea

Prof. Bernard Cousin, PhD., IRISA/Universite de Rennes 1, France

Dr. Feng CHENG, Hasso Plattner Institute at University of Potsdam, Germany

Prof. El-Sayed M. El-Alfy, PhD., King Fahd University of Petroleum and Minerals, Saudi Arabia

Mr. Nicolai Kuntze, Fraunhofer Institute for Secure Information Technology, Germany

Dr. Min-Hong Yun, ETRI, Korea

Dr. Jin Woo HONG, Electronics and Telecommunications Research Inst., Korea

Dr. Heeseok Choi, KISTI(Korea Institute of Science and Technology Information), korea

Dr. Ho-Jin CHOI, KAIST(Univ), Korea

Dr. Su-Cheng HAW, Multimedia University, Faculty of Information Technology, Malaysia

Dr. Myoung-Jin Kim, Soongsil University, Korea

Dr. Gyu Myoung Lee, Institut Mines-Telecom, Telecom SudParis, France

Prof. Yoonhee Kim, PhD., Sookmyung Women s University, Korea

Prof. Dimiter G. Velev, PhD., UNWE(University of National and World Economy), Bulgaria

Prof. Jun-Kyun Choi, PhD., KAIST (Univ.), Korea

Dr. kyeong kang, University of technology sydney, faculty of engineering and IT, Australia

Dr. Nirmalya Thakur, University of Cincinnati in Ohio, USA

Prof. K.L. Sudha, DSCE, VTU, Bangalore, Karnataka, India

Prof. Taku Yamazaki, Shibaura Institute of Technology, Japan

Dr. leila mohammadi, ITRC, Iran

Dr. Ealwan Lee, GCT Semiconductor, Inc., Korea

Dr. Abidah Mat Taib, Universiti Teknologi MARA, Malaysia

Prof. sohail jabbar, CIIT, Pakistan

Prof. Youngil Kwon, Hoseo University, Korea

Prof. Viet-Vu Vu, PhD, Vietnam National University, Vietnam

Dr. Youssef SAID, National Engineering School of Tunis/ Tunisie Telecom, Tunis

Prof. Yong-Ik Yoon, PhD, Sookmyung University, Korea

Prof. Yao-Chung Chang, PhD, National Taitung University, Taiwan

Prof. Francis C.M. Lau, Hong Kong Polytechnic University, Hong Kong

Prof. Do Nang Toan, PhD, ITI Vietnam National University, Vietnam

Prof. Ali Abdulwahhab Mohammed, PhD, KUS, Iraq

Prof. K.L. Sudha, Dayananda Sagar College of Engineering, Bengaluru

Prof. Ahn Hyun, PhD, Hanshin University , Korea

Dr. Jeong-Ju Yoo, ETRI, Korea

Prof. Khaled R. Ahmed, PhD, Southern Illinois University, USA

Dr. Chin-Feng Lin, NTOU University, Taiwan

Prof. Minjae Park, PhD, Computer Software Dept. Daelim University, Korea

Dr. Mostafa Zaman Chowdhury, KUET, BD

Prof. Radhakrishna Bhat, Ph.D,, Manipal Institute of Technology, India

Dr. Tae-Gyu Lee, PTU, Korea

Prof. Yong Jin, TITECH, Japan

Local Arrangement Committee

Prof. Kim Hyun-Ah, PhD, (LAC Chair) Kyonggi University, Korea

Dr. Myung-Won Song, (LAC Co-Chair) National Information Society Agency, Korea

Mr. Yoon-seok Ko, National Information Society Agency, Korea

Dr. Hyung-soon Kim, National Information Society Agency, Korea

Prof. Ahn Hyun, Hanshin University , Korea

Ms. Young-Ae Park, National Information Society Agency, Korea

Mr. Joo Hyung-Joo, Kyonggi University, Korea

Mr. Hong Seok-Woo, Kyonggi University, Korea

Mr. Seo Yoon-Deuk, Kyonggi University, Korea

Ms. Jae-Eun Song, Kyunghee University, Korea

Mr. Kim Sun-Pil, Kyonggi University, Korea

Mr. Lee Sun-Ho, Kyonggi University, Korea

Ms. Ji-Eun LEE, Ehwa Woman s University, Korea

Mr. Jeong Hyeon Il, Kyonggi University, Korea

Ms. Kim Mee Sun, Kyonggi University, Korea

Mr. Lee Do Kyeong, Kyonggi University, Korea

Ms. Joo-Hee Yoo, Kyunghee University, Korea

ⓒ Copyright 2024. ICACT (ISSN 1738-9445) & Global IT Research Institute (GiRI) All rights Reserved . Contact: office (at) icact . org Tel:+82-70-4146-4991

1713 Obelisk, 216 Seohyunno, Bundang-gu, Sungnam Kyunggi-do, Republic of Korea 13591

Business License Number : 220-82-07506, President: Thomas ByeongNam Yoon Ph.D.

Outstanding Paper Awards

Paper ID: 20240040

Title: Performance Evaluation of UAV-based NOMA for 5G and Beyond

Topic : Mobile Communication, 5G, 6G, Cloud & Mobile Computing

Keyword : 6G, EE, NOMA, OMA, UAV, SE, SNR.

Institute : NITW

Country : India

Author : Ms. MOUNIKA NEELAM, Prof. Anuradha Sundru,

Paper ID: 20240058

Title: Design of Ka-band Chip Antenna Based on Slot Antenna

Topic : Wireless Communication, Satellite Communications, AESA

Keyword : Antenna on-chip, millimeter-wave, CMOS

Institute : Electronic Engineering

Country : Taiwan

Author : Prof. Ming-An Chung, Mr. Kai-Xiang Chen, Mr. Kuo-Chun Tseng,

Paper ID: 20240070

Title: QPSO-based Beamforming in Dual RIS-assisted Uplink Anti-jamming Communication System

Topic : Wireless Communication, Satellite Communications, AESA

Keyword : anti-jamming, RIS, QPSO, beamforming, SINR

Institute : Shandong University

Country : China

Author : Ms. Di Zhou, Prof. Zhiquan Bai, Ms. Jinqiu Zhao, Mr. Zeyu Liu, Prof. Dejie Ma, Prof. KyungSup Kwak,

Paper ID: 20240076

Title: A Test Method for the Convergence of the Metaverse and Blockchain

Topic : Metaverse, Computer Vision, Graphics & Image Processing, XR, AR, VR, HMD

Keyword : Blockchain, Metaverse, Cryptocurrency, Testing, Evaluation

Institute : Pyeongtaek University

Country : Korea(South)

Author : Prof. Tae-gyu Lee,

Paper ID: 20240199

Title: Utilizing Machine Learning for Sensor Fault Detection in Wireless Sensor Networks

Topic : Communication Network, Optical, Internet, Router, Networks-on-Chip

Keyword : fault detection, WSN, SVM, ANN, machine learning

Institute : Nazarbayev University

Country : Kazakhstan

Author : Mr. Abubakar Abdulkarim, Mr. Israel Ehile, Prof. Refik Caglar Kizilirmak,

Paper ID: 20240241

Title: Dual-RIS Assisted 3D Positioning and Beamforming Design in ISAC System

Topic : Mobile Communication, 5G, 6G, Cloud & Mobile Computing

Keyword : Integrated sensing and communication (ISAC), reconfigurable intelligent surface (RIS), stepwise matching pursuit (SMP), beamforming design

Institute : Shandong University

Country : China

Author : Mr. Dejie Ma, Prof. Zhiquan Bai, Dr. Jinqiu Zhao, Mr. Hao Xu, Mr. Zeyu Liu, Dr. Di Zhou, Prof. Mingyan Jiang, Prof. KyungSup Kwak,

Paper ID: 20240269

Title: Hybrid Clustering Mechanisms for High-Efficiency Intrusion Prevention

Topic : Information & Network Security, Vulnerability, OWASP, DDoS

Keyword : K-means Algorithm, MITRE ATTACK, Snort, Locality Sensitive Hashing, Malicious Packet

Institute : www.ntust.edu.tw

Country : Taiwan

Author : Ms. Pin-Shan Lin, Mr. Yi-Cheng Lai, Ms. Man-Ling Liao, Ms. Shih-Ping Chiu, Prof. Jiann-Liang Chen,

Paper ID: 20240309

Title: The development of new system for generating training data of AI-based anomaly detection

Topic : Information & Network Security, Vulnerability, OWASP, DDoS

Keyword : anomaly detection, artificial intelligence (AI), dataset, training data, cybersecurity

Institute : Chungbuk National University

Country : Korea(South)

Author : Ms. Thi My Truong, Dr. Won Seok Choi, Mr. Jang Hyeon Jeong, Prof. Seong Gon Choi,

Paper ID: 20240342

Title: Utterance-Level Incongruity Learning Network for Multimodal Sarcasm Detection

Topic : Recommender System, AI, Deep Learning, Big Data, Data Mining

Keyword : multimodal sarcasm detection, utterance-level attention, incongruity learning

Institute : Computer Network Information Center

Country : China

Author : Dr. Liujing Song, Dr. Zefang Zhao, Prof. Yuxiang Ma, Dr. Yuyang Liu, Prof. Jun Li,

Paper ID: 20240372

Title: Overview of the potentials of multiple instance learning in cancer diagnosis: Applications, challenges, and future directions

Topic : Metaverse, Computer Vision, Graphics & Image Processing, XR, AR, VR, HMD

Keyword : Weakly supervised learning, Multiple Instance Learning, Bag of instances, Cancer diagnosis

Institute : Inje University

Country : Korea(South)

Author : Mr. Tagne Poupi Theodore Armand, Mr. Subrata Bhattacharjee, Prof. Hee-Cheol Kim,

Paper ID: 20240396

Title: Generalized Parabola Chaotic map for Pseudorandom Random Number Generator

Topic : Authentication, Bio-metric, Private Security, Facial Recognition

Keyword : chaotic map, discrete-time chaotic, parabola function, pseudo random number generator, NIST

Institute : King Mongkut University of Technology

Country : Thailand

Author : Dr. Nattagit Jiteurtragool,

Paper ID: 20240414

Title: Multicore Packet Distribution method using Multicore Network Interface Card based on Tile-gx72 Network Processor

Topic : Communication Network, Optical, Internet, Router, Networks-on-Chip

Keyword : Network Interface Card, Tile-gx72, Network Processor, HPC

Institute : Chungbuk National Univertisy

Country : Korea(South)

Author : Dr. Choi Won Seok, Mr. Lee Sang Ju , Mr. Kim Jong Oh, Prof. Choi Seong Gon,

Paper ID: 20240417

Title: Micro-services internal load balancing for Ultra Reliable Low Latency 5G Online charging system

Topic : System Work Method, Software Development

Keyword : Service mesh, Connection library, Internal load balancing, Benchmark, 5G, URLLC, Online-charging system

Institute : Viettel High Technology

Country : Viet Nam

Author : Mr. Ngoc Tien Nguyen, Mr. Thanh Son Pham, Mr. Van Duong Nguyen, Dr. Cong Dan Pham, Mr. Duc Hai Nguyen,

Paper ID: 20240426

Title: Flexible Localization Method with Motion Estimation for Underwater Wireless Sensor Networks

Topic : Communication Network, Optical, Internet, Router, Networks-on-Chip

Keyword : Underwater Wireless Sensor Networks (UWSN), Localization, Movement Estimation, Tracking, Sensor Deployment

Institute : University of Science and Technology of China

Country : China

Author : Dr. Abdelrahman Samy, Prof. Ammar Hawbani, Prof. Xingfu Wang, Dr. Samah Abdel Aziz, Prof. Liang Zhao, Prof. Nasir Saeed,

Paper ID: 20240441

Title: Design of Calibration Algorithms for Fully-Activated Millimeter-Wave Phased Array Antennas

Topic : Wireless Communication, Satellite Communications, AESA

Keyword : millimeter-wave phased array, rotating element electric field vector, software defined radio platform, calibration.

Institute : Yuan Ze University

Country : Taiwan

Author : Prof. Juinn-Horng Deng, Mr. Xiang-He Huang, Dr. Chung-Lien Ho, Ms. Yu-Chien Wu,

Paper ID: 20240459

Title: Intelligent Anomaly Detection System Based on Ensemble and Deep Learning

Topic : Hacking & Defense Security, Malware, Macro, Ransomware

Keyword : overfitting, ensemble, recall, precision, classification accuracy, generalizing

Institute : Bradley University

Country : USA

Author : Dr. Babu Baniya, Mr. Thomas Rush,

Paper ID: 20240469

Title: A Review of Detection-related Multiple Object Tracking in Recent Times

Topic : Recommender System, AI, Deep Learning, Big Data, Data Mining

Keyword : Multi-object Tracking (MOT), Object detection, Deep learning, Research progress

Institute : Henan University

Country : China

Author : Ms. Suya Li, Ms. Ying Cao, Ms. Xin Xie,

Paper ID: 20240475

Title: An Enhanced Topic Modeling Method in Educational Domain by Integrating LDA with Semantic

Topic : Recommender System, AI, Deep Learning, Big Data, Data Mining

Keyword : Topic Modeling, Online Discussion, Text Mining, Machine Learning

Institute : South China Normal University

Country : China

Author : Mr. Ruofei Ding, Mr. Pucheng Huang , Ms. Shumin Chen, Mr. Jiale Zhang, Mr. Jingxiu Huang, Mr. Yunxiang Zheng,

Paper ID: 20240480

Title: Location based Data-centric Forwarding for Mobile Ad-hoc Networks

Topic : Future Network, Information Centric Network

Keyword : MANET, Named-data Networking, Forwarding strategy, network simulation

Institute : University of Colorado Denver

Country : USA

Author : Mr. Hieu Nguyen, Prof. Ilkyeun Ra,

Paper ID: 20240482

Title: A Study on Real-time Evaluation of Uncertainty of PM-10 Concentration Determined by Tele-measuring Instrument

Topic : System Work Method, Software Development

Keyword : Real-time uncertainty evaluation. Measurement uncertainty, Particulate matter, PM-10, Air quality monitoring instrument, T ele-measuring system

Institute : Konkuk University

Country : Korea(South)

Author : Mr. Jeeho Kim, Dr. Jin-Chun Woo, Prof. Young Sunwoo,

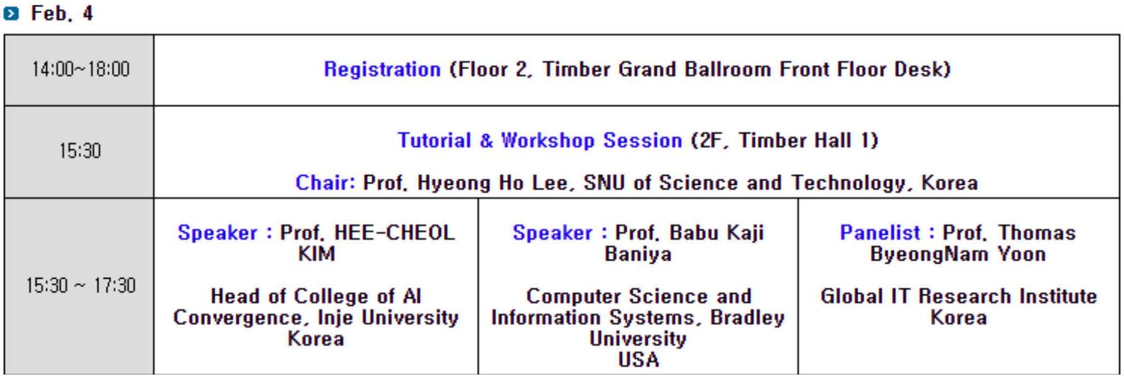
Author Index

Name	Paper No	Session	Organization	nationality
Dr. Abdelrahman Samy	20240426	5A	University of Science and	Egypt
Prof. Ammar Muthanna	20240130	3A	The Bonch-Bruevich Saint	Russia
Prof. Ammar Muthanna	20240202	3A	The Bonch-Bruevich Saint	Russia
Dr. Artem Volkov	20240366	3A	Department of	Russia
Dr. Hao Yang	20240321	4B	Huawei	China
Dr. Hyeonguk Jang	20240037	2B	ETRI	Korea(South)
Dr. Jennifer Llovido	20240471	3C	Bicol University	Philippines
Dr. Jin-Chun Woo	20240482	5B	KOSTEC, Inc	Korea(South)
Dr. Lea Austero	20240444	3B	Bicol University	Philippines
Dr. Michael Angelo	20240438	5B	Bicol University	Philippines
Dr. Nattagit Jiteurtragool	20240396	1C	KMUTNB	Thailand
Dr. Won Seok Choi	20240414	5A	Chungbuk National	Korea(South)
Mr. Bangwei He	20240369	4A	Shandong University	China
Mr. Birahim BABOU	20240393	4B	University of Dakar - UCAD	Senegal
Mr. Cheng-He Wang	20240079	2C	National Formosa University	Taiwan
Mr. Dejie Ma	20240241	3A	Shandong University	China
Mr. Harish V	20240467	2C	CDAC	India
Mr. Hieu Nguyen	20240480	4A	University of Colorado	Viet Nam
Mr. HyeokSoo Lee	20240049	2B	Sungkyunkwan University	Korea(South)
Mr. Israel Ehile	20240199	5A	Nazarbayev University	Nigeria
Mr. Israel Ehile	20240287	2A	Nazarbayev University	Nigeria
Mr. Jeeho Kim	20240482	5B	Konkuk University	Korea(South)
Mr. Juno Choi	20240312	4C	Pukyong National University	Korea(South)
Mr. Maisam Ali	20240384	4B	Inje University	Pakistan
Mr. Md Ariful Islam	20240479	4B	Inje University	Bangladesh
Mr. Min Gu Kang	20240333	5B	Chungbuk National Univ.	Korea(South)
Mr. Ming-Hsun Tsai	20240034	2C	National Formosa University	Taiwan
Mr. Minhwa Hong	20240299	5B	Chungbuk National	Korea(South)
Mr. Muhammad Yaseen	20240327	5B	Inje University	Pakistan
Mr. Ngoc Tien Nguyen	20240417	3C	Viettel High Technology	Viet Nam
Mr. Ruofei Ding	20240475	2B	South China Normal	China
Mr. Shah Muhammad	20240405	5B	Inje University	Bangladesh
Mr. Sikandar Ali	20240402	3B	Inje University	Pakistan
Mr. Tagne Poupi	20240345	5B	Inje University	Cameroon
Mr. Tagne Poupi	20240372	5B	Inje University	Cameroon

Mr. Tagne Poupi	20240411	4B	Inje University	Cameroon
Mr. Thanh Son Pham	20240417	3C	Viettel High Technology	Viet Nam
Mr. Wenyi Li	20240387	4A	Beijing Institute of	China
Mr. Woo-Hyeon Kim	20240357	1B	Kyonggi University	Korea(South)
Mr. Yincai CAI	20240073	5B	Changan University	China
Mr. Yincai CAI	20240339	5B	Changan University	China
Mr. Youngchul Kim	20240330	3C	ETRI	Korea(South)
Mr. Yuan Sun	20240375	3A	Shandong University	China
Ms. Di Zhou	20240070	2A	Shandong University	China
Ms. Hyekyoung Hwang	20240217	1B	Sungkyunkwan University	Korea(South)
Ms. Kouayep Sonia	20240453	3B	Inje university	Cameroon
Ms. Kounen Fathima	20240477	4B	Inje University	India
Ms. Liujing Song	20240342	1B	Computer Network	China
Ms. Narantuya	20240470	4A	Mongolian Defence	Mongolia
Ms. Pin-Shan Lin	20240269	1C	www.ntust.edu.tw	Taiwan
Ms. Suya Li	20240469	2B	Henan University	China
Ms. Tai-Ying Chiu	20240272	1C	National Taiwan University of	Taiwan
Ms. Thi My Truong	20240309	1C	Chungbuk National	Viet Nam
Ms. Tianzhu Hu	20240378	5A	USTC	China
Ms. Vatcharavaree	20240324	1C	Mahidol University	Thailand
Ms. Xiang DONG	20240061	4C	Changan University	China
Ms. Xiang DONG	20240348	5B	Changan University	China
Ms. Xiaoling Niu	20240456	3C	The China Academy of	China
Prof. Anuradha S	20240067	5A	NIT Warangal	India
Prof. Anuradha Sundru	20240040	4A	NITW	India
Prof. Babu Baniya	20240459	2C	Bradley University	Nepal
Prof. Bing-Yuh Lu	20240429	4C	Guangdong University of	China
Prof. Bing-Yuh Lu	20240435	4C	Guangdong University of	China
Prof. Christian Sy	20240447	3B	Bicol University	Philippines
Prof. Dal-Hwan Yoon	20240336	1B	Semyung University	Korea(South)
Prof. Francis C.M. Lau	20240009	2A	Hong Kong Polytechnic	Hong Kong
Prof. Heeyoul Kim	20240031	2C	Kyonggi University	Korea(South)
Prof. Hyejin S. Kim	20240043	1B	ETRI	Korea(South)
Prof. Jiann-Liang Chen	20240248	2B	National Taiwan University of	Taiwan
Prof. Juinn-Horng Deng	20240441	1A	Yuan Ze University	Taiwan
Prof. Ming-An Chung	20240052	1A	Electronic Engineering	Taiwan
Prof. Ming-An Chung	20240055	1A	Electronic Engineering	Taiwan
Prof. Ming-An Chung	20240058	1A	Electronic Engineering	Taiwan
Prof. Narantuya	20240473	2A	SICT, MUST	Mongolia
Prof. Otgonbayar Bataa	20240473	2A	SICT, MUST	Mongolia
Prof. Seung Yeob Nam	20240293	3C	Yeungnam University	Korea(South)

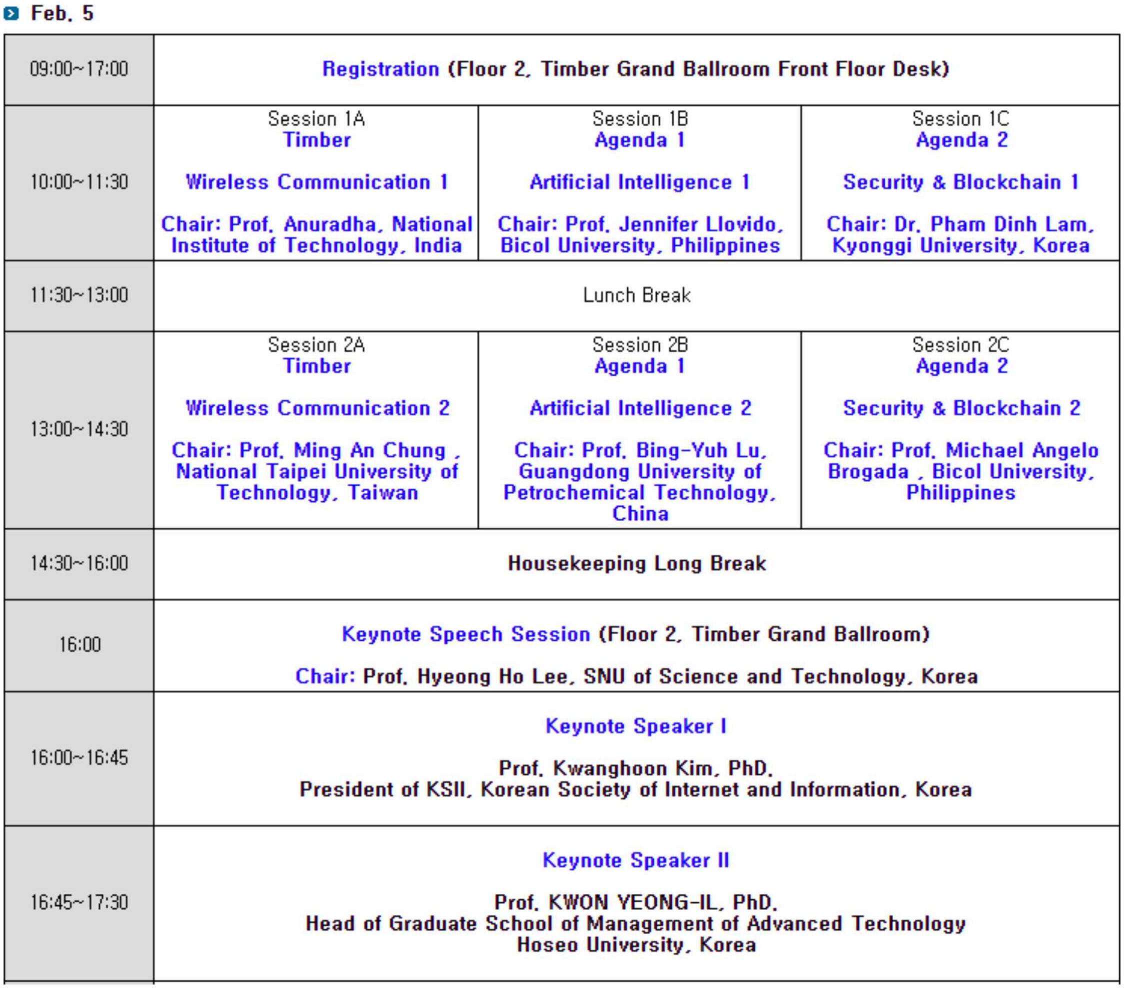
Prof. Tae-gyu Lee	20240076	4C	Pyeongtaek University	Korea(South)
Prof. Tin-Yu Wu	20240381	1A	NPUST	Taiwan
Prof. Vakula Damera	20240022	2A	NITW	India
Prof. Yao-Chung Chang	20240064	2C	National Taitung University	Taiwan
Prof. Yao-Chung Chang	20240208	1B	National Taitung University	Taiwan

ICACT2024 Conference Program



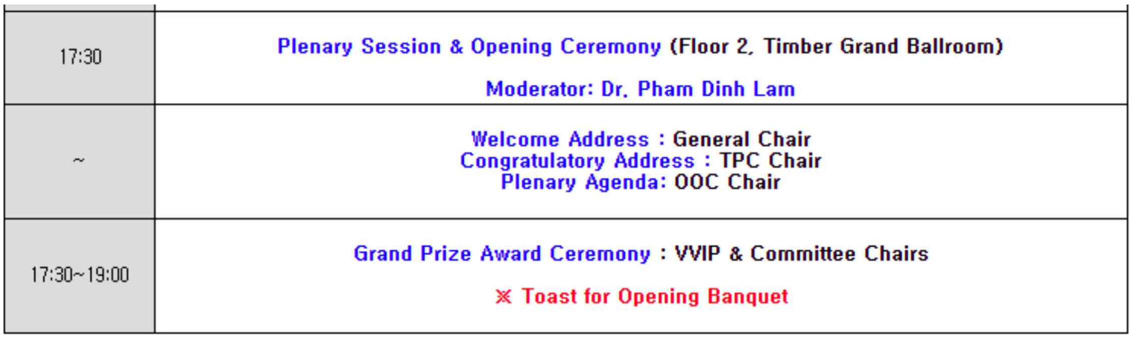
Feb. 4

14:00~18:00	Registration (Floor 2, Timber Grand Ballroom Front Floor Desk)		
15:30	Tutorial & Workshop Session (2F, Timber Hall 1) Chair: Prof. Hyeong Ho Lee, SNU of Science and Technology, Korea		
15:30 ~ 17:30	Speaker : Prof. HEE-CHEOL KIM Head of College of AI Convergence, Inje University Korea	Speaker : Prof. Babu Kaji Baniya Computer Science and Information Systems, Bradley University USA	Panelist : Prof. Thomas ByeongNam Yoon Global IT Research Institute Korea



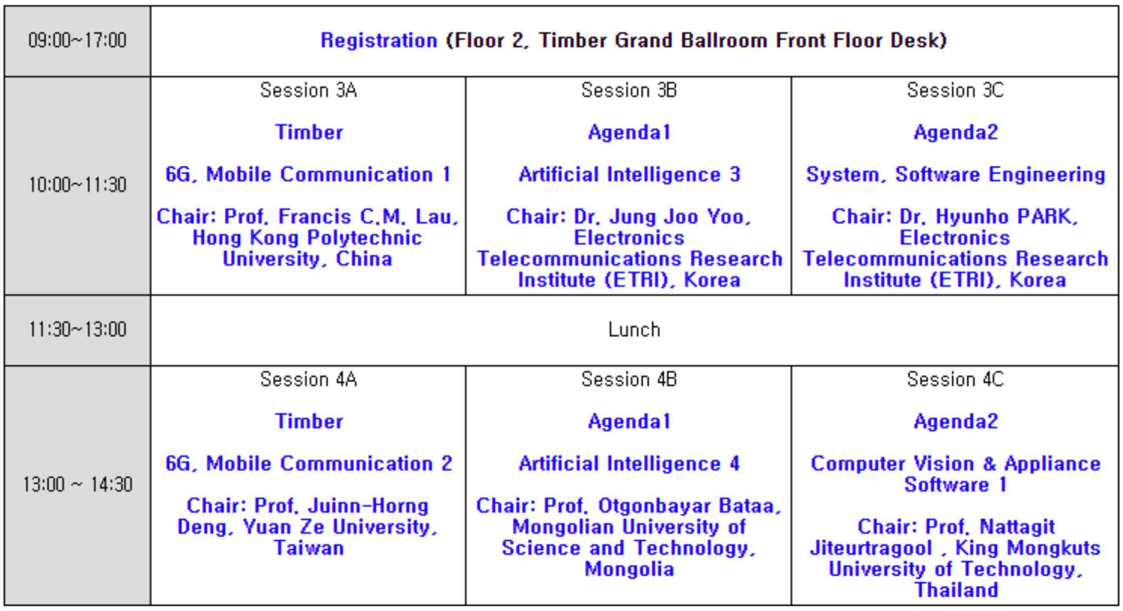
Feb. 5

09:00~17:00	Registration (Floor 2, Timber Grand Ballroom Front Floor Desk)		
10:00~11:30	Session 1A Timber Wireless Communication 1 Chair: Prof. Anuradha, National Institute of Technology, India	Session 1B Agenda 1 Artificial Intelligence 1 Chair: Prof. Jennifer Llovido, Bicol University, Philippines	Session 1C Agenda 2 Security & Blockchain 1 Chair: Dr. Pham Dinh Lam, Kyonggi University, Korea
11:30~13:00	Lunch Break		
13:00~14:30	Session 2A Timber Wireless Communication 2 Chair: Prof. Ming An Chung , National Taipei University of Technology, Taiwan	Session 2B Agenda 1 Artificial Intelligence 2 Chair: Prof. Bing-Yuh Lu, Guangdong University of Petrochemical Technology, China	Session 2C Agenda 2 Security & Blockchain 2 Chair: Prof. Michael Angelo Brogada , Bicol University, Philippines
14:30~16:00	Housekeeping Long Break		
16:00	Keynote Speech Session (Floor 2, Timber Grand Ballroom) Chair: Prof. Hyeong Ho Lee, SNU of Science and Technology, Korea		
16:00~16:45	Keynote Speaker I Prof. Kwanghoon Kim, PhD. President of KSII, Korean Society of Internet and Information, Korea		
16:45~17:30	Keynote Speaker II Prof. KWON YEONG-IL, PhD. Head of Graduate School of Management of Advanced Technology Hoseo University, Korea		



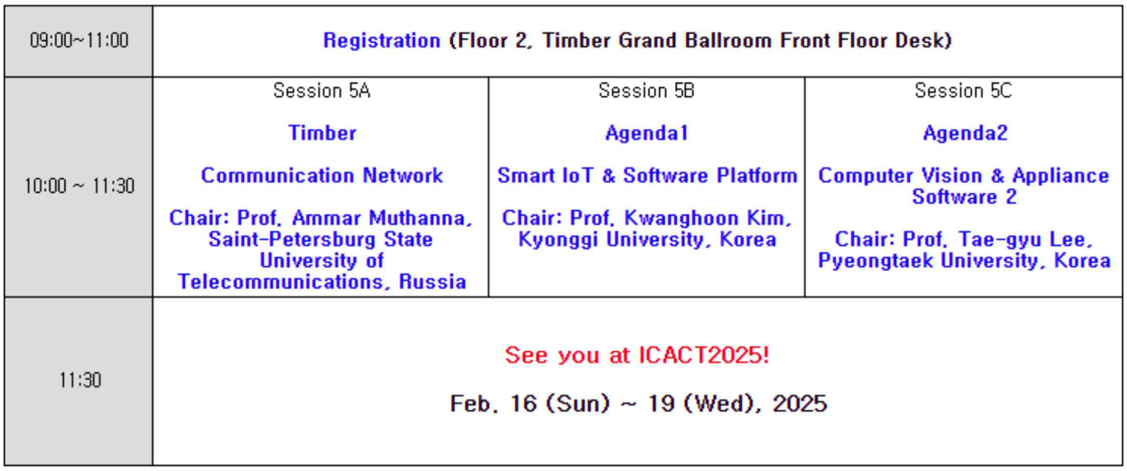
17:30	Plenary Session & Opening Ceremony (Floor 2, Timber Grand Ballroom) Moderator: Dr. Pham Dinh Lam
~	Welcome Address : General Chair Congratulatory Address : TPC Chair Plenary Agenda: OOC Chair
17:30~19:00	Grand Prize Award Ceremony : VVIP & Committee Chairs ※ Toast for Opening Banquet

Feb. 6



09:00~17:00	Registration (Floor 2, Timber Grand Ballroom Front Floor Desk)		
10:00~11:30	Session 3A Timber 6G, Mobile Communication 1 Chair: Prof. Francis C.M. Lau, Hong Kong Polytechnic University, China	Session 3B Agenda1 Artificial Intelligence 3 Chair: Dr. Jung Joo Yoo, Electronics Telecommunications Research Institute (ETRI), Korea	Session 3C Agenda2 System, Software Engineering Chair: Dr. Hyunho PARK, Electronics Telecommunications Research Institute (ETRI), Korea
11:30~13:00	Lunch		
13:00 ~ 14:30	Session 4A Timber 6G, Mobile Communication 2 Chair: Prof. Juinn-Horng Deng, Yuan Ze University, Taiwan	Session 4B Agenda1 Artificial Intelligence 4 Chair: Prof. Otgonbayar Bataa, Mongolian University of Science and Technology, Mongolia	Session 4C Agenda2 Computer Vision & Appliance Software 1 Chair: Prof. Nattagit Jiteurtragool , King Mongkuts University of Technology, Thailand

Feb. 7



09:00~11:00	Registration (Floor 2, Timber Grand Ballroom Front Floor Desk)		
10:00 ~ 11:30	Session 5A Timber Communication Network Chair: Prof. Ammar Muthanna, Saint-Petersburg State University of Telecommunications, Russia	Session 5B Agenda1 Smart IoT & Software Platform Chair: Prof. Kwanghoon Kim, Kyonggi University, Korea	Session 5C Agenda2 Computer Vision & Appliance Software 2 Chair: Prof. Tae-gyu Lee, Pyeongtaek University, Korea
11:30	See you at ICACT2025! Feb. 16 (Sun) ~ 19 (Wed), 2025		

Table of Contents
(Conference)

Session 1A: Wireless Communication 1

Chair: Prof. Anuradha, National Institute of Technology, India

1 Paper ID: 20240055, 1~4

Broadband On-Chip antenna array design in CMOS technology for D-band application

Prof. Ming-An Chung, Mr. Shang-Jui Huang, Mr. Yu-Hsun Chen,

National Taipei University of Technology, Taiwan

2 Paper ID: 20240052. 5~8

Chip Antenna with Vivaldi-Like Structure for W-Band Design Prof.

Ming-An Chung, Mr. Ming-Chun Hsieh, Mr. Kuo-Chun Tseng,

National Taipei University of Technology, Taiwan

3 Paper ID: 20240058, 9~12

Design of Ka-band Chip Antenna Based on Slot Antenna

Prof. Ming-An Chung, Mr. Kai-Xiang Chen, Mr. Kuo-Chun Tseng,

National Taipei University of Technology. Taiwan

4 Paper ID: 20240441, 13~18

Design of Calibration Algorithms for Fully-Activated Millimeter-Wave Phased Array Antennas

Prof. Juinn-Horng Deng, Mr. Xiang-He Huang, Dr. Chung-Lien Ho, Ms. Yu-Chien Wu,

Yuan Ze University. Taiwan

5 Paper ID: 20240381, 19~22

The Seamless Connection between Underwater and Terrestrial Communication for 6G

Dr. Tin-Yu Wu, Mr. Yi-Kai Chen, Mr. Fu-Jie Tey,

NPUST. Taiwan

Session 1B: Artificial Intelligence 1

Chair: Prof. Jennifer Llovido, Bicol University, Philippines

1 Paper ID: 20240217, 23~28

Test case prioritization with z-Score based neuron coverage

Ms. Hyekyoung Hwang, Prof. Jitae Shin,

Sungkyunkwan University. Korea(South)

2 Paper ID: 20240208, 29~32

Physics-Informed Neural Networks for solving Blood Flows

Prof. Yao-Chung Chang, Prof. Yu-Shan Lin, Dr. Jeu-Jiun Hu,

National Taitung University. Taiwan

3 Paper ID: 20240336, 33~37

Implementation of IoT-based Control System for Maintenance Operation of Long-distance Air Pollution Prevention Device RTO

Prof. DAL-HWAN YOON,

Semyung University. Korea(South)

4 Paper ID: 20240357, 38~42

Search and Recommendation Systems with Metadata Extensions

Mr. Woo-Hyeon Kim, Dr. Joo-Chang Kim,

Kyonggi University. Korea(South)

5 Paper ID: 20240342, 43~49

Utterance-Level Incongruity Learning Network for Multimodal Sarcasm Detection

Dr. Liujing Song, Dr. Zefang Zhao, Prof. Yuxiang Ma, Dr. Yuyang Liu, Prof. Jun Li,

Computer Network Information Center. China

6 Paper ID: 20240043, 50~52

Anomaly Detection During Additive Processes for DLP 3D Printing

Prof. Hyejin S. Kim,

ETRI. Korea(South)

Session 1C: Security & Blockchain 1

Chair: Dr. Pham Dinh Lam, Kyonggi University, Korea, ,

1 Paper ID : 20240396, 53~56

Generalized Parabola Chaotic map for Pseudorandom Random Number Generator

Dr. Nattagit Jiteurtragool,

King Mongkut's University of Technology North Bang. Thailand

2 Paper ID : 20240324, 57~62

SECURITY ANALYSIS OF ANDROID APPLICATIONS FOR HOTEL AND FLIGHT BOOKING APPLICATIONS

Ms. Vatcharavaree Wongsuna, Prof. Sudsanguan Ngamsuriyaroj,

Faculty of ICT, Mahidol University. Thailand

3 Paper ID : 20240309, 63~66

The development of new system for generating training data of AI-based anomaly detection

Ms. Thi My Truong, Dr. Won Seok Choi, Mr. Jang Hyeon Jeong, Prof. Seong Gon Choi,

Chungbuk National University. Korea(South)

4 Paper ID : 20240272, 67~72

Router Penetration Testing Based on CEM Vulnerability Assessment Criteria

Ms. Tai-Ying Chiu, Mr. Bor-Yao Tseng, Mr. Bagus ATMAJA, Prof. Jiann-Liang Chen,

National Taiwan University of Science & Technology. Taiwan

5 Paper ID : 20240269, 73~78

Hybrid Clustering Mechanisms for High-Efficiency Intrusion Prevention

Ms. Pin-Shan Lin, Mr. Yi-Cheng Lai, Ms. Man-Ling Liao, Ms. Shih-Ping Chiu, Prof. Jiann-Liang Chen,

National Taiwan University of Science & Technology. Taiwan

Session 2A: Wireless Communication 2

Chair: Prof. Ming An Chung , National Taipei University of Technology, Taiwan

1 Paper ID: 20240287, 79~83

Enhancing Inter-Satellite Data Relay in Dynamic Space Communication

Prof. REFIK CAGLAR KIZILIRMAK, Mr. Israel Ehile, Mr. Bekzat kabdrashev, Mr. Sergey Khvan,

Nazarbayev University. Kazakhstan

2 Paper ID: 20240473, 84~88

A Study on the evaluation of the ICT development indexes and some results

Ms. Narantuya Erkhembaatar, Prof. Otgonbayar Bataa,

SICT, MUST. Mongolia

3 Paper ID: 20240009, 89~93

Decoding Convolutional Hadamard Codes and Turbo Hadamard Codes using Recurrent Neural Networks

Dr. Sheng Jiang, Prof. Francis C.M. Lau,

The Hong Kong Polytechnic University. Hong Kong

4 Paper ID: 20240070, 94~99

QPSO-based Beamforming in Dual RIS-assisted Uplink Anti-jamming Communication System

Ms. Di Zhou, Prof. Zhiquan Bai, Ms. Jinqiu Zhao, Mr. Zeyu Liu, Prof. Dejie Ma, Prof. KyungSup Kwak,

Shandong University. China

5 Paper ID: 20240022, 100~103

A compact dual-band metamaterial absorber using square split rings for C-band and X-band sensors applications

Mr. Ramesh Amugothu, Prof. Vakula Damara,

NITW. India

Session 2B: Artificial Intelligence 2

Chair: Prof. Bing-Yuh Lu, Guangdong University of Petrochemical Technology, China

1 Paper ID: 20240037, 104~106

Optimizing Implementation of SNN for Embedded System

Dr. Hyeonguk Jang, Dr. Jae-Jin Lee, Dr. Kyuseung Han,

Electronics and Telecommunications Research Instit. Korea(South)

2 Paper ID: 20240049, 107~115

Multivariate PCA-based Composite Criteria Evaluation Method for Anomaly Detection in Manufacturing Data

Mr. HyeokSoo Lee, Mr. Youngki Jo, Prof. Jongpil Jeong,

Department of Smart Factory Convergence, Sungkyunkwan University, Korea(South)

3 Paper ID: 20240248, 116~121

Pitching-Motion: Pose-Based Pitch Trajectory Overlay System

Mr. Bor-Yao Tseng, Mr. Hung-Tse Chiang, Prof. Jiann-Liang Chen, Mr. Han-Chuan Hsieh,

National Taiwan University of Science & Technology. Taiwan

4 Paper ID: 20240469, 122~130

A Review of Detection-related Multiple Object Tracking in Recent Times

Ms. Suya Li, Ms. Ying Cao, Ms. Xin Xie,

Henan University. China

5 Paper ID: 20240475, 131~136

An Enhanced Topic Modeling Method in Educational Domain by Integrating LDA with Semantic

Mr. Ruofei Ding, Mr. Pucheng Huang , Ms. Shumin Chen, Mr. Jiale Zhang, Mr. Jingxiu Huang, Mr. Yunxiang Zheng,

South China Normal University. China

Session 2C: Security & Blockchain 2

Chair: Prof. Michael Angelo Brogada , Bicol University, Philippines

1 Paper ID: 20240459, 137~142

Intelligent Anomaly Detection System Based on Ensemble and Deep Learning

Dr. Babu Baniya, Mr. Thomas Rush,

Bradley University. USA

2 Paper ID: 20240064, 143~146

A Private Blockchain System based on Zero Trust Architecture

Prof. Yao-Chung Chang, Prof. Yu-Shan Lin, Dr. Hsin-Te Wu, Prof. Arun Kumar Sangaiah,

National Taitung University. Taiwan

3 Paper ID: 20240031, 147~151

Novel Design of Blockchain based IIoT Framework for Smart Factory

Ms. Ahyun Song, Prof. Euiseong Seo, Prof. Heeyoul Kim,

Sungkyunkwan University. Korea(South)

4 Paper ID: 20240467, 152~157

A Reference Implementation of Blockchain Interoperability Architecture Framework

Mr. Harish V, Ms. Swathi R, Mr. Satyanarayana N,

CDAC. India

5 Paper ID: 20240034, 158~162

Multiple Merged Structures Based on Image Recognition for Converting Application of Natural Language Artificial Intelligence Service

Prof. Hui-Yu Huang, Mr. Ming-Hsun Tsai, Dr. Ming Shen Jian,

National Formosa University. Taiwan

6 Paper ID: 20240079, 163~167

Pipeline Based Genetic Algorithm for Patient Scheduling in Hospital Outpatient Department and Laboratory

Dr. Ming Shen Jian, Mr. Cheng-He Wang, Mr. Wei-Siou Wu, Mr. Tzu-Wei Hunag,

National Formosa University. Taiwan

Session 3A: 6G, Mobile Communication 1

Chair: Prof. Francis C.M. Lau, Hong Kong Polytechnic University, China

1 Paper ID: 20240202, 168~ 173

Traffic Type Recognition in 6G Software-Defined Networking for Telepresence Services

Dr. Artem Volkov, Ms. Varvara Mineeva, Dr. Ammar Muthanna, Prof. Andrey Koucheryavy,

The Bonch-Bruevich Saint Petersburg State Universi. Russia

2 Paper ID: 20240130, 174~182

Microservice-Based Fog Testbed for 6G Applications

Ms. Ekaterina Kuzmina, Ms. Meriem Tefikova, Dr. Artem Volkov, Dr. Ammar Muthanna, Prof. Andrey Koucheryavy,

The Bonch-Bruevich Saint Petersburg State Universi. Russian Federation

3 Paper ID: 20240366, 183~186

Migration routing algorithm for microservice based Fog computing system

Ms. Ekaterina Kuzmina, Ms. Meriem Tefikova, Dr. Artem Volkov, Dr. Ammar Muthanna, Prof. Andrey Koucheryavy,

Department of Telecommunication Networks and Data . Russia

4 Paper ID: 20240241, 187~192

Dual-RIS Assisted 3D Positioning and Beamforming Design in ISAC System

Mr. Dejie Ma, Prof. Zhiquan Bai, Dr. Jinqiu Zhao, Mr. Hao Xu, Mr. Zeyu Liu, Dr. Di Zhou, Prof. Mingyan Jiang, Prof. KyungSup Kwak,

Shandong University. China

5 Paper ID: 20240375, 193~198

Deep Reinforcement Learning Based Beamforming in RIS-assisted MIMO System Under Hardware Loss

Mr. Yuan Sun, Mr. Zhiquan Bai, Ms. Jinqiu Zhao, Mr. Dejie Ma, Ms. Zhaoxia Xian, Mr. KyungSup Kwak,

Shandong University. China

Session 3B: Artificial Intelligence 3

Chair: Dr. Jung Joo Yoo, Electronics Telecommunications Research Institute (ETRI), Korea

1 Paper ID: 20240447, 199~205

Transforming Education Policy: Evaluating UAQTE Program Implementation through LDA, BoW and TF-IDF Techniques

Mr. Christian Sy, Dr. Lany Maceda, Dr. Thelma Palaoag, Dr. Mideth Abisado,

Bicol University. Philippines

2 Paper ID: 20240444, 206~210

Leveraging Machine Learning to Uncover Key Factors Influencing Satisfaction Among Free Tertiary Education Recipients in the Philippines

Prof. John Raymund Barajas, Dr. Lea Austero, Dr. Jennifer Llovido, Dr. Lany Maceda, Dr. Mideth Abisado,

Bicol University. Philippines

4 Paper ID: 20240402, 211~215

Classifying Gastric Cancer carcinoma stages with deep semantic features and GLCM Texture Features

Mr. Sikandar Ali, Ms. Samman Fatima, Mr. Ali Hussain, Mr. Maisam Ali , Mr. Muhammad Yaseen, Mr. Tagne Poupi Theodore Armand, Prof. Hee Cheol Kim,

Inje University. Korea(South)

5 Paper ID: 20240453, 216~220

Enhanced Experiences: Benefits of AI-powered recommendation systems

Ms. Kouayep Sonia Carole, Mr. Tagne Poupi Theodore Armand, Prof. Hee-Cheol Kim,

inje university. Korea(South)

Session 3C: System, Software Engineering

Chair: Dr. Hyunho PARK, Electronics Telecommunications Research Institute (ETRI), Korea

1 Paper ID: 20240330, 221~225

Evaluation of |Y> Magic State Distillation Circuit

Mr. Youngchul Kim, Dr. Soo-Cheol Oh, Ms. Sangmin Lee, Mr. Ki-Sung Jin, Mr. Gyuil Cha,

ETRI. Korea(South)

2 Paper ID: 20240293, 226~231

DB Workload Management through Characterization and Idleness Detection

Dr. Abdul Mateen, Mr. Khawaja Tahir Mahmood, Dr. Seung Yeob Nam,

Federal Urdu University of Arts, Science & Technol. Pakistan

3 Paper ID: 20240417, 232~240

Micro-services internal load balancing for Ultra Reliable Low Latency 5G Online charging system

Mr. Ngoc Tien Nguyen, Mr. Thanh Son Pham, Mr. Van Duong Nguyen, Dr. Cong Dan Pham, Mr. Duc Hai Nguyen,

Viettel High Technology. Viet Nam

4 Paper ID: 20240471, 241~247

AppTest: Assessing the Usability and Performance Efficiency of BOSESKO for Digital Participation

Dr. Jennifer Llovido, Dr. Michael Angelo Brogada, Dr. Lany Maceda, Dr. Mideth Abisado,

Bicol University. Philippines

5 Paper ID: 20240456, 248~251

Research on the transformation path of DevOps in the digital era

Ms. XIAOLING NIU, Ms. LINGLING YANG, Ms. KAILING LIU, Mr. ZHAOWEI LIU,

The China Academy of Information and Communication. China

Session 4A: 6G, Mobile Communication 2

Chair: Prof. Juinn-Horng Deng, Yuan Ze University, Taiwan

1 Paper ID: 20240470, 252~262

An Efficient Resource Allocation Algorithm for Traffic of a Content Streaming in Non-standalone OFDM-Based 5G NR

Ms. Narantuya VANDANTSEREN ,

Mongolian Defense University. Mongolia

2 Paper ID: 20240387, 263~270

Design of Communication Countermeasure Simulation Model and Data Interaction Interface for Battlefield Network Based on QualNet

Mr. Wenyi Li, Prof. Peng Gong, Mr. Weidong Wang, Mr. Yu Liu, Mr. Jianfeng Li, Dr. Xiang Gao,

Beijing Institute of Technology. China

3 Paper ID:20240480, 271~274

Location based Data-centric Forwarding for Mobile Ad-hoc Networks

Mr. Hieu Nguyen, Prof. Ilkyeun Ra,

University of Colorado Denver. USA

4 Paper ID: 20240369, 275~280

Optimization of Downlink Power Allocation in NOMA-OTFS based Cross-Domain Vehicular Networks

Mr. Hao Xu, Mr. Zhiquan Bai, Ms. Jinqiu Zhao, Mr. Dejie Ma, Mr. Bangwei He, Mr. KyungSup Kwak,

Shandong University. China

5 Paper ID: 20240040, 281~285

Performance Evaluation of UAV-based NOMA for 5G and Beyond

Ms. MOUNIKA NEELAM, Prof. Anuradha Sundru,

NITW. India

Session 4B: Artificial Intelligence 4

Chair: Prof. Otgonbayar Bataa, Mongolian University of Science and Technology, Mongolia

1 Paper ID: 20240477, 286~289

Multi-Class Document Classification using LayoutLMv1 and V2

Ms. Kounen Fathima, Mr. Athar Ali, Prof. Hee Cheol Kim,

Inje University. Korea(South)

2 Paper ID: 20240384, 290~294

Machine learning based techniques for the Prediction of axillary lymph node metastases in early breast cancer

Mr. Maisam Ali, Mr. Muhammad Yaseen, Mr. Sikandar Ali, Prof. Hee-Cheol Kim,

Inje University. Korea(South)

3 Paper ID: 20240479, 295~299

AI-based logistics system overview and a conceptual framework for digital freight forwarding in logistics

Mr. Md Ariful Islam Mozumder, Mr. Rashedul Islam Sumon, Mr. Ziaullah Khan, Mr. Shah Muhammad Imtiyaj Uddin, Mr. Muhammad Omair Khan, Prof. Hee-Cheol Kim,

Inje University. Korea(South)

4 Paper ID: 20240411, 300~304

The benefits of integrating AI, IoT, and Blockchain in healthcare supply chain management: A multi-dimensional analysis with case study

Mr. Tagne Poupi Theodore Armand, Ms. Kouayep Sonia Carole, Mr. Subrata Bhattacharjee, Mr. Md Ariful Islam Mozumder, Dr. Austin Oguejiofor Amaechi, Prof. Hee-Cheol Kim,

Inje University. Korea(South)

5 Paper ID: 20240321, 305~310

Knowledge-Prompted Estimator:A Novel Approach to Explainable Machine Translation Assessment

Dr. hao yang,

huawei. China

6 Paper ID: 20240393, 311~314

Integration of a Chatbot to facilitate access to educational content in digital universities

Mr. Birahim BABOU, Dr. Khalifa SYLLA, Mr. Mouhamadou Yaya Sow, Prof. Samuel OUYA,

UCAD. Senegal

Session 4C: Computer Vision & Appliance Software 1

Chair: Prof. Nattagit Jiteurtragool , King Mongkuts University of Technology, Thailand

1 Paper ID: 20240312, 315~320

Explainable Rip Current Detection and Visualization with XAI EigenCAM

Mr. Juno Choi, Mr. Muralidharan Rajendran, Mr. Yong Cheol Suh,

Pukyong National University. Korea(South)

2 Paper ID: 20240076, 321~326

A Test Method for the Convergence of the Metaverse and Blockchain

Prof. Tae-gyu Lee,

Pyeongtaek University. Korea(South)

3 Paper ID: 20240429, 327~331

Time-frequency Analysis for Validating Prognostics Algorithms of Rolling Element Bearings

Dr. Guanhua Zhu, Dr. Xiaoling Xu, Mr. Qing Zhong , Dr. Bing-Yuh Lu, Mr. Yushen Lu, Mr. Guangming Xu, Mr. Yumeng Zhou, Mr. Ziyi Jiang, Mr. Kai Sun, Mr. Minhao Wang,

Guangdong University of Petrochemical Technology. China

4 Paper ID: 20240435, 332~336

Evaluation System for Dancing Enlightenment Posture Training Using the Skeleton Tracking of Microsoft Common Objects in Context

Mr. Ruilong Huang, Ms. Huifang Deng, Prof. Ruei-Yuan Wang, Prof. Bing-Yuh Lu, Prof. Hongwei Ren, Mr. Yiheng Chen, Mr. Jianwen Ye, Mr. Jinhui Chen, Mr. Yingbo Jia, Ms. Leyang Lang,

Guangdong University of Petrochemical Technology. China

5 Paper ID: 20240061, 337~342

Computer Vision-Based Structural Deformation Monitoring System on Android Smartphones: Design and Implementation

Ms. Xiang DONG, Mr. Maokai LAI, Ms. Hui LIANG, Mr. Peng WU, Ms. Chaoxia WANG, Mr. Ting PENG,

Chang'an University. China

Session 5A: Communication Network

Chair: Prof. Ammar Muthanna, Saint-Petersburg State University of Telecommunications, Russia

1 Paper ID: 20240199, 343~349

Utilizing Machine Learning for Sensor Fault Detection in Wireless Sensor Networks

Mr. Abubakar Abdulkarim, Mr. Israel Ehile, Prof. Refik Caglar Kizilirmak,

Nazarbayev University. Kazakhstan

2 Paper ID: 20240414, 350~353

Multicore Packet Distribution method using Multicore Network Interface Card based on Tile-gx72 Network Processor

Dr. Choi Won Seok, Mr. Lee Sang Ju , Mr. Kim Jong Oh, Prof. Choi Seong Gon,

Chungbuk National Univertisy. Korea(South)

3 Paper ID: 20240426, 354~359

Flexible Localization Method with Motion Estimation for Underwater Wireless Sensor Networks

Dr. Abdelrahman Samy, Prof. Ammar Hawbani, Prof. Xingfu Wang, Dr. Samah Abdel Aziz, Prof. Liang Zhao, Prof. Nasir Saeed,

University of Science and Technology of China. China

4 Paper ID: 20240378, 360~364

A reliable routing method for remote entanglement distribution under limited resources

Ms. Tianzhu Hu,

USTC. China

5 Paper ID: 20240067, 365~368

Energy Efficiency Analysis of novel Index Modulation-based Non-Orthogonal Multiple Access (IMNOMA) system for 5G Networks

Ms. Shwetha H M, Prof. Anuradha Sundru,

Department of Electronics & communication Engineer. India

Session 5B: Smart IoT & Software Platform

Chair: Prof. Kwanghoon Kim, Kyonggi University, Korea, ,

1 Paper ID: 20240333, 369~372

The Data Alignment Method between GPS and IMU based on ICP for Indoor Positioning

Mr. Yong Hee Park, Mr. Min Gu Kang, Mr. Jang Hyeon Jeong, Prof. Seong Gon Choi,

Pabat. Korea(South)

2 Paper ID: 20240299, 373~375

Incorporation of waypoint following logic into ROS publish and subscribe mechanism

Mr. Minhwa Hong, Prof. Seonggon Choi, Dr. Heonjong Yoo,

Chungbuk National University. Korea(South)

3 Paper ID: 20240438, 376~383

Development and Implementation of BOSESKO: A Synoptic Multi-platform Digital Citizen Participatory System

Dr. Jennifer Llovido, Dr. Michael Angelo Brogada, Dr. Floradel Relucio, Dr. Lea Austero, Dr. Lany Maceda, Dr. Mideth Abisado,

Bicol University. Philippines

4 Paper ID: 20240348, 384~388

An IoT-Based Early Warning System for Settlement Monitoring Using Differential Pressure Static Level

Mr. Tieyan CHAO, Ms. Hui LIANG, Mr. Yuwei GE, Mr. Kai HOU, Ms. Xiang DONG, Mr. Ting PENG,

Shaanxi Huashan Road and Bridge Group Co., LTD. China

5 Paper ID: 20240073, 389~394

Tunnel Construction Site Monitoring and Digital Twin System

Mr. Wei CHENG, Mr. Yuxing PAN, Mr. Zhi MA, Mr. Yincai CAI, Dr. Yuan LI, Prof. Ting PENG,

Sichuan Chuanjiao Road and Bridge Co., LTD. China

6 Paper ID: 20240339, 395~401

Research on LSTM-based Model for Predicting Deformation of Tunnel Section During Construction Period

Mr. Jiwen ZHANG, Mr. Kai YUAN, Mr. Jianjun MAO, Mr. Yincai CAI, Mr. Dongfeng LEI, Mr. Jinyang DENG, Mr. Ting PENG,

Sichuan Chuanjiao Road and Bridge Co., LTD. China

Session 5C: Computer Vision & Appliance Software 2

Chair: Prof. Tae-gyu Lee, Pyeongtaek University, Korea, ,

1 Paper ID: 20240405, 402~407

Leveraging Deep Learning for Automated Analysis of Colorectal Cancer Histology Images to Elevate Diagnosis Precision

Mr. Shah Muhammad Imtiyaj Uddin, Mr. Md Ariful Islam Mozumder, Mr. Rashedul Islam Sumon, Prof. Joo Moon-il, Prof. Hee-Cheol Kim,

Inje University. Korea(South)

2 Paper ID: 20240327, 408~412

Deep Learning Based Cervical Spine Bones Detection: A case study using YOLO

Mr. Muhammad Yaseen , Mr. Maisam Ali, Mr. Sikandar Ali, Mr. Ali Hussain, Mr. Ali Athar, Prof. Hee-Cheol Kim,

Inje University. Korea(South)

3 Paper ID: 20240345, 413~418

Vision transformer-based model for gastric cancer detection and classification using weakly annotated histopathological images

Mr. Tagne Poupi Theodore Armand, Mr. Subrata Bhattacharjee, Mr. Hyun-Joong Kim, Mr. Ali Hussain, Mr. Sikandar Ali, Mr. Heung-Kook Choi, Dr. Hee-Cheol Kim,

Inje University. Korea(South)

4 Paper ID: 20240372, 419~425

Overview of the potentials of multiple instance learning in cancer diagnosis: Applications, challenges, and future directions

Mr. Tagne Poupi Theodore Armand, Mr. Subrata Bhattacharjee, Prof. Hee-Cheol Kim,

Inje University. Korea(South)

5 Paper ID: 20240482, 426~432

A Study on Real-time Evaluation of Uncertainty of PM-10 Concentration Determined by Tele-measuring Instrument

Mr. Jeeho Kim, Dr. Jin-Chun Woo, Prof. Young Sunwoo,

Konkuk University. Korea(South)

Session 1A: Wireless Communication 1

Chair: Prof. Anuradha, National Institute of Technology, India

Broadband On-Chip antenna array design in CMOS technology for D-band application

Ming-An Chung*
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
mingannchung@ntut.edu.tw

Shang-Jui Huang
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
eric24588@gmail.com

Yu-Hsun Chen
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
t111368143@ntut.org.tw

***Abstract*— A patch array antenna working in the D-band is proposed in TSMC's 180 nm CMOS. The antenna structure applies a top metal layer (Metal6) as the patch antenna and a bottom metal layer (Metal1) for the ground. Simulation results demonstrate that the on-chip patch antenna has a peak gain of -5.95 dBi at 141 GHz. The measurement of the proposed antenna demonstrates bandwidth from 140 GHz to 202.5 GHz. The array on-chip patch antenna proposed can use in Smart cities technology and automotive applications.**

***Keywords*— D-band, CMOS, Antenna-on-chip, Broadband**

I. Introduction

Higher data transmission requirements have been realized in the 5G mmWave frequency band, such as 28 GHz and 40 GHz. This frequency band has a sub-span size of 800 MHz and a maximum transmission rate of 10 Gbit/s.With the increasing demand for high-speed transmission, greater transmission rates and bandwidth are required to meet the demand. For example, antennas operating in the D-band (110-170 GHz) frequency range greater than 100 GHz can provide highly aggregated bandwidth, enabling transmission rates up to 100 Gbit/s[1, 2].

The wavelength in the D-band frequency band is shorter, so the antenna can be applied to the chip at a relatively low price. The process relationship makes the gain of the chip antenna poor, which may lead to insufficient received signal strength. To avoid the problem of insufficient antenna gain, it is necessary to increase the transmit power to solve the negative impact of the communication system[3, 4].

Recently, there has been lots of research for mmWave chip antenna. The reference [5] utilized a patch antenna array and a planar monopole antenna. The planar monopole antenna incorporated multiple membrane supports and a polymer cavity. Under the patch array antenna configuration, at 130 GHz has an 8.66 dBi maximum gain and 126.5 to 138 GHz bandwidth. To raise the radiation effect of the slot antenna, a cavity is introduced, which can improve the gain to a maximum value of -2 dBi [6]. In the 130 GHz meandering slot antenna, a stacked dielectric resonator (DR) can improve the antenna's high efficiency and gain. Experimental results show that 130 GHz has a 43% radiation efficiency and a 4.7 dBi gain [7]. The dual-frequency helical monopole chip antenna utilizes the hybrid structure of the helix and the package substrate, and the linear monopole antenna can raise the antenna radiation's high efficiency and obtain a dual-frequency band. The antenna gets an efficiency of 45% at 28 GHz and 30% at 60 GHz [8]. A square ring resonator (SRR) can also be used to address the gain drop problem. Experimental results show that a gain reduction of about 7 dBi can be compensated at 85 GHz; at 81.5 GHz, a maximum gain of 1.61 dBi can be compensated [9]. The experimental results indicate that approximately 7 dBi of gain reduction can be compensated for at 85 GHz, while 1.61 dBi, a maximum gain, can be compensated for at 81.5 GHz.

This paper proposes a broadband on-chip array antenna in the D-band frequency. The antenna is implemented using CMOS technology. The simulation results demonstrate that at 141 GHz, the antenna can obtain a peak gain at -5.95 dBi. The bandwidth measured below the -10 dB standard ranges from 140 to 202.5 GHz. It can be applied in Industrial, healthcare, and automotive application [10].

II. Antenna Design

This paper introduces the CMOS broadband on-chip antenna in D-band implemented. Figure 1 shows the structure of the on-chip patch array antenna using the top metal layer (Metal6) with a thickness of 2.34 μm as the antenna and the bottom metal layer (Metal1) with a thickness of 0.53 μm as the ground. The size of the patch array antenna is 1200 × 1200 μm². Figure 2 and Table 1, illustrate the antenna geometry. Preliminary simulation results indicate that the chip antenna exhibits two frequencies resonant at 141 GHz and 180 GHz. In Figure 3, The lowest points of the reflection coefficient are -34.3 dB and -15 dB, respectively. The gain, depicted in Figure 4, reaches its peak at 137 GHz with a value of -5.41 dBi.

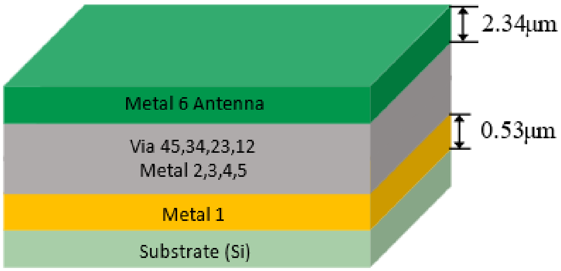


Figure 1. Process stack diagram

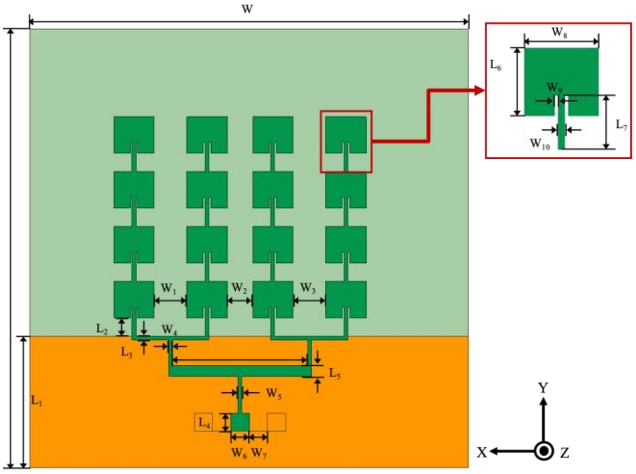


Figure 2. Antenna geometry

TABLE 1. SUMMARY OF GEOMETRIC DIMENSIONS

Unit: μm			
Parameter	**Value**	**Parameter**	**Value**
W	1200	L	1200
W1	90	L1	360
W2	70	L2	60
W3	90	L3	10
W4	10	L4	50
W5	10	L5	30
W6	50	L6	100
W7	50	L7	80
W8	110		
W9	5		
W10	10		

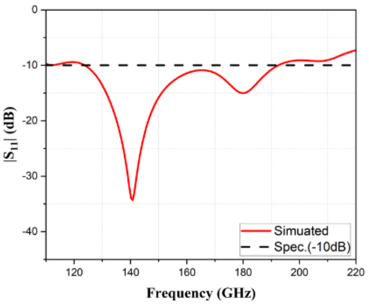


Figure 3. Reflection coefficient (S11)

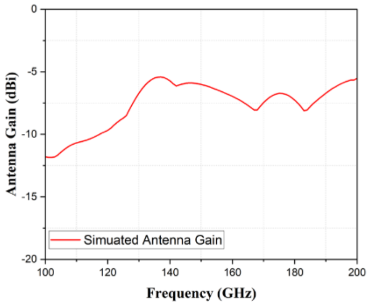


Figure 4. Gain plot

III. SIMULATION AND MEASUREMENT RESULTS

The prototype of a broadband on-chip antenna fabricated by the CMOS technology. The proposed antenna was designed for D-band operation. To prevent metal layers from interfering with each other and reduce losses, the antenna structure uses a top metal layer (Metal6) and a bottom metal layer (Metal1). The size of the chip antenna is 1200 ×1200μm². Figure 5 shows a micrograph of the chip antenna.

The measurement environment adopts the following equipment settings: Keysight PNA-X N5242B (10 MHz~26.5 GHz) vector network analyzer, and G-band frequency extension module (140 GHz~220 GHz), as shown in Figure 6.

During measurement, the chip is placed on the signal source system platform with absorbing material and fed using GSG probes. In Figure 7, According to the simulation, that can be observed that the chip antenna exhibits two frequencies resonant at 141 GHz and 180 GHz. The bandwidth simulated is 43.2%, with a range of 124 GHz to 192.4 GHz. The measured bandwidth is 36.5%, ranging from 140 to 202.5 GHz. Since the instrument can only measure 140 - 220GHz, measurement results below 140GHz cannot be obtained. Based on the trend of measurement results from 140 - 220GHz, it is speculated that the results below 140GHz should be similar to the simulation results.

Figure 8 and Figure 9 show the 2D antenna radiation pattern planes of xz and yz at 141 GHz. The antenna can be observed that the peak gain in the xz plane is -6.14 dBi. The peak gain is -5.95 dBi in the yz plane.

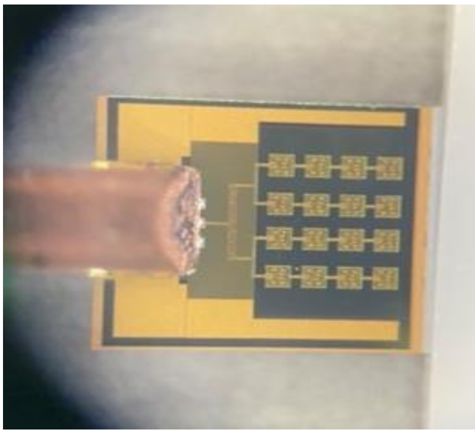


Figure 5. Chip photograph

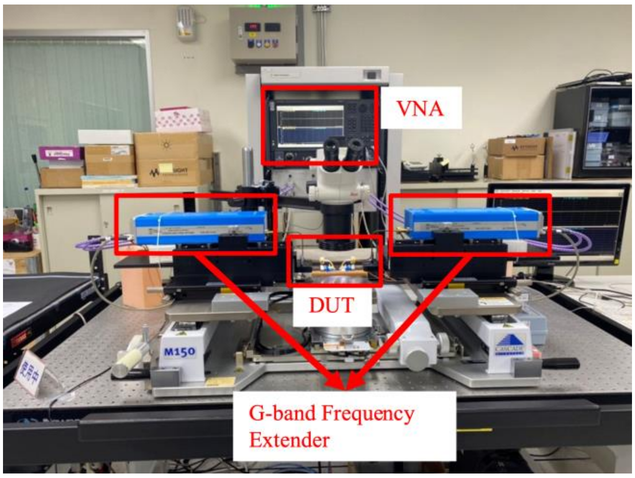


Figure 6. Measurement equipment diagram

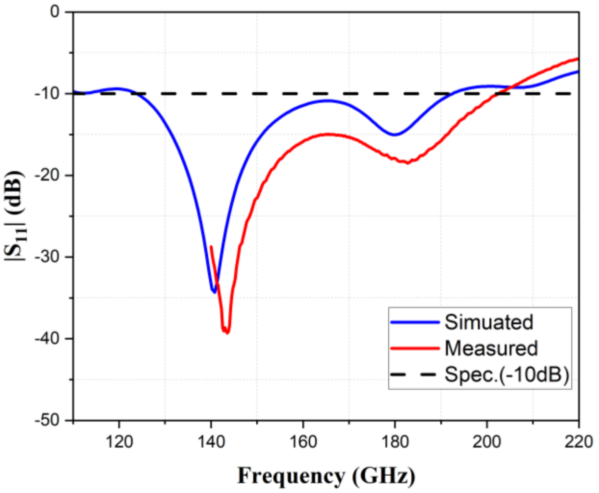


Figure 7. Simulation and measurement overlay plot

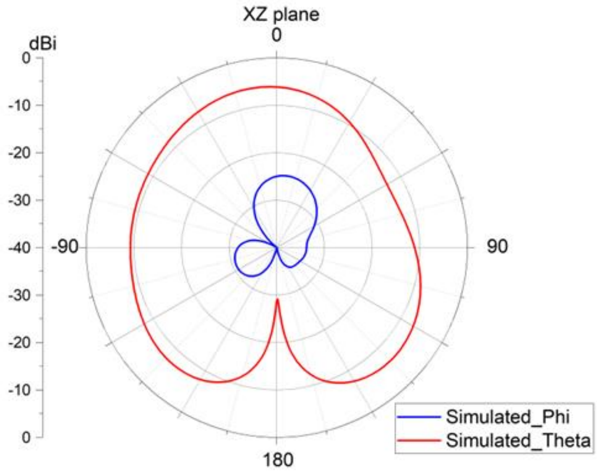


Figure 8. XZ plane

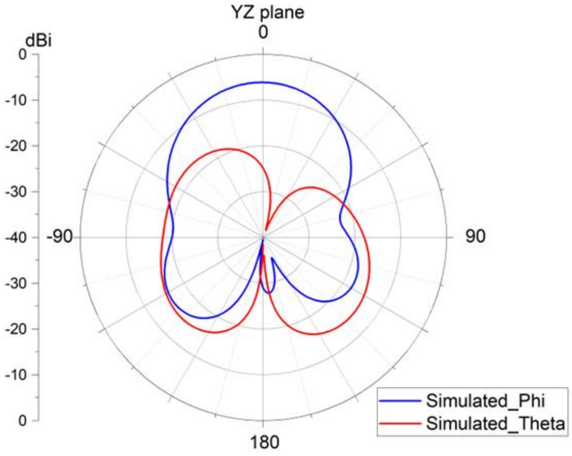


Figure 9. YZ plane

IV. Conclusions

This study proposed a broadband on-chip antenna array that functions within the D-band. The antenna is implemented by the 180 nm CMOS process of TSMC. The simulation result shows that 141 GHz has a -5.95 dBi maximum gain, with a bandwidth percentage of 43.2% ranging from 124 GHz to 192.4 GHz. The measured bandwidth is 36.5%, range of bandwidth from 140 GHz to 202.5 GHz. The proposed antenna can be applied in Smart cities, smart homes, electronic healthcare, intelligent transportation, and smart factory[11-13].

References

[1] T. Maiwald et al., "A broadband zero-IF down-conversion mixer in 130 nm SiGe BiCMOS for beyond 5G communication systems in D-Band," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 68, no. 7, pp. 2277-2281, 2021.

[2] T. Maiwald et al., "A Review of Integrated Systems and Components for 6G Wireless Communication in the D-Band," Proceedings of the IEEE, 2023.

[3] W. A. Ahmad, M. Kucharski, A. Di Serio, H. J. Ng, C. Waldschmidt, and D. Kissinger, "Planar highly efficient high-gain 165 GHz on-chip antennas for integrated radar sensors," IEEE Antennas and Wireless Propagation Letters, vol. 18, no. 11, pp. 2429-2433, 2019.

[4] M. Kucharski, W. A. Ahmad, H. J. Ng, and D. Kissinger, "Monostatic and bistatic G-band BiCMOS radar transceivers with on-chip antennas and tunable TX-to-RX leakage cancellation," IEEE Journal of Solid-State Circuits, vol. 56, no. 3, pp. 899-913, 2020.

[5] H. Chu, Y.-X. Guo, T.-G. Lim, Y. M. Khoo, and X. Shi, "135-GHz micromachined on-chip antenna and antenna array," IEEE transactions on antennas and propagation, vol. 60, no. 10, pp. 4582-4588, 2012.

[6] S. Pan and F. Capolino, "Design of a CMOS on-chip slot antenna with extremely flat cavity at 140 GHz," IEEE Antennas and Wireless Propagation Letters, vol. 10, pp. 827-830, 2011.

[7] D. Hou, Y.-Z. Xiong, W.-L. Goh, S. Hu, W. Hong, and M. Madihian, "130-GHz on-chip meander slot antennas with stacked dielectric resonators in standard CMOS technology," IEEE Transactions on Antennas and Propagation, vol. 60, no. 9, pp. 4102-4109, 2012.

[8] P. Burasa, T. Djerafi, and K. Wu, "A 28 GHz and 60 GHz dual-band on-chip antenna for 5G-compatible IoT-served sensors in standard CMOS process," IEEE Transactions on Antennas and Propagation, vol. 69, no. 5, pp. 2940-2945, 2020.

[9] C. Mustacchio, L. Boccia, E. Arnieri, and G. Amendola, "A gain levelling technique for on-chip antennas based on split-ring resonators," IEEE Access, vol. 9, pp. 90750-90756, 2021.

[10] M. de Kok, A. B. Smolders, and U. Johannsen, "A review of design and integration technologies for D-band antennas," IEEE Open Journal of Antennas and Propagation, vol. 2, pp. 746-758, 2021.

[11] A. Moglia et al., "5G in healthcare: from COVID-19 to future challenges," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 8, pp. 4187-4196, 2022.

[12] L. Chettri and R. Bera, "A comprehensive survey on Internet of Things (IoT) toward 5G wireless systems," IEEE Internet of Things Journal, vol. 7, no. 1, pp. 16-32, 2019.

[13] E. A. Oyekanlu et al., "A review of recent advances in automated guided vehicle technologies: Integration challenges and research areas for 5G-based smart manufacturing applications," IEEE access, vol. 8, pp. 202312-202353, 2020.

MING-AN CHUNG (Member, IEEE) received the B.Eng. and M.Eng. degrees in electronic engineering from the Chang Gung University, Taoyuan, Taiwan and the D.Eng. degree in electrical engineering from the National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan, in 2003, 2005, and 2016, respectively. He is currently an Associate Professor with the Department of Electronic Engineering, National Taipei University of Technology (NTUT), where he also serves as the Leader of the Innovation Wireless Communication and Electromagnetic Applications Laboratory. His research interests include wireless communication propagation, intelligent robotics, self-driving vehicles, antenna design for various mobile and wireless communications, electromagnetic theory, and applications. He is also a Reviewer of many scientific journals, including the IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, IEEE Transactions on Industrial Informatics, Journal of Intelligent & Robotic Systems, IET Microwaves, Antennas and Propagation, IEEE Antennas and Wireless Propagation Letters, International Review of Electrical Engineering, International Journal on Communications Antenna and Propagation and AEÜ - International Journal of Electronics and Communications, and many international conferences, including ICRA, ICCE-TW, RFIT, ICBEB, EMCAR and SNSP.

Shang-Jui Huang (HUANG, SHANG-JUI) received his Bachelor's degree in Electronic Engineering from National Taipei University of Technology (NTUT). He has a rich and diverse professional background, having previously worked at Inventec, where he gained valuable industry experience. Currently, he is pursuing a Master's degree at Taipei Tech with a focus on antenna technology, along with a strong interest in AI, edge devices, the Internet of Things (IoT), and power supply design.

Yu-Hsun Chen received the B.S. degree in electrical engineering from the Tunghai University, Taichung, Taiwan, in 2022. His current research interests include the design of phased array of antennas and microwave couplers for the applications of the fifth generation of mobile communications.

Chip Antenna with Vivaldi-Like Structure for W-Band Design

Ming-An Chung*
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
mingannchung@ntut.edu.tw

Ming-Chun Hsieh
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
t112368523@ntut.edu.tw

Kuo-Chun Tseng
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
t110368157@ntut.org.tw

***Abstract*— This paper proposes an application for a W-band chip antenna using the CMOS 180nm process. The chip size is 800 µm×1200 µm, and it employs by Vivaldi-like antenna for design and performance analysis. The bandwidth of simulation is 99.3 - 114.8 GHz. The result of measurement shows that the reflection coefficient (S11) is 91.1 - 107.1 GHz. The antenna gain is simulated by high-frequency electromagnetic simulation software, which is about -5 dBi to -4 dBi, and the deepest resonance point is at 108 GHz, showing a gain of -4.6 dBi.**

***Keywords*— Chip antenna, millimeter-wave, CMOS, fifth generation, W-band**

I. Introduction

In current 5G communication technology, higher transmission speeds, lower latency, and increased network capacity are achieved primarily through the addition of frequency bands and enhanced spectral efficiency. The millimeter-wave frequency range, known for its wide bandwidth, offers excellent data transmission speeds, making it a crucial tool in the field of high-speed wireless communication [1]. Millimeter-wave antennas are suitable for miniaturized solutions due to their high frequency, allowing for a significant reduction in antenna size [2, 3]. Chip antennas, which are integrated into chips, offer a viable solution for miniaturization. They can be applied to various small electronic devices such as smartphones, wearable devices, and IoT devices [4, 5].

In CMOS processes, the distance between metal layers is typically short, which may result in lower gain and radiation efficiency. The fabrication process is also influenced by the silicon substrate, leading to interference and inefficient propagation of antenna radiation, thus causing a decrease in antenna gain [6, 7]. Previous literature has addressed these issues and proposed solutions. These techniques include using meander-line monopole antennas combined with packaging substrates to improve radiation efficiency and gain performance [8]. Another approach involves integrating square Split Ring Resonators (SRR) as parasitic couplers with the feed antenna to increase gain and improve impedance matching [9]. In reference [10], cavity-backed stacked patch antennas were added to chip antennas to improve radiation performance, frequency range and minimize loss. There are also studies that incorporate metallic posts and embedded guiding structures to reduce unnecessary losses [11].

The antenna proposed in this paper adopts Vivaldi-like antenna. The V-shaped slot design of the Vivaldi antenna provides wide bandwidth and is directional. However, implementing a true V-shaped slot in a chip antenna faces limitations in manufacturing and size. A rectangular slot is used as an alternative solution to overcome these issues. The S11 of the proposed chip antenna is lower than the -10 dB standard, covering 99.3 to 114.8 GHz. The bandwidth percentage for this range is 14.08%. The antenna achieves a gain of approximately -4.5 dBi in this frequency range. The measured data shows a bandwidth ranging from 91.1 - 107.1 GHz, with a bandwidth percentage of 16.14%.

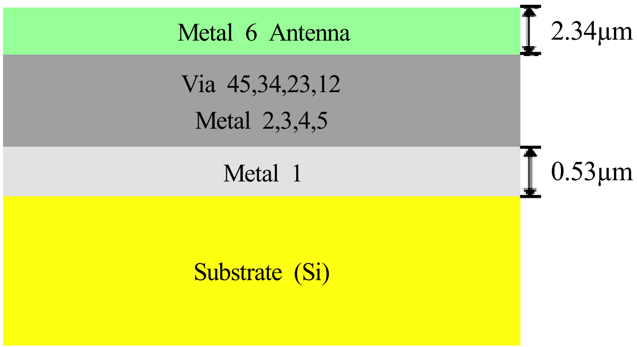


Figure 1. CMOS process stack diagram

II. Antenna Design

The proposed chip antenna is developed using CMOS process. Fig. 1 shows the fabrication process, which includes six metallization layers. The thicknesses of metal 6 and metal 1 are 2.34 µm and 0.8 µm. Due to the dielectric constant and resistivity of the substrate, it is challenging to achieve high-gain chip antennas compared to other antennas [12, 13]. Therefore, the antenna is constructed by connecting Metal 6 and Metal 1, which helps reduce losses.

Fig. 2 and Table I illustrate the structure and parameters of the chip antenna, with a chip size of 800 µm×1200 µm. In the antenna design, to mitigate the issues of low gain and efficiency caused by the CMOS process, a structure Vivaldi-like antenna is employed, using a rectangular slot as an approximation of the V-shaped slot. The Vivaldi antenna structure can provide higher gain and directivity, enabling a longer communication range. Additionally, it can reduce scattering and effectively suppress interference sources [14]. Due to process density limitations, slot cutting was performed on the Metal 6 layer of the antenna. After simulation testing, it has been verified that there is almost no difference between the results with and without the slot cutting.

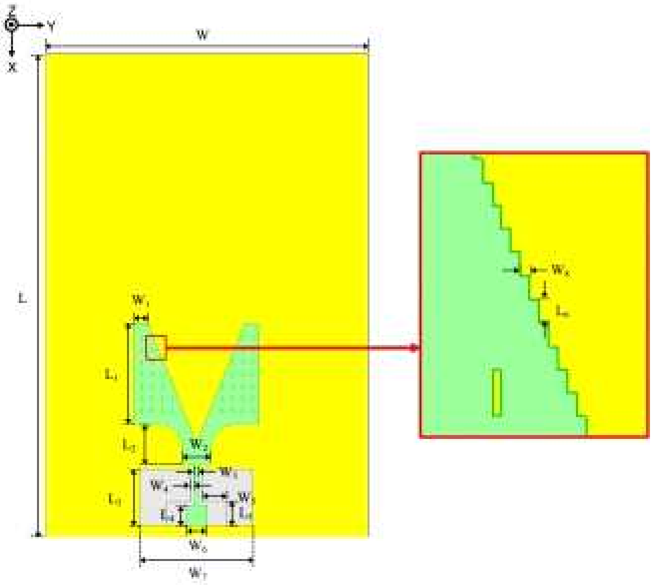


Figure 2. Geometric diagram of the Vivaldi-like antenna.

TABLE 1. ANTENNA ARCHITECTURE PARAMETERS

Parameter	**Value(μm)**	**Parameter**	**Value(μm)**
W	800	W8	2
W1	37	L	1200
W2	70	L1	250
W3	10	L2	100
W4	10	L3	140
W5	60	L4	50
W6	50	L5	60
W7	280	L6	5

III.RESULT

Figure 3 illustrates the actual chip layout, which is designed using the Taiwan Semiconductor Manufacturing Company 180 nm CMOS process. The measurements were conducted at the Taiwan Semiconductor Research Institute (TSRI) using the Agilent E7350A vector network analyzer (VNA). The experimental setup and test environment are depicted in Fig 4. During the testing, the chip was placed on the signal source system platform. GSG probes with a spacing of 100 μm were used to make connections, and the chip was connected to the VNA for measurement.

The simulated reflection coefficient is below -10 dB standard for the 99.3 - 114.8 GHz frequency range, and reaches the lowest resonance point at 108 GHz with a reflection coefficient of -30.6 dB. The measured data shows a bandwidth between 91.1 GHz and 107.1 GHz, with the lowest resonance point at 100 GHz and a reflection coefficient of -22.1 dB. Due to the limitations of the VNA used, which has a measure frequency range from 2 - 110 GHz, it was not possible to measure beyond 110 GHz and therefore it is not shown in the graph. The measured lowest resonance point shows better results compared to the simulation. However, due to fabrication limitations, there is a frequency offset observed in the measured results towards lower frequencies. Nonetheless, the overall trend of the measured results is similar to the simulated values. Therefore, it can be inferred that the measured results beyond 150 GHz would also align with the simulated data. The antenna exhibits a gain of approximately -5 dBi to -4 dBi. At the resonance point of 108 GHz, the gain is -4.6 dBi, as shown in Fig 5.

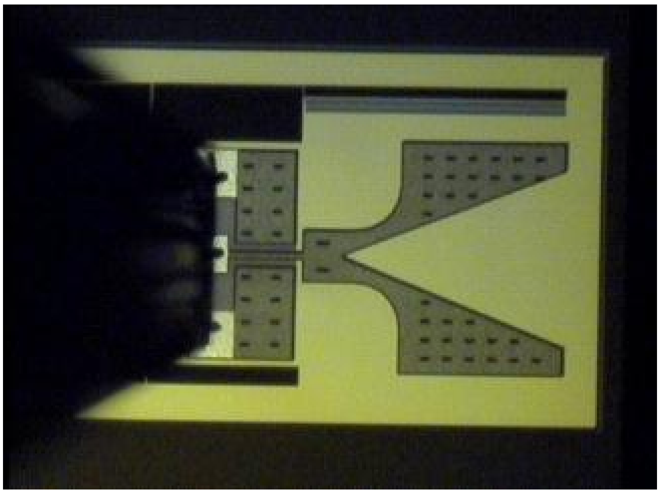


Figure 3. Proposed antenna chip diagram.

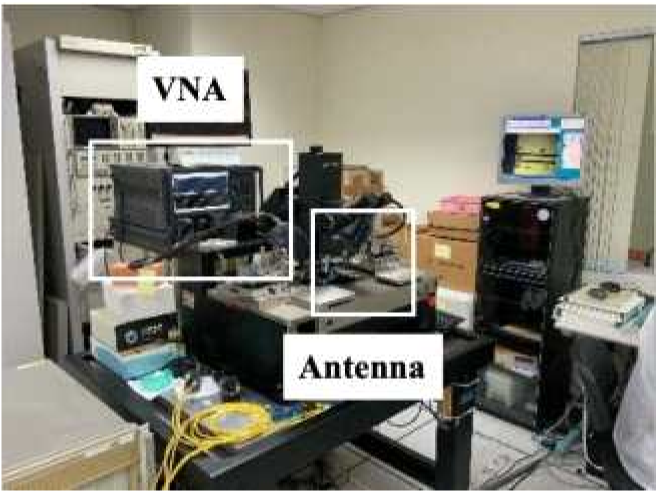


Figure 4. Measurement environment setup

Figure 7(a), 7(b), and 7(c) show that the analysis of surface current distribution for three frequency bands: 99 GHz, 108 GHz (at the lowest resonance point), and 114 GHz. Fig 7(a) illustrates the current distribution at 99 GHz, showing a uniform distribution of surface currents. Fig 7(b) represents the current distribution at 108 GHz, where the currents are primarily excited at the edges of the antenna. Fig 7(c) illustrates the distribution of current at 114 GHz, exhibiting a similar trend to Fig 7(a), with currents approximately evenly distributed across the antenna surface.

High-frequency electromagnetic simulation software was used to simulate the 3D radiation pattern of the antenna at frequencies of 99 GHz, 108 GHz, and 114 GHz, as depicted in Figure 8(a), 8(b), and 8(c) consecutively. In Fig 8(a), the radiation pattern illustrates that the main radiation direction is concentrated on the left side of the chip antenna, with maximum gain of -4.9 dB. In Fig 8(b), the radiation direction shifts towards the top of the antenna, with maximum gain of -4.6 dB. Fig 8(c) exhibits a radiation pattern similar to Fig 8(b), with maximum gain of -4.5 dB.

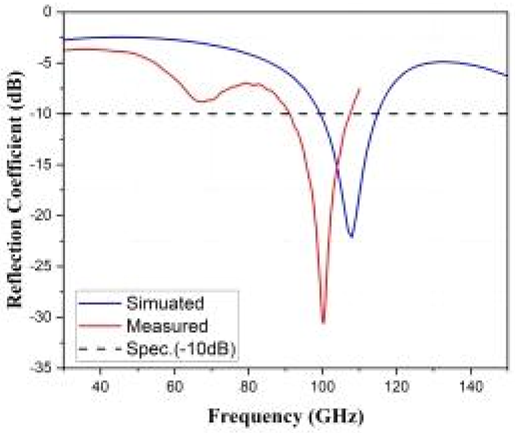


Figure 5. Simulated and measured reflection coefficient data

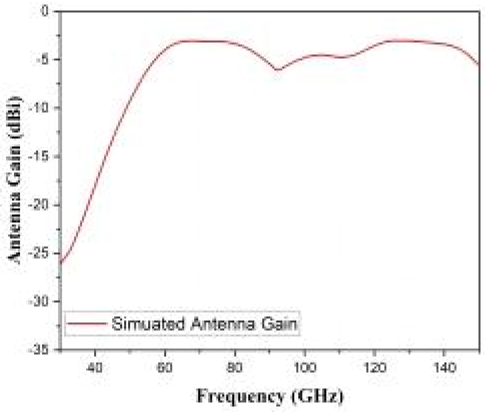


Figure 6. Simulated antenna gain plot

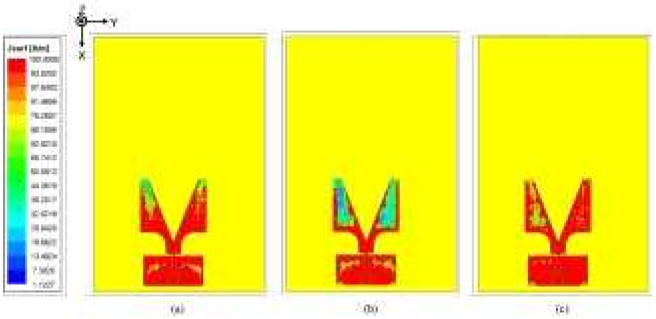


Figure 7. Current distribution (a) 99 GHz (b) 108 GHz and (c) 114 GHz

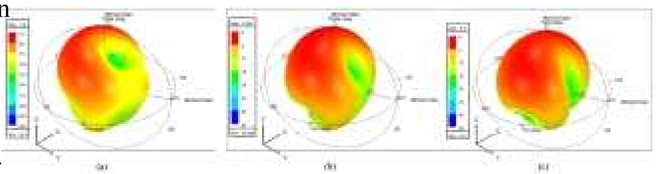


Figure 8. 3D radiation patterns (a) 99 GHz (b) 108 GHz and (c) 114 GHz

IV. Conclusions

This paper proposes a W-band chip antenna with a Vivaldi-like antenna structure and fabricated in CMOS technology. The size of the chip is 800 μm×1200 μm. After simulation and measurement, the simulated data, based on a reflection coefficient standard of -10 dB, covers a frequency range between 99.3 and 114.8 GHz., with a bandwidth percentage of 14.08%. The antenna exhibits a gain of approximately -4.5 dBi in this frequency range. The measured data show a bandwidth ranging from 91.1 - 107.1 GHz, with a bandwidth percentage of 16.14%. In the future, it will be possible to explore the impact of different CMOS processes on chip antennas. By employing different structural designs, it will be feasible to achieve improved performance

References

[1] C.-W. Chiang, C.-T. M. Wu, N.-C. Liu, C.-J. Liang, and Y.-C. Kuan, "A Cost-Effective W-Band Antenna-in-Package Using IPD and PCB Technologies," *IEEE Transactions on Components, Packaging and Manufacturing Technology,* vol. 12, no. 5, pp. 822-827, 2022.

[2] M. de Kok, A. B. Smolders, and U. Johannsen, "A review of design and integration technologies for D-band antennas," *IEEE Open Journal of Antennas and Propagation,* vol. 2, pp. 746-758, 2021.

[3] L. Marnat, A. A. Carreno, D. Conchouso, M. G. Martı, I. G. Foulds, and A. Shamim, "New movable plate for efficient millimeter wave vertical on-chip antenna," *IEEE Transactions on Antennas and Propagation,* vol. 61, no. 4, pp. 1608-1615, 2013.

[4] M. Alibakhshikenari *et al.*, "A comprehensive survey on antennas on-chip based on metamaterial, metasurface, and substrate integrated waveguide principles for millimeter-waves and terahertz integrated circuits and systems," *IEEE Access,* vol. 10, pp. 3668-3692, 2022.

[5] M. R. Karim, X. Yang, and M. F. Shafique, "On chip antenna measurement: A survey of challenges and recent trends," *IEEE Access,* vol. 6, pp. 20320-20333, 2018.

[6] M. K. Hedayati *et al.*, "Challenges in on-chip antenna design and integration with RF receiver front-end circuitry in nanoscale CMOS for 5G communication systems," *IEEE Access,* vol. 7, pp. 43190-43204, 2019.

[7] B. B. Adela, P. T. van Zeijl, U. Johannsen, and A. B. Smolders, "On-chip antenna integration for millimeter-wave single-chip FMCW radar, providing high efficiency and isolation," *IEEE Transactions on Antennas and Propagation,* vol. 64, no. 8, pp. 3281-3291, 2016.

[8] P. Burasa, T. Djerafi, and K. Wu, "A 28 GHz and 60 GHz dual-band on-chip antenna for 5G-compatible IoT-served sensors in standard CMOS process," *IEEE Transactions on Antennas and Propagation,* vol. 69, no. 5, pp. 2940-2945, 2020.
[9] C. Mustacchio, L. Boccia, E. Arnieri, and G. Amendola, "A gain levelling technique for on-chip antennas based on split-ring resonators," *IEEE Access,* vol. 9, pp. 90750-90756, 2021.
[10] Q. Van den Brande *et al.*, "A hybrid integration strategy for compact, broadband, and highly efficient millimeter-wave on-chip antennas," *IEEE Antennas and Wireless Propagation Letters,* vol. 18, no. 11, pp. 2424-2428, 2019.
[11] Y. Yu, Z. Akhter, and A. Shamim, "Ultra-Thin Artificial Magnetic Conductor for Gain Enhancement of Antenna-on-Chip," *IEEE Transactions on Antennas and Propagation,* vol. 70, no. 6, pp. 4319-4330, 2022.
[12] H. Zhu, X. Li, W. Feng, J. Xiao, and J. Zhang, "235 GHz on-chip antenna with miniaturised AMC loading in 65 nm CMOS," *IET Microwaves, Antennas & Propagation,* vol. 12, no. 5, pp. 727-733, 2018.
[13] M.-A. Chung, Y.-H. Chen, and I.-P. Meiy, "Antenna-on-Chip for Millimeter Wave Applications Using CMOS Process Technology," *Telecom,* vol. 4, no. 1, pp. 146-164, 2023.
[14] L. Tripodi *et al.*, "Broadband CMOS millimeter-wave frequency multiplier with vivaldi antenna in 3-D chip-scale packaging," *IEEE transactions on microwave theory and techniques,* vol. 60, no. 12, pp. 3761-3768, 2012.

MING-AN CHUNG (Member, IEEE) received the B.Eng. and M.Eng. degrees in electronic engineering from the Chang Gung University, Taoyuan, Taiwan and the D.Eng. degree in electrical engineering from the National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan, in 2003, 2005, and 2016, respectively. He is currently an Associate Professor with the Department of Electronic Engineering, National Taipei University of Technology (NTUT), where he also serves as the Leader of the Innovation Wireless Communication and Electromagnetic Applications Laboratory. His research interests include wireless communication propagation, intelligent robotics, self-driving vehicles, antenna design for various mobile and wireless communications, electromagnetic theory, and applications. He is also a Reviewer of many scientific journals, including the IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, IEEE Transactions on Industrial Informatics, Journal of Intelligent & Robotic Systems, IET Microwaves, Antennas and Propagation, IEEE Antennas and Wireless Propagation Letters, International Review of Electrical Engineering, International Journal on Communications Antenna and Propagation and AEÜ - International Journal of Electronics and Communications, and many international conferences, including ICRA, ICCE-TW, RFIT, ICBEB, EMCAR and SNSP.

MING-CHUN HSIEH received the B.Eng. degrees in electronic engineering from the National Taipei University of Technology (NTUT). Previously, he worked at HTC, where his gained valuable industry experience. At present, he is pursuing a Master's degree at Taipei Tech with a focus on the antenna field. His research interests encompass the fields of AI, edge devices, the Internet of Things (IoT), and antenna technology.

KUO-CHUN TSENG received the B.S. degree in Computer and Communication from the National Pingtung University, Pingtung, Taiwan, in 2021. He is pursuing an M.S. degree in electronic engineering with the National Taipei University of Technology, Taiwan. His current research interest includes the design of CMOS RF/microwave integrated circuits and antenna phased arrays for the applications of the fifth generation of mobile communications.

Design of Ka-band Chip Antenna Based on Slot Antenna

Ming-An Chung*
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
minganngchung@ntut.edu.tw

Kai-Xiang Chen
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
t112368524@ntut.edu.tw

Kuo-Chun Tseng
Department of Electronic Engineering
National Taipei University of Technology
Taipei, Taiwan
t110368157@ntut.edu.tw

***Abstract*— A broadband chip antenna designed for millimeter wave (mmW) can be used in fifth-generation (5G) frequency bands. The on-chip mmW antenna designed for the Ka-band utilizes standard Complementary Metal-Oxide-Semiconductor (CMOS) technology. In this design, the mmW antenna structure is formed by connecting the top layer (Metal6) and the bottom metal layer (Metal1) to reduce losses. Many architectures have been proposed for chip antennas to overcome the metal thickness during fabrication and improve the gain of chip antennas. Therefore, this paper proposes an on-chip antenna for the fifth-generation mobile communication millimeter-wave frequency band. Using high-frequency electromagnetic simulation software, the chip antenna exhibits a minimum return loss of -18dB at 31GHz and a peak gain of -7 dB. The measured reflection coefficient is below -10dB from 18.8 GHz to 32.5 GHz.**

***Keywords*— Antenna on-chip, millimeter-wave, CMOS**

I. Introduction

The 5G communication millimeter wave frequency band has recently attracted the attention of academia and the industry and has also become the focus of discussion in the field of the Internet of Things and future communication frequency band technology. It not only provides faster data connections but also enables services with virtually no latency characteristics [1, 2].

Antennas play a vital role in communication systems and can be used to send and receive signals, improving communication quality by compensating for propagation losses [3]. Due to the higher frequency range used in mmW or future terahertz (THz) applications, the antennas used are very small in size [4]. Furthermore, with the development of high system integration, mmW antenna, and feed circuit technologies have become increasingly mature, making chip antennas a favorable choice.

In [5], the study investigates how chip antennas achieve high gain and wideband performance. The application of a pair of switches is used to design frequency reconfigurability on the antenna, enhancing the antenna's bandwidth. The chip utilizes CMOS technology with high resistivity (HR) to improve antenna gain. A chip-integrated spherical dielectric resonator is used, and the antenna is excited by an on-chip annular resonator. Two vertical microstrip line feeds generate radiation, increasing the antenna's bandwidth [6].

This paper designs a broadband on-chip array antenna for Ka-band. The antenna is implemented using CMOS technology. It exhibits a minimum reflection coefficient of -18 dB at 31 GHz and -7 dBi peak gain. The measured reflection coefficient is under -10dB, and the frequency range is 18.8 GHz to 32.5 GHz, making it highly suitable for applications in the 5G mobile communication mmW frequency band.

II. Antenna Design

This paper presents an on-chip wideband antenna designed using CMOS technology for the 38 GHz millimeter-wave frequency band. The overall on-chip antenna comprises Metal6 with a thickness of 2.34 μm for the design of the antenna body and a bottom layer (Metal1) with a thickness of 0.53 μm for grounding, as shown in Figure 1.

Figure 2 and Table 1 depict the geometric structure and parameters of the antenna. The chip antenna occupies an area of 1150 × 1000 μm2. Simulation results show that the return loss of the design is 18.8-32.5 GHz in the frequency range below -10 dB. In Figure 5, the lowest point of the reflection coefficient is -27dB at 26 GHz. The gain is illustrated in Figure 6, reaching a peak value of -7 dBi at approximately 40 GHz.

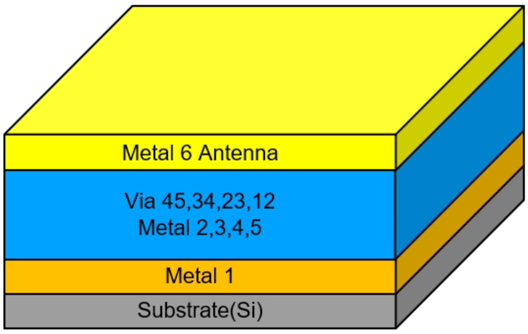


Figure 1. Layer structure of 0.18-μm CMOS.

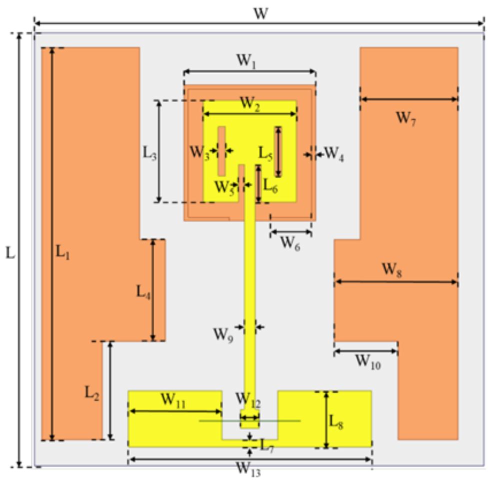


Figure 2. Slot antenna CMOS architecture.

TABLE 1. ON-CHIP ANTENNT PARAMETERS.

Parameter	Value(μm)	Parameter	Value(μm)
W	1200	L	1150
W_1	350	L_1	1040
W_2	250	L_2	260
W_3	20	L_3	270
W_4	10	L_4	270
W_5	15	L_5	130
W_6	110	L_6	100
W_7	260	L_7	20
W_8	330	L_8	150
W_9	30		
W_{10}	170		
W_{11}	250		
W_{12}	50		
W_{13}	650		

III.RESULT

This article describes an antenna design on-chip. The proposed antenna is intended for the mmW frequency band of fifth-generation mobile communication. Figure 3 shows a microscopic image of the chip. During measurement, the on-chip antenna is placed on the signal source platform using a suction pen, and the GSG RF probe feeds the chip antenna while being externally connected to a network analyzer for measurement. Prior to measurement, LRRM calibration is performed, followed by verification using microstrip lines (Thru, Open, Short). The chip measurement is conducted after confirming that all verifications are within the acceptable range of errors. The equipment used for chip antenna measurement is the Network Analyzer (Keysight N5247A), which has a measurement bandwidth of 10MHz to 67GHz. The software used for chip antenna measurement is Cascade WinCal XE and Agilent IC-CAP. Figure 4 illustrates the measurement environment setup.

Observing the actual measured S11 values, as shown in Figure 5, the simulation trend slightly deviates towards lower frequencies. The simulated antenna is a 38GHz slot chip antenna. The measured reflection coefficient meets the standard below -10dB, and the bandwidth is 18.8-32.5GHz. The area of the chip antenna is 1150×1000 μm. In the simulation, the M1 layer is used as the ground to realize the reflection band, and the slot effect is used to enhance the antenna gain so that the patch antenna can obtain the maximum gain. The simulated gain of the patch CMOS on-chip antenna is about -7dBi, as shown in Figure 6. The gain of an mmW patch antenna cannot be measured directly due to the current limitations of the instrumentation, so the gain graphs are based on simulated values.

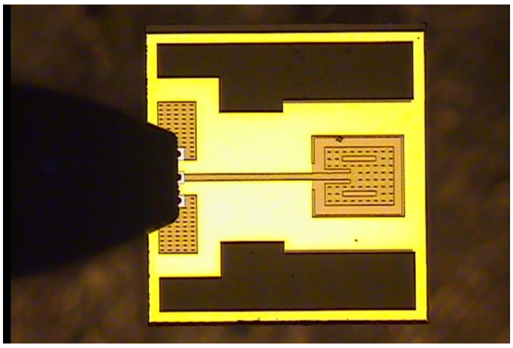


Figure 3. Micrograph of the proposed chip antenna.

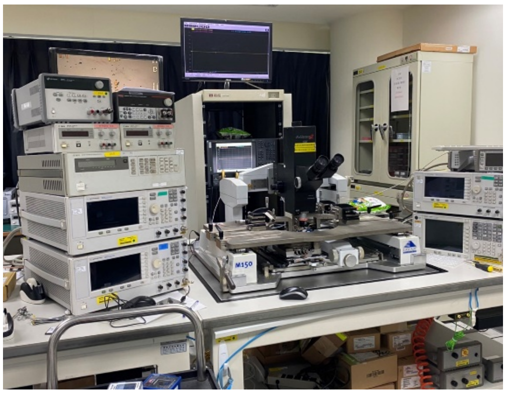


Figure 4. Measurement environment architecture.

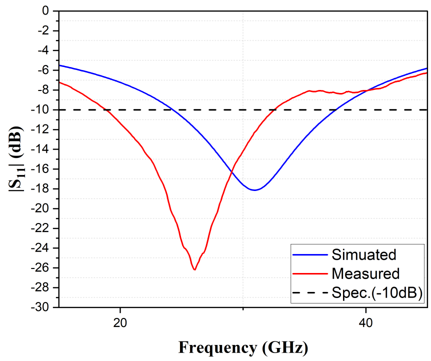


Figure 5. Hi

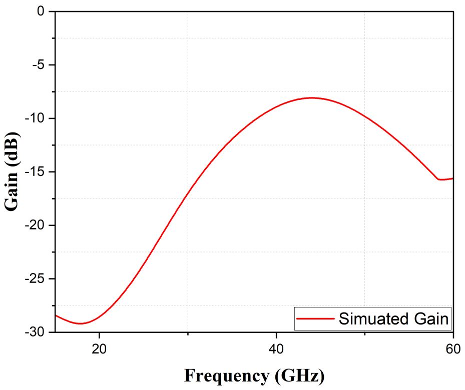


Figure 6.

Figure 7 shows the current distribution on the surface of the CMOS antenna. As can be seen, the current is largely stimulated at the antenna's feeding location when the frequency is at 26GHz, and coupling takes place between the antenna and its surroundings as well as between the Metal 1 ground plane and the antenna's central slot to produce bandwidth. The antenna's overall current distribution is uniform.

The 2D radiation patterns of this on-chip antenna are shown in Figure 8. The figure displays the antenna xz-plane and yz-plane of the antenna at 26GHz. The 3D radiation pattern in Figure 9 shows that at 26GHz, the overall radiation is uniformly distributed directly above the chip antenna, and the radiation energy is highly concentrated without any radiation divergence.

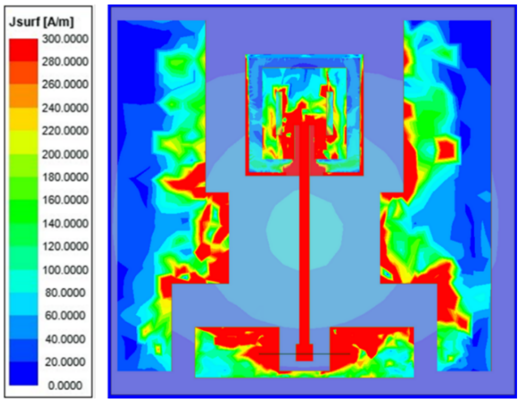


Figure 7. Current distribution (a) 99 GHz (b) 108 GHz and (c) 114 GHz

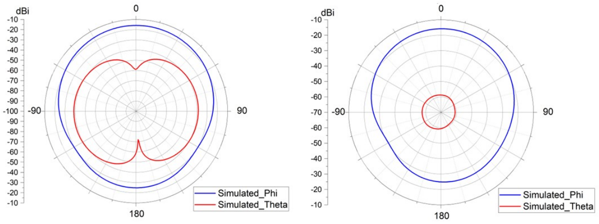


Figure 8. 3D radiation patterns (a) 99 GHz (b) 108 GHz and (c) 114 GHz

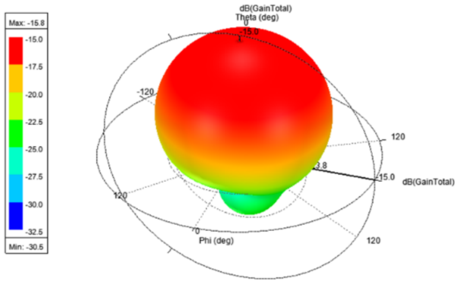


Figure 9.

IV. Conclusions

This research suggests a Ka-band chip antenna with a slot antenna-like structure that can be produced using the TSMC 0.18-m CMOS technology process. The CMOS on-chip antenna measures 1150 by 1000 m2. Through the simulations and measurements, the antenna's simulated frequency range of 25-38 GHz with a bandwidth percentage of 52% and an antenna gain of approximately -6 dBi, satisfying the reflection coefficient standard below -10 dB. The measured data demonstrates a bandwidth range of 18.8-32.5 GHz with a bandwidth percentage of 54.8%. In order to raise the antenna bandwidth and gain performance and realize diversified applications, it will be possible to study the development and impact of various CMOS processes on-chip antennas in the future.

References

[1] P. Burasa, T. Djerafi, and K. Wu, "A 28 GHz and 60 GHz dual-band on-chip antenna for 5G-compatible IoT-served sensors in standard

CMOS process," IEEE Transactions on Antennas and Propagation, vol. 69, no. 5, pp. 2940-2945, 2020.

[2] P. He et al., "A W-band 2× 2 rectenna array with on-chip CMOS switching rectifier and on-PCB tapered slot antenna for wireless power transfer," IEEE Transactions on Microwave Theory and Techniques, vol. 69, no. 1, pp. 969-979, 2020.

[3] C. Ma, S. Y. Zheng, Y. M. Pan, and Z. Chen, "Millimeter-wave fully integrated dielectric resonator antenna and its multi-beam application," IEEE Transactions on Antennas and Propagation, vol. 70, no. 8, pp. 6571-6580, 2022.

[4] S. Sahin, N. K. Nahar, and K. Sertel, "Noncontact Characterization of Antenna Parameters in mmW and THz Bands," IEEE Transactions on Terahertz Science and Technology, vol. 12, no. 1, pp. 42-52, 2021.

[5] Y. Song et al., "An on-chip frequency-reconfigurable antenna for Q-band broadband applications," IEEE Antennas and Wireless Propagation Letters, vol. 16, pp. 2232-2235, 2017.

[6] Z. Ahmad and J. Hesselbarth, "On-chip dual-polarized dielectric resonator antenna for millimeter-wave applications," IEEE antennas and wireless propagation letters, vol. 17, no. 10, pp. 1769-1772, 2018.

MING-AN CHUNG (Member, IEEE) received the B.Eng. and M.Eng. degrees in electronic engineering from the Chang Gung University, Taoyuan, Taiwan and the D.Eng. degree in electrical engineering from the National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan, in 2003, 2005, and 2016, respectively. He is currently an Associate Professor with the Department of Electronic Engineering, National Taipei University of Technology (NTUT), where he also serves as the Leader of the Innovation Wireless Communication and Electromagnetic Applications Laboratory. His research interests include wireless communication propagation, intelligent robotics, self-driving vehicles, antenna design for various mobile and wireless communications, electromagnetic theory, and applications. He is also a Reviewer of many scientific journals, including the IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, IEEE Transactions on Industrial Informatics, Journal of Intelligent & Robotic Systems, IET Microwaves, Antennas and Propagation, IEEE Antennas and Wireless Propagation Letters, International Review of Electrical Engineering, International Journal on Communications Antenna and Propagation and AEÜ - International Journal of Electronics and Communications, and many international conferences, including ICRA, ICCE-TW, RFIT, ICBEB, EMCAR and SNSP.

KAI-XIANG CHEN received the B.Eng. from National Taipei University of Technology (NTUT), Taipei, Taiwan, in 2023. Prior to this, he worked at Inventec as a firmware engineer, where he contributed to the development of OpenBMC, gaining substantial experience in both hardware and software domains. Currently, he is pursuing a Master's degree at National Taipei University of Technology, specializing in the field of electromagnetics. His research interests encompass antenna technology, artificial intelligence (AI), unmanned devices, and the Internet of Things (IoT).

KUO-CHUN TSENG received the B.S. degree in Computer and Communication from the National Pingtung University, Pingtung, Taiwan, in 2021. He is pursuing an M.S. degree in electronic engineering with the National Taipei University of Technology, Taiwan. His current research interest includes the design of CMOS RF/microwave integrated circuits and antenna phased arrays for the applications of the fifth generation of mobile communications.

Design of Calibration Algorithms for Fully-Activated Millimeter-Wave Phased Array Antennas

Juinn-Horng Deng*, Xiang-He Huang*, Chung-Lien Ho**, Yu-Chien Wu*
*Department of Electrical Engineering, Yuan Ze University, 135 Yuan-Tung Road, Chung-Li, Taoyuan, Taiwan
** Industrial Technology Research Institute, 195, Sec. 4, Chung Hsing Rd., Chutung, Hsinchu, Taiwan
E-mails: jh.deng@saturn.yzu.edu.tw, s1104804@mail.yzu.edu.tw, clho@itri.org.tw, s1114813@mail.yzu.edu.tw

***Abstract*—This paper focuses on the study of the calibration technology for fully activated millimeter-wave phased array antennas. It utilizes a self-developed 1x8 array antenna module in conjunction with the M3-Force software defined radio (SDR) platform. By initiating the calibration mode with all antennas fully activated, the array antenna tuning is accomplished using the rotating element electric field vector (REV) algorithm. However, this algorithm involves the issues of ambiguity, and the paper proposes a technique to strengthen the synthesis of vectors to stabilize the estimation of antenna calibration parameters. Furthermore, an enhanced version of the REV method is introduced for incrementally joint block tuning of subarrays, facilitating the calibration of large array antennas. The paper overcomes the limitations of the traditional REV method in large array antenna calibration. Computer simulation results confirm the superior performance of this technique in overcoming issues with traditional REV methods and its applicability to large array antenna calibration. Finally, the paper validates the proposed technique through hardware testing with an array antenna and an SDR platform, confirming that the enhanced REV technology can calibrate a 1x8 millimeter-wave phased array antenna and suggesting its potential extension to the calibration of larger array antennas in the future.**

***Keywords*—millimeter-wave phased array, rotating element electric field vector, software defined radio platform, calibration.**

I. Introduction

To cope with the significant growth in the Internet of Things (IoT) and mobile devices in the current 5G era, the bandwidth in the low-frequency bands is no longer sufficient to accommodate such a large number of devices. Therefore, the application of millimeter-wave frequencies is undoubtedly a trend. Millimeter waves, i.e., high-frequency signals, induce the rapid energy attenuation with increasing distance in transmission. Directional antennas can overcome this by focusing signal power, thereby extending the transmission distance. In directional antennas, phased array antennas play a crucial role in millimeter-wave transmission. Their high directionality compensates for the substantial path loss in millimeter-wave transmission.

Typically, each antenna element in a phased array antenna has independent phase shifters and attenuators. Electronic control of phase shifters allows the effective beam to be steered in a specific direction, while control of attenuators helps form the desired radiation pattern. However, due to the small wavelength of millimeter waves, phased array antennas are highly sensitive to small deviations. Phase responses and position responses of individual antenna elements may be affected by inevitable errors and manufacturing tolerances, and the performance of the antenna can also be influenced by temperature changes over time. If these deviations and errors are not corrected, unwanted beams may form in various directions, causing interference and limiting the gain of the antenna array as well as the overall system performance. This study addresses the issues of phase and position inconsistency among antenna elements in phased array antennas, as well as the nonlinearity of the control curves of phase shifters and attenuators. First, it explores feasible calibration techniques [1]-[7] to adjust the phase and position of antenna elements, overcoming the nonlinearity in the control curves. Next, it employs a Software Defined Radio (SDR) platform in conjunction with millimeter-wave phased array antennas for practical testing. The study validates the feasibility of the calibration technique by measuring antenna radiation patterns and confirms that this technology effectively enhances the performance of antenna beamforming.

II. Explanation of the Integration of Millimeter-Wave Active Phased Array Antennas with Software-Defined Radio Platforms

This paper focuses on the calibration of millimeter-wave phased array antennas, with the details of the calibration technique presented in the third section. This section introduces the integration and development of the active array antenna with SDR platform. To begin with, the array antenna module employed in this thesis is a self-manufactured 1x8 millimeter-wave phased array antenna module [8] operating at 28GHz by Yuan Ze University. The physical appearance of the array antenna module is illustrated in Figure 1.

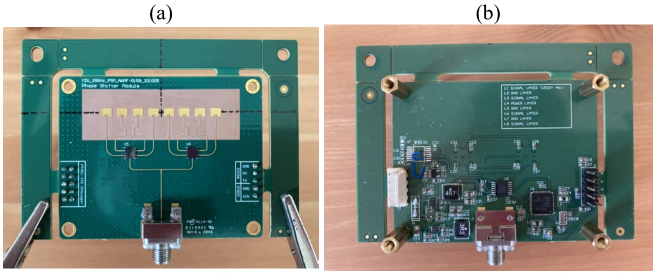


Figure 1. Homemade millimeter-wave active array module. (a) Physical front view. (b) Physical rear view.

The module incorporates an ARM (MCU) chip and an Anokiwave 0158 Beamforming IC, which is controlled by ARM to adjust the 4-bit Gain and 6-bit Phase Shifter. The array module is designed to be calibrated with SDR platform.

Next, for the explanation, the integration involves calibrating the antenna module using both a PC and the M3-

Force SDR. Based on the design module, the PCB design incorporates high-frequency microstrip circuits and low-frequency ARM control circuits. This design coexists on an eight-layer PCB. Grounding and interference elimination are crucial considerations in the design, leading to the incorporation of numerous large-area perforations in the PCB.

Regarding the development platform, it will be used to calibrate the Gain/Phase of the 1x8 millimeter-wave phased array antenna PCB mentioned above. It can execute functions related to coherent wave pointing. Traditional network analyzer calibration platforms are known to be expensive. Such platforms include probes and network analyzers integrated to measure single-frequency signals transmitted by different antennas (in switching mode). Through network analyzer computations, stable Gain and Phase values for different antennas can be obtained. This stable result is widely adopted in the industry for millimeter-wave array calibration.

Another cost-effective measurement platform is the M3-Force software defined radio (SDR) platform. This device includes two sets of AD9361 RF modules, capable of simultaneous control for four transmitters and four receivers. It supports a maximum bandwidth of 56MHz and frequencies up to 6GHz, with transmit attenuation and receive gain support of up to 40dB. When paired with MATLAB programs for signal processing, it can measure antenna phase and position. This SDR platform also incorporates millimeter-wave up/down-conversion components and a computer. Unlike traditional network analyzers (common in millimeter-wave band measurements), this platform combines physical BaseBand I/Q signals, upconverts them to millimeter-wave frequencies, retrieves the I/Q signals, calculates the Gain and Phase for the 1x8 antenna paths, and then performs compensation and calibration. This approach closely mimics the "full system calibration requirements for practical millimeter-wave communication applications." Therefore, the calibration method [8] using the SDR platform has gained recent attention. The process diagram of this SDR measurement platform is illustrated in Figure 2.

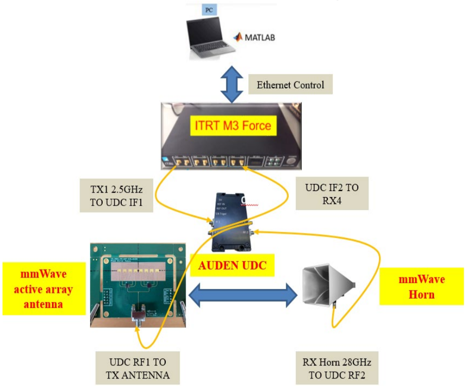


Figure 2. Flowchart of the SDR Measurement Platform.

Next, this section is dedicated to performing beam pattern measurements for the 1x8 millimeter-wave module (after compensation for calibration parameters). The measurement platform is illustrated in the following diagram, comprising the M3-Force SDR platform for signal transmission, a non-reflective chamber, and the chamber involves a 1x8 array antenna mounted on a 360-degree rotator. This setup allows for beam pattern measurements. Moreover, after the calibration in subsequent sections, beam pattern measurements will be conducted using this non-reflective chamber setup.

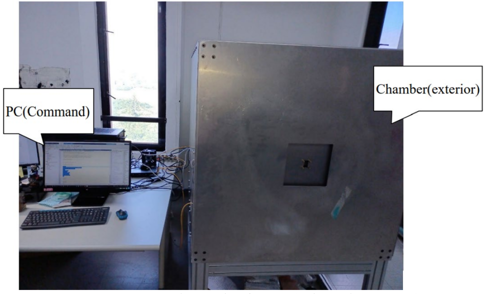


Figure 3. Testing platform (Chassis & PC).

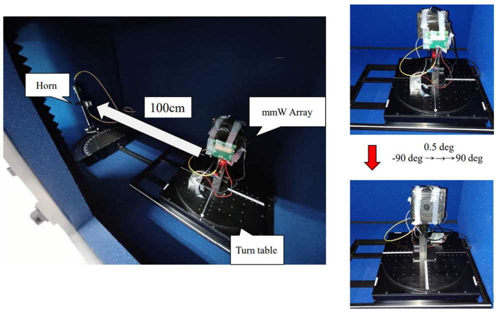


Figure 4. Internal arrangement of the chassis.

Furthermore, Figure 3 depicts the calibration chamber and the computer platform. Then, the external components, including the millimeter-wave up/down-conversion module and the M3-Force SDR platform, are connected to the calibration platform. Subsequently, Figure 4 illustrates a conceptual diagram for future testing of REV antenna inter-calibration (with a distance of more than 40cm between the Horn and Array), representing far-field testing. Finally, after antenna array calibration, Figure 4 outlines the utilization of this calibration for beam pattern measurements ranging from -90 degrees to 90 degrees.

III. Research on Algorithms for Estimating Phase and Alignment of Phased Array Antennas

This section explores the research on algorithms for estimating the phase and alignment offsets of each antenna element in a phased array, aiming to obtain the calibration table for the antennas and verify the array antenna's pointing performance. Firstly, the discussion focuses on the Rotating Element Electric Field Vector (REV) method [1], which is internationally recognized for its ability to identify the phase/alignment deviations of antennas using only amplitude/power (real-number) measurement techniques. Moreover, the operational procedure for REV antenna element field measurements involves simultaneous testing of different phases across the array antenna. In other words, the REV technique incorporates real-world environmental factors into the calibration process. This includes factors such as changes in the T/R module, variations in the feed circuit, and diffraction effects caused by the antenna structure, all of which are adjusted simultaneously by the REV technique. This approach provides a more comprehensive calibration.

The REV technique involves measuring all antennas to be calibrated, but only one antenna's phase is altered to measure the overall power variation. This approach takes into account the interaction coupling factors between antennas. Such a comprehensive adjustment of all coupling factors is expected to be more widely accepted in practical applications.

Having reviewed relevant literatures, the author in reference [2] mentions that through theoretical analysis of measurement errors, the amplitude and phase differences of the excited antenna elements will decrease as the number of antennas increases, especially when combined with the full activation of REV. This indicates that the calibration measurement errors for array antennas will decrease. In reference [3], the author extends the REV method to an expanded version, where multiple antennas rotate their phases together, allowing for faster calibration. Reference [4] introduces the development of the array antenna power synthesis response using high-order Fourier series, which can enhance REV performance. Reference [5] proposes the Fast REV method, which offers the advantage of accelerating the calibration speed of REV. Commonly used simple calibration methods, such as initial calibration stages, subsequent calibration stages, near-field (far-field) calibration stages, and even the need for extensive measurement data and time-consuming processes during on-board system self-diagnostic calibration, are addressed. Therefore, there is a need for fast and accurate calibration methods [5][6] to achieve the calibration goals of array antennas.

A. Description of REV Techniques

This section aims to provide a brief overview of the core principles of the REV technology. The paper will also extend and improve the "REV confusion detection" issue associated with this technology. Furthermore, it will introduce an enhanced version of the REV tailored for large arrays.

Firstly, REV is a calibration method for synthesizing array field vectors in a specified direction. This method possesses the synthetic characteristic of summing all antenna elements. The synthesized field vector undergoes changes when one antenna element is rotated from 0 degrees to 360 degrees (controlled by a phase shifter), resulting in variations in the synthesized vector. Based on the changing amplitude values of this synthesis, the amplitude and phase of the antenna element can be determined. The behaviour of the power representation of the changing synthesized amplitude will exhibit a sinusoidal curve. The schematic diagram of REV is as follows:

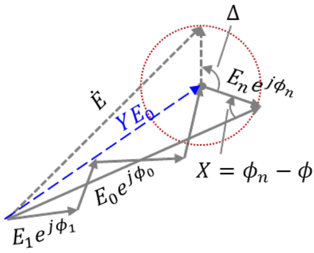


Figure 5. Schematic diagram of overall synthesized vectors in REV [1].

Figure 5 is an illustrative diagram of the synthesized vectors for all antennas. At this point, the mathematical expression for the synthesized vector field in Figure 5 (with the *n*th antenna changing phase Δ) is expressed as follows [1]:

$$\dot{E} = (E_0 e^{j\phi_0} - E_n e^{j\phi_n}) + E_n e^{j(\phi_n + \Delta)} \tag{1}$$

In the above formula, E_0 represents the synthesized field of all antennas, ϕ_0 denotes the phase rotation of the all antennas and ϕ_n denotes the phase rotation of the *n*th element. By removing the vector field E_n of the *n*th element and then introducing the change in phase Δ for the *n*th element (with a 360-degree variation), the overall dynamic synthesized quantity $\dot{E}$ can be obtained. Next, this technique aims to determine the amplitude ratio and phase deviation between the vector field of the *n*th element E_n and the original synthesized vector E_0, as expressed in the following equation [1]:

$$k = \frac{E_n}{E_0} \quad \text{and} \quad X = \phi_n - \phi_0 \tag{2}$$

Furthermore, explaining the reason for this sine wave curve, when Δ changes, there will be a variation in $\dot{E}$. At this point, the ratio (*Q*) relative to the original synthesized vector field E_0 can be obtained (measured), as expressed in the following equation [1]:

$$Q \equiv \frac{|\dot{E}|^2}{E_0^2} = (Y^2 + k^2) + 2kY\cos(\Delta + \Delta_0) \tag{3}$$

where $Y^2 = (\cos X - k)^2 + \sin^2 X$ and $\tan\Delta_0 = \sin X / (\cos X - k)$. In (3), *Q* performs the sine wave curve when the Δ undergoes a 360-degree rotation. Moreover, (3) represent the interactive performance of *Q* at its maximum value with *X* and *k*. Next, further exploration is conducted on the ratio relationship between the maximum and minimum values of *Q*, as expressed in the following [1]:

$$r = \pm\frac{Y + k}{Y - k} > 1 \tag{4}$$

(4) has the representation indicating the need to determine whether to adopt + or adopt - in estimating *X* and *k* with confusion problem. Through the integration calculation of the aforementioned equations (3) and (4), values for *k* and *X* can be obtained (under two conditions) [1]. The explanation of the above REV technology reveals that the relative phase difference and amplitude ratio can be obtained, but there is an additional risk of confusion in judgment. A proposed method to avoid this risk is presented subsequently. The total number of measurements in the above scenario can be calculated, considering *N* antennas (for example, *N*=8 antennas), each having N_b bits to change the phase (for example, N_b =6 bits). Based on the above, it can be determined that there are $N \cdot 2^{N_b}$ changes in *Q* values to be measured. If *N*=8 and N_b =6, there are *Q*=512 ($8 \cdot 2^6 = 512$) value variations to be measured. After obtaining these *Q* values and considering the conditions *Y*>*k* or *Y*<*k* as mentioned above, the relative amplitude ratio (*k*) and phase difference (*X*) can be determined.

With these values, the relative phase calibration between antennas (i.e., $X_n - X_1$ for n=2, 3, ..., N) and amplitude calibration between antennas (i.e., k_n/k_1 for n=1, 2, ..., N) can be completed. These parameters form a table, which can be used for compensation and beamforming verification on the SDR and array antenna platform in the subsequent Section III.B. Further details will be explained in the following experiments.

B. Description of the Proposed Enhanced REV Technology

In the previous section, it was explained that the REV technology involves confusion issues due to the possibility of the overall synthesized vector E_0 undergoing phase cancellation or being lower in magnitude than the field vector of the nth element being tested. Conversely, the overall synthesized vector E_0 remains stable, maintaining a strength greater than the field vector of the nth element. At this point, existing literature suggests additional steps to assess the potential situations of the current synthesized vector E_0. The author [6] proposes altering the size of the field vector of the nth element to aid in distinguishing the relationship between E_0 and the new E_n. While this approach improves the confusion issue, it indirectly increases the number of Q values that need to be measured by up to twice. As illustrated in the example explained in the previous section, it involves Q=1024 ($2 \cdot 8 \cdot 2^6 = 1024$) values to be measured, which results in a higher level of complexity. In this section, we enhance the energy E_0 of the REV to reduce the probability of confusion, achieving strengthened $E_0 > E_n$. This significantly diminishes the likelihood of confusion. The indirect approach we propose involves increasing the measurement E_0 by a factor of α. In other words, we can perform α measurements and summations of the overall synthesized vector, with the mathematical explanation of this summation's effectiveness provided below:

$$\begin{aligned}\dot{E}' &= (\alpha E_0 e^{j\phi_0} - E_n e^{j\phi_n}) + E_n e^{j(\phi_n+\Delta)}, \alpha > 1 \\ &= (E_0 e^{j\phi_0} - E_n e^{j\phi_n}) + E_n e^{j(\phi_n+\Delta)} + (\alpha - 1)\cdot E_0 e^{j\phi_0} \qquad (5) \\ &= \dot{E} + (\alpha-1)\cdot E_0 e^{j\phi_0} \approx \dot{E} + \alpha \cdot E_0 e^{j\phi_0}, \alpha \gg 1\end{aligned}$$

The above formula illustrates that the original REV method had a value $\dot{E}$, which could lead to confusion. By strengthening E_0 the value α times, we can significantly achieve the result of Y>k, satisfying the stable estimation performance of the sum E_0 being greater than E_n. Therefore, with a high number of α, a lower probability of confusion can be obtained. The value of α can be chosen to be less than the total number of antennas. Following the explanation above, this paper chooses an enhanced REV method with α=7, which is compared with several literature methods such as traditional REV [1], Fast REV [5], and improved confusion method [7] for stability. The criterion for determining stability is based on calculating the Mean Square Error (MSE) between the REV Beam Pattern and the Ideal Beam. If the MSE is greater than 0.3dB, it is considered an erroneous result. For example, if the MSE for the enhanced REV is less than 0.3dB, it is considered a successful calculation; otherwise, it is considered a failure. Using this approach, we conduct 5000 simulations and calculate the error rate as $\text{error rate} = \text{number of failures}/5000 \times 100\%$. The results for different methods are presented in the following figure:

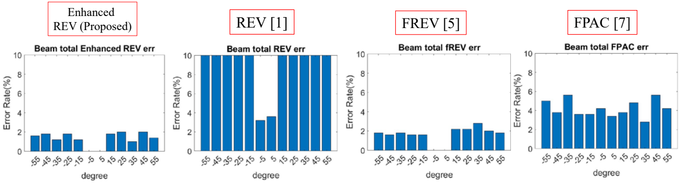


Figure 6. Comparison of stability among different methods.

The results indicate that the enhanced REV has excellent performance, while other methods show poor performance or require too many REV calculations. Therefore, the enhanced REV method proposed in this paper has a lower number of Q calculations. With α=7, the estimation is $8\times(2^6+\alpha)$=512+56 (only an additional 56 calculations, much lower than the previous estimate of 1024 Q values).

C. The Proposed Enhanced REV Technology for Large-Scale Arrays

Building upon the enhanced REV technology discussed above, this section extends to the array calibration of the Group-Based REV (Block-Based REV) in a Sub-Array configuration. The goal is to make it adaptable to large arrays and expedite the calibration process. Additionally, this approach considers factors related to mutual coupling between antennas, allowing for their simultaneous adjustment. The paper proposes the following two viable solutions.

1) Method 1: *First, simultaneously execute the REV technique within each block, waiting for completion within each block after REV. Then, coordinate the REV adjustments between blocks outside the respective blocks.*

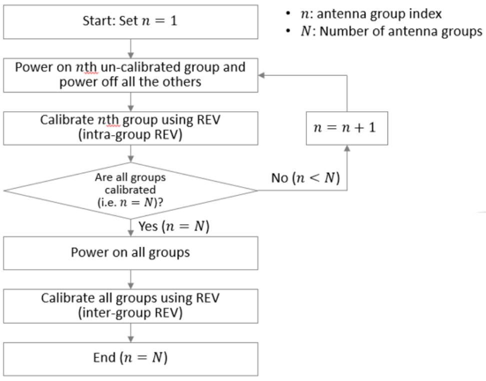

Start: Set $n = 1$
• n: antenna group index
• N: Number of antenna groups
Power on nth un-calibrated group and power off all the others
Calibrate nth group using REV (intra-group REV)
$n = n + 1$
Are all groups calibrated (i.e. $n = N$)?
No ($n < N$)
Yes ($n = N$)
Power on all groups
Calibrate all groups using REV (inter-group REV)
End ($n = N$)

Figure 7. Adjustment process diagram for Method 1.

This method involves calibrating both the intra-block and inter-block REV. It covers the entire large array of antennas, meaning that all array antennas are divided into N blocks (for example, 4 blocks, each containing 8 sub-antennas, totalling 32 antennas to be calibrated). Each element within each sub-block must undergo calibration using the enhanced REV method. As shown in the internal loop in the diagram of Figure 7, all internal calibrations within a group are completed. Once each group is aligned, as if integrating into a synthesized element, with N elements representing N groups, inter-group calibration is performed. Therefore, a total of N+1 REV operations are completed for the calibration of the entire array of antennas in that direction. The flowchart for this process is shown in Figure 7, and for the internal enhanced REV method, please refer to the explanations in Section III.B.

*2) **Method 2:** First, sequentially execute the REV adjustments within each block, then perform the cumulative REV adjustments between blocks. This adjustment utilizes the alignment strength energy provided by the preceding block to enhance the intensity value for REV within the next group block, thereby improving the tuning efficiency in the subsequent block.*

Compared to Method 1, this second method only requires N REV operations to complete the calibration of the entire array of antennas (lower than the N+1 REV operations in Method 1). The detailed flowchart for this method is shown in Figure 8. In this flowchart, the left loop represents the internal calibration process for Group 1, while the right loop represents the calibration process for Group 2. When calibrating Group 2, Group 1, which has already been calibrated, needs to be used for reference, and all internal calibration parameters of Group 2 are adjusted accordingly. However, the calibration table for Group 1 remains fixed, and the internal calibration for Group 2 is completed during the current REV operation. This approach allows aligning the internal calibrations of Group 2 and correcting the deviations between Group 1 and Group 2 simultaneously, providing several advantages. This process is extended to the 3rd, 4th, ..., Nth groups, completing the alignment of all antenna elements.

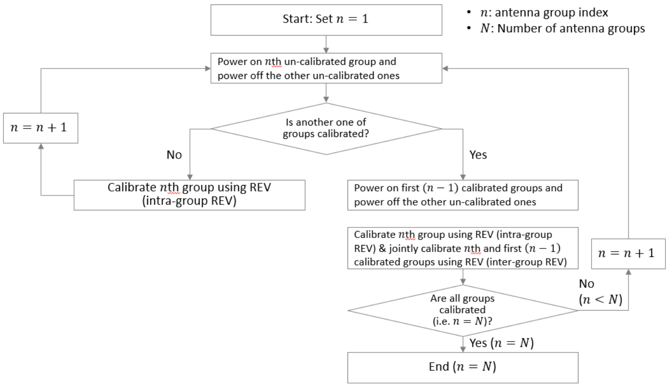

Start: Set $n = 1$
n: antenna group index
N: Number of antenna groups
Power on nth un-calibrated group and power off the other un-calibrated ones
$n = n + 1$
Is another one of groups calibrated?
No
Yes
Calibrate nth group using REV (intra-group REV)
Power on first $(n-1)$ calibrated groups and power off the other un-calibrated ones
Calibrate nth group using REV (intra-group REV) & jointly calibrate nth and first $(n-1)$ calibrated groups using REV (inter-group REV)
$n = n + 1$
No ($n < N$)
Are all groups calibrated (i.e. $n = N$)?
Yes ($n = N$)
End ($n = N$)

Figure 8. Adjustment process diagram for Method 2.

IV. Computer Simulation and Hardware Platform Testing

A. Computer Simulation Verification

This section focuses on computer program simulations for the Enhanced REV method proposed in Section III.B and the large array calibration method proposed in Section III.C. Firstly, the baseline antenna configuration is set with 16 antennas, inter-antenna spacing of the half wavelength, 6-bit phase shifters, 4-bit amplitude attenuators, and an SNR of 10dB. The simulation includes the influence of random variables on the amplitude and phase of each element, representing different antennas with various microstrip line distortion factors.

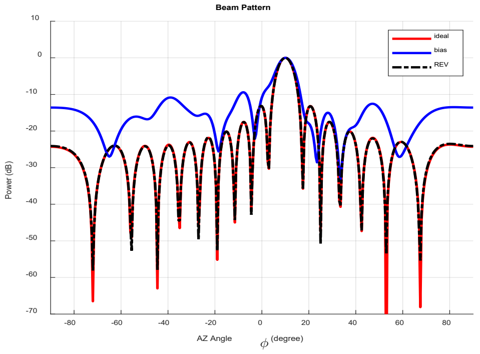


Figure 9. Simulation results of the enhanced REV method in Section III.B.

After running the simulation based on the above settings, the results are illustrated in Figure 9. The red curve represents the ideal beam pattern and the black dashed line represents the beam pattern with the Enhanced REV method (close to ideal). The blue curve represents the original simulation result with added gain/phase bias and SNR=10dB effects that have not been compensated yet. In the Figure 9, it can be observed that the proposed enhanced REV method exhibits good performance.

Next, simulations were conducted to validate the two different methods proposed for the large array Group REV technology in Section III.C. Firstly, the baseline antenna parameters are similar to the explanation in the previous section, with the only difference being an increased total number of antennas to 32. This section will explain the adjustments needed for Group X and Group Y with total 32 antenna elements, where Group X involves adjustments for 1x8 antennas, and Group Y involves adjustments for 1x4 blocks, as described in Section III.C. REV X represents the adjustment method for Group X, REV Y represents the adjustment method for Group Y, namely the method 1 proposed in Section III.C. Furthermore, the simulation results include AccSub, indicating the adjustment combined with the previously adjusted Group block, as the method 2 proposed in Section III.C. The simulation results are shown in the Figure 10. Both methods, i.e., Method 1 (REV X and REV Y) and Method 2 (REV X and REV Y with AccSub), successfully calibrate the array with total 32 antennas, and the results show a well-directed beam with good performance.

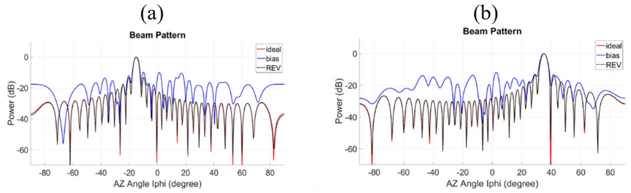

(a) (b)

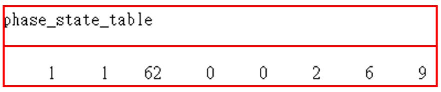
Figure 10. Simulation results in section III.C. (a) Method 1 (REV X & REV Y). (b) Method 2 (REV X & REV Y & AccSub).

Furthermore, from the above results, it can be inferred that the proposed enhanced REV calibration for large arrays, as outlined in the method 2 of Section III.C, achieves superior beamforming performance. In particular, the beam pattern after block-wise joint calibration closely approximates the ideal beam pattern depicted in red.

B. Hardware Platform Testing

This section involves the use of the antenna array and SDR platform from Section II, along with the enhanced REV algorithm proposed in Section III for practical calibration table and beam pattern verification. In other words, the chamber platform remains the same, and the SDR with UDC (Up-Down Converter) platform is also the same. The initial step is to perform antenna array calibration at a distance of 40 cm. The calibration algorithm follows the enhanced REV method in Section III.B. Thus, the red text indicates the scenario where 1x8 antennas are fully activated. In this case, the phase of the remaining 7 antennas remains unchanged, while the phase of one antenna varies by 360 degrees (utilizing 64 states, changing every 5.625 degrees). This process is repeated for each antenna, resulting in the recording of 64 Q values for each antenna (corresponding to the peak values recorded by the SDR platform for QPSK in each packet after time synchronization). This sequence is performed for all 8 antennas, resulting in a total of 8x(64+7) = 512+56 = 568 recorded Q values, including the 7 records for the enhanced state of being fully activated. Once all Q values are obtained, the REV algorithm is applied to calculate the k values and X values, representing amplitude ratio and phase difference. The phase table is then organized as shown in the table below:

TABLE 1. HARDWARE TESTING TABLE RESULTS



```
phase_state_table

     1     1    62     0     0     2     6     9
```

After obtaining the calibration table through the enhanced REV testing described above, a beam pattern diagram can be generated using a fully activated antenna array at 100 cm distance with a -90 degrees to 90 degrees rotation table, as shown in Figure 11 (pointing only towards 0 degrees).

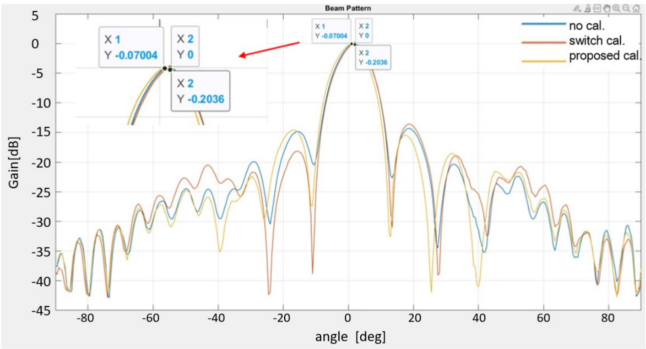


Figure 11. Diagram of field patterns from hardware testing.

From the results of this diagram, it can be observed that only the Proposed cal. (enhanced REV method) and the red Switch cal. calibration method [8] exhibit lower Null Pattern diagrams compared to the uncalibrated (original 1x8 array) beam pattern. This result confirms the feasibility of the calibration algorithm proposed in this paper, integrated with the array antennas and SDR platform. In the future, the approach will be extended to practical calibration tests with a larger number of antennas using the method outlined in Section III.C. It is important to note that, currently, the laboratory at Yuan Ze University does not have a large array antenna, so the method described in Section III.C cannot be practically tested and verified at this time.

V. CONCLUSIONS

The algorithm development for millimeter-wave antenna calibration is crucial, and this paper has achieved significant progress. When all antennas are fully activated, the method of validating the variation of the Radiating Element Electric Field Vector (REV) using broadband signals has been implemented. This method has gained international attention as it allows obtaining phase/amplitude calibration values of antennas while considering coupling factors between antennas, with measurements relying solely on amplitude (power) and all antennas being active. However, the REV technique has inherent challenges, and in Section III, an enhanced REV algorithm was proposed to address these issues. The enhanced REV technology has been validated through simulations, demonstrating advantages such as reduced computational requirements and faster estimation compared to other algorithms from existing literature. Furthermore, the paper extends the proposed methodology to the estimation rules for large array Group-based antennas in Section III.C, showing excellent performance. Finally, in Section IV, the developed 1x8 array antenna platform, SDR platform, and calibration algorithm were jointly tested, confirming the practicality and effectiveness of the proposed methodology.

ACKNOWLEDGMENT

This work was sponsored by the National Science Council, R.O.C., under Contract NSTC 112-2221-E-155-015.

REFERENCES

[1] S. Mano and T. Katagi, “A method for measuring amplitude and phase of each radiating element of a phased array antenna,” *Trans. IEICE*, vol. J65-B, no. 5, pp. 555-560, May 1982.

[2] A. M. Shitikov and A. V. Bondarik, “Multi-element PAA calibration with REV method,” *International Conference on Antenna Theory and Techniques*, vol. 2. pp. 761-764, Sept. 2003.

[3] T. Xie, J. Zhu, and J. Luo, “The simplified REV method combined with Hadamard group division for phased array calibration,” *IEICE Trans. Commun.*, vol. E101-B, no. 3, pp. 847-855, Mar. 2018.

[4] M. Liu and Z. Feng, “Combined rotating-element electric-field vector method for near field calibration of phased array antenna,” *IEEE International Conf. on Microwave and Millimeter Wave Tech.*, 2007.

[5] R. Long, *et al.*, “Fast amplitude-only measurement method for phased array calibration,” *IEEE Trans. on Antennas and Propagation*, vol. 65, no. 4, pp. 1815-1822, Apr. 2017.

[6] H. J. Yoon and B. W. Min, “Improved rotating-element electric-field vector method for fast far-field phased array calibration,” *IEEE Trans. Antenna Propagation*, vol. 69, no. 11, pp. 8021-8026, Nov. 2021.

[7] Zhai Yu, Su Donglin, “Ambiguity of rotating-element electric-field vector method and elimination method,” *Journal of Beijing University of Aeronautics and Astronautics*, vol. 38, no. 11, pp. 1450-1453, Nov. 2012.

[8] Juinn-Horng Deng, *et al.*, “Design of Millimeter Wave Active Array Antenna Module with Embedded System and Calibration of Software Defined Radio Platform,” *IEEE VTS 17th Asia Pacific Wireless Communications Symposium* (APWCS), pp. 1-5, Aug. 2021.

The Seamless Connection between Underwater and Terrestrial Communication for 6G

Tin-Yu Wu[1], Yi-Kai Chen[1], Fu-Jie Tey[2]

[1] Department of Management Information Systems, National Pingtung University of Science and Technology, Pingtung, Taiwan

[2] Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

tyw@mail.npust.edu.tw, yikaie@g4e.npust.edu.tw, d10907001@gapps.ntust.edu.tw

Abstract— **Underwater communication was chiefly employed in military affairs and underwater exploration in the early days. In recent years, thanks to the progressive advancement of communication networks, many protocols have already been developed for underwater communication. However, due to technological factors, underwater communication is still looking for a seamless link to terrestrial communication. This paper proposes a novel frame format to improve the frame head of underwater communication, and presents a corresponding operation process of underwater nodes with the hope that the data link layer of the OSI 7-layer model provides the Media Access Control (MAC) protocol to the corresponding underwater network. This proposed method can be provided as a standard solution to create a seamless connection between underwater and terrestrial communication.**

Keywords— **Underwater communication, 6G, Gateway for underwater communication.**

I. Introduction

To enable the space-air-ground integrated network for the emerging 6G networks, many countries have initiated the research on quantum communication technology to improve communication performance. However, the standardisation of quantum communication is still in its infancy. Compared with the OSI 7-layer network architecture, quantum communication does not have its specific data transmission method in the physical layer.

Since various Media Access Control (MAC) protocols have been designed for underwater environment and characteristics of sounds have been fully studied in the classical physics, this paper proposes an underwater communication gateway to enable the seamless connection between underwater communication and terrestrial communication. Based on Wireless Ad Hoc Network (WANET) and ALOHA, we modify the Ethernet frame format widely used in terrestrial communication, improve the drawbacks of using the Ethernet frame format or the IEEE802.11 frame format in underwater communication, and present a novel frame format for underwater communication.

The proposed gateway for underwater communication can translate the frame format of terrestrial communication and sends the results to designated underwater nodes. In addition, by introducing the idea of backbone network and last mile network, this paper classifies underwater nodes to increase communication network reliability but reduce network deployment cost.

II. Method

To propose a new frame head format for underwater communication, we first modify the Ethernet frame format to match the properties of underwater communication and satisfy its transmission requirements. To put the protocol into practice, this paper also gives the interaction details between frames and nodes so that nodes and frame format can be compatible and collaborative with each other.

Because the underwater transmission of signals is far slower than the propagation of EM waves in terrestrial communication, the Ethernet frame format cannot be directly adopted. For this reason, in light of the existing frame format of terrestrial communication and the features of underwater environment, this paper improves the frame header so that underwater and terrestrial communication protocols can be compatible and interfaced to suit the requirements of the space-air-ground integrated 6G network.

Figure 1 compares our proposed underwater communication frame format with the Ethernet frame format and marks the differences, including:

- Group Number: This column is used to tag the group number of the frame. The details about groups of frames will be given below.
- Next-Hop Node MAC Address: MAC address of the next-hop node.
- Previous-Hop Node MAC Address: MAC address of the previous-hop node.

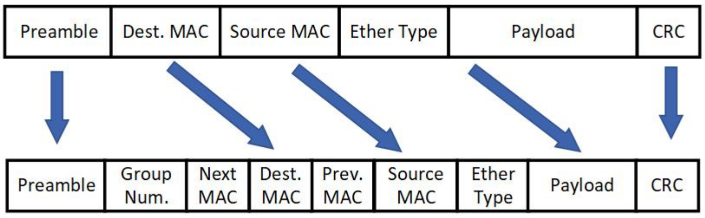


Figure 1. Frame Format Comparison between our proposed method with the Ethernet

The Group Number is used to tag the group number of devices. In our design, this column requires 4 bits, numbered from 00 to 15, among which 01 is allocated to the sink node gateway, 02 is allocated to the bridge nodes and 03-15 are allocated to the normal nodes. Next, we will introduce the use of bridge nodes and normal nodes, and how the two types of nodes process the frames while receiving.

Compared with replacing or recharging the node batteries in terrestrial communication, the power supply to underwater communication nodes is rather difficult. Therefore, one of the major concerns in underwater communication is the power supply and power consumption. To deal with the problem, the proposed method classifies the nodes into different types according to their uses and properties. By designating certain nodes to take responsibility for forwarding traffic, the proposed method can extend the battery life and make a charge last longer so as to help improve the overall network reliability. According to our design, each normal node will stop listening to frames while receiving signals that do not belong to its group, in order to save energy required to analyze the frames once receiving.

● **Sink Node Gateway**– Group 01:

The sink node gateway needs to exchange frames with other communication protocols and is responsible for managing all nodes' communications in the region, making it very busy and consuming a lot of power. Therefore, in our design, the sink node gateway is built on the water surface and is equipped with power supply models like cables or solar panels to provide unlimited power supply. Since the sink node gateway has limitless power sources, it is not especially designed in power-saving design and will always continue to listen to frames.

◆ **Bridge Node** – Group 02:

Bridge nodes need to transfer the frames captured. So, bridge nodes will be built under the water surface and work as nodes in the underwater network. Since bridge nodes are responsible for transferring frames, they need high energy storage capacity to bridge the connections.

◆ **Normal Node** –Group 03-15:

In our design, normal nodes do not transfer frames and therefore consume less power. However, underwater nodes perform different tasks, resulting in different transmission frequencies. An underwater environment monitoring system includes different functional sensors to measure underwater noises, temperature, salinity and so on. Since different sensors have different properties, the required times for transmission are also different. For example, while detecting the changes of sounds, sound monitoring nodes irregularly report the status of sounds to the network, but underwater temperature and salinity sensor may need to report salinity changes every half an hour. Traditionally, while encountering frames bound for sound sensors or from sound sensors, the temperature and salinity sensors need to receive the frames first. Once finding itself not the destination of the frames, the temperature and salinity sensors abandon the frames, leading to unnecessary power consumption. If energy consumption can be reduced, the temperature and salinity sensors will not require such large capacity power storage so as to reduce the manufacturing cost of building a node and extend the lifetime of underwater nodes.

Let us use Fig. 2 as an example and see how the system operates when a frame is transmitted from external network to N1. First, the frame enters the sink node gateway from an external network and the frame is translated into the format for underwater communication. The sink node gateway is its starting point in underwater network. When the frame is transmitted to B2, its group number is changed to 2. Because the group number of the sink node gateway and green normal nodes near B2 is 3, they stop receiving the frame and go in idle status, instead of wasting power in analyzing the frame. If a routing protocol can build a routing table for all bridge nodes and the sink node gateway, including the information of all network nodes, the bridge nodes B2, B3 and B4 in group 2 all have the detailed routing table of the underwater network and the routing table can bind normal nodes and bridge nodes together. When receiving frames from other bridge nodes, each bridge node can search for the destination address to modify the next-hop node MAC address of the header.

For example, while being transmitted from B3 to B4, the frame's destination MAC address is always the MAC address of N1. From the routing table, B3 knows that N1 has been bound with B3 and the next-hop MAC address of the frame is therefore changed to B3. After B3 receives the frame, B3 confirms its destination MAC address and changes the next-hop MAC address to N1 to send the frame. Moreover, when the frame is broadcasted from B3 to B4, B2 receives the frame also and B2 will use the frame to detect if the transmission is completed. After B2 sends out the frame, B2 starts the timer and temporarily saves data. If the frame from B3 is successfully received by B4 within the time counting range, the timer between B2 and B3 is stopped and the data temporarily saved is eliminated. If the frame from B3 to B4 is not detected within the time counting range, the transmission fails. Then, the system reads the temporary data and sends the frame again until it is successfully delivered.

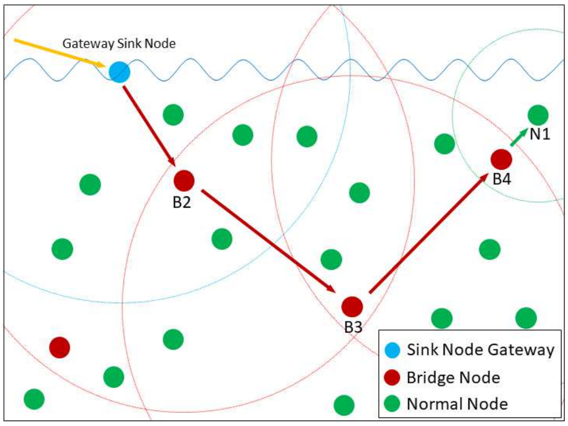


Figure 2. Node placement and Routing path of the frame

III. Experiment Design and Simulation Results

A. Experiment Design

This paper runs a simulation for underwater transmission and places fixed nodes in the simulated underwater environment. In the experimental group, each group of nodes has the following characteristics:

- Sink Node Gateway: has unlimited source of power and analyses every frame.
- Bridge Node: With high capacity power storage, bridge nodes receive and analyze the frames according to the properties of bridge nodes.
- Normal Node: With less capacity power storage, normal nodes receive and analyse the frames according to the properties of normal nodes.

In the control group, the frame receive procedure is conventional: nodes receive the frames and consume all energy to transfer. Nodes are not divided into groups.

- Sink Node Gateway: has unlimited source of power and analyses all frames.
- Bridge Node: With high capacity power storage, bridge nodes analyse all frames.
- Normal Node: with less capacity power storage, normal nodes analyses all frames.

This experiment ignores the power consumption of other systems but nodes in the experimental group need to consume corresponding energy for decision making in addition to fundamental power consumption. Because of different elements in different systems, the power consumption in analysing the same data amount will be different. Therefore, this paper takes consumption counts as unit of measurement.

B. Simulation Result

A simulated environment is created to analyze the performance of MAC frame and protocols. Figure 3 shows the placement of nodes in the simulated environment and compares node power consumption. Table 1 and 2 display the result of ten times of simulations, in which T1, the node in red square, is set as the destination node.

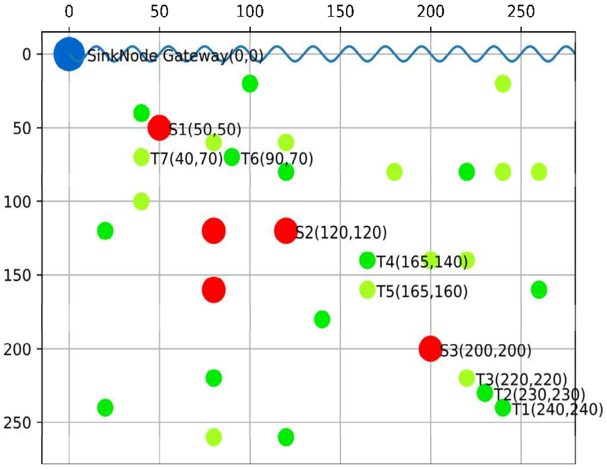


Figure 3. Node placement in the simulated environment

Table1 and 2 show that the proposed new frame format can effectively reduce the power consumption of the nodes. For example, the bridge node S1 is very close to the sink node gateway and therefore receives lots of frames, including those that are destined for S1 and those that are destined for other nodes. In the conventional method, such frames must be abandoned or fully analyzed before being transferred, increasing the power consumption of the bridge nodes. By using our proposed method, S1 obviously saves a lot of energy. As far as normal nodes are concerned, normal nodes that are divided into groups effectively reduces power consumption. Take T7 as an example. T7 is close to the network center where many frames pass by. In the conventional method, T7 is not designated as the receiving node but consumes lots of power. By using our proposed method, nodes can detect and immediately ignore those packets that are bound for other destinations, extending the lifetime of normal nodes in the network.

IV. Conclusions

This paper proposes a solution to improve the frame format header and modify the frame format so that the MAC protocol can be more suitable for underwater communication. By introducing the idea of backbone network and classifying the nodes, energy of the nodes can be efficiently utilized. Thus, the novel method allows underwater and terrestrial communication protocols to be compatible and interfaced and provides a solution for 6G to connect underwater and terrestrial communication.

Acknowledgment

Thanks to the National Science and Technology Council for providing research projects to complete the paper: NSTC 112-2221-E-020-008-MY2。

References

[1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] Jie Zhang, Guangjie Han, Jianfa Sha, Yujie Qian, Jun Liu, "AUV-Assisted Subsea Exploration Method in 6G Enabled Deep Ocean Based on a Cooperative Pac-Men Mechanism", *IEEE Transactions on Intelligent Transportation Systems*, vol.23, issue 2, pp. 1649-1660, Feb.2022.

[3] L. Gupta, R. Jain and G. Vaszkun, "Survey of Important Issues in UAV Communication Networks," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1123-1152, doi: 10.1109/COMST.2015.2495297, Second Quarter 2016.

[4] "IEEE Standard for Information Technology--Telecommunications and Information Exchange between Systems - Local and Metropolitan Area Networks--Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," *IEEE Std 802.11-2020* (Revision of IEEE Std 802.11-2016), vol., no., pp.1-4379, doi: 10.1109/IEEESTD.2021.9363693, 26 Feb. 2021.

[5] Y. Song, "Underwater Acoustic Sensor Networks with Cost Efficiency for Internet of Underwater Things," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 2, pp. 1707-1716, doi: 10.1109/TIE.2020.2970691, Feb. 2021.

[6] M. Jouhari, K. Ibrahimi, H. Tembine and J. Ben-Othman, "Underwater Wireless Sensor Networks: A Survey on Enabling Technologies, Localization Protocols, and Internet of Underwater Things," *IEEE Access*, vol. 7, pp. 96879-96899, doi: 10.1109/ACCESS.2019.2928876, 2019.

[7] A. Darehshoorzadeh and A. Boukerche, "Underwater sensor networks: a new challenge for opportunistic routing protocols," *IEEE Communications Magazine*, vol. 53, no. 11, pp. 98-107, doi: 10.1109/MCOM.2015.7321977, November 2015.

[8] Wilson, W. D., "Equation for the Speed of Sound in Sea Water. *Journal of the Acoustical Society of America*", 32(10), 1357-1357, 1960.

[9] Nurul Huda Mahmood, Hirley Alves, Onel Alcaraz López, Mohammad Shehab, Diana P. Moya Osorio, Matti Latva-Aho,"Six Key Features of Machine Type Communication in 6G", *2nd 6G Wireless Summit (6G SUMMIT 2020)*, Mar. 2020.

[10] Siyuan Zheng, Xiuling Cao, F. Tong, Gangqiang Zhang, Yangze Dong, "Performance Evaluation of Acoustic Network for Underwater Autonomous Vehicle in Confined Spaces", *IEEE 8th International Conference on Underwater System Technology: Theory and Applications* (USYS2018), Dec. 2018.

[11] Antonio Vasilijević; Đula Nađ; Nikola Mišković, "Autonomous Surface Vehicles as Positioning and Communications Satellites for the Marine Operational Environment-Step toward Internet of Underwater Things", *IEEE 8th International Conference on Underwater System Technology: Theory and Applications (USYS2018)*, Dec. 2018.

TABLE 1. RESULTS OF 10 TIMES OF SIMULATION BY USING THE NOVEL FRAME FORMAT

Node	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Group	2	2	2	3	3	4	3	4	3	4
Type	Bridge	Bridge	Bridge	Normal	Normal	Normal	Normal	Normal	Normal	Normal
Initial Power	1M	1M	1M	100K	100K	100K	100K	100K	100K	100K
Residual Power	988180	988250	985730	38730	97800	99340	99240	99760	98800	99640

TABLE 2. RESULTS OF 10 TIMES OF SIMULATION BY USING THE CONVENTIONAL FRAME FORMAT

Node	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Group	2	2	2	3	3	4	3	4	3	4
Type	Bridge	Bridge	Bridge	Normal	Normal	Normal	Normal	Normal	Normal	Normal
Initial Power	1M	1M	1M	100K	100K	100K	100K	100K	100K	100K
Residual Power	588196	885856	911604	98360	95736	92620	88356	86716	51784	27676

Session 1B: Artificial Intelligence 1

Chair: Prof. Jennifer Llovido, Bicol University, Philippines

1 Paper ID: 20240217, 23~28

Test case prioritization with z-Score based neuron coverage

Ms. Hyekyoung Hwang, Prof. Jitae Shin,

Sungkyunkwan University. Korea(South)

2 Paper ID: 20240208, 29~32

Physics-Informed Neural Networks for solving Blood Flows

Prof. Yao-Chung Chang, Prof. Yu-Shan Lin, Dr. Jeu-Jiun Hu,

National Taitung University. Taiwan

3 Paper ID: 20240336, 33~37

Implementation of IoT-based Control System for Maintenance Operation of Long-distance Air Pollution Prevention Device RTO

Prof. DAL-HWAN YOON,

Semyung University. Korea(South)

4 Paper ID: 20240357, 38~42

Search and Recommendation Systems with Metadata Extensions

Mr. Woo-Hyeon Kim, Dr. Joo-Chang Kim,

Kyonggi University. Korea(South)

5 Paper ID: 20240342, 43~49

Utterance-Level Incongruity Learning Network for Multimodal Sarcasm Detection

Dr. Liujing Song, Dr. Zefang Zhao, Prof. Yuxiang Ma, Dr. Yuyang Liu, Prof. Jun Li,

Computer Network Information Center. China

6 Paper ID: 20240043, 50~52

Anomaly Detection During Additive Processes for DLP 3D Printing

Prof. Hyejin S. Kim,

ETRI. Korea(South)

Test case prioritization with z-score based neuron coverage

Hyekyoung Hwang*, Jitae Shin*

* *Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, 16419, Korea*

ristar1234@skku.edu, jtshin@skku.edu

Abstract— **Deep neural networks (DNNs) have been widely used in various applications, such as autonomous driving, healthcare, etc. However, despite achieving high accuracy, DNNs have exhibited quality issues in various aspects such as vulnerability to data corruptions, adversarial attacks, and data dependencies. To ensure the integrity and reliability of these systems, the use of systematic verification and validation methodologies before the deployment of DNN is considered an indispensable technique. Test case prioritization techniques reduce the cost of DNN verification by prioritizing test cases that could induce mispredictions of DNN. In this paper, we propose a neuron coverage-based test case prioritization technique for the DNN classifier that assigns sample priorities based on the ratio of outlier-valued neurons among total neurons in DNN. We evaluate the proposed method with three publicly accessible datasets with different sizes of DNN. The experimental results demonstrate that the proposed method outperforms the existing state-of-the-art neuron coverage-based approach both in error-inducing sample prioritization effectiveness and inference time efficiency.**

Keywords— **Test case prioritization, Neuron coverage, Deep neural network**

I. Introduction

Deep learning (DL) has achieved remarkable accomplishments, showing its potential capabilities in tackling intricate tasks in diverse advanced real-world applications. These include tasks such as image classification [1], natural language processing [2]. In addition, there is a growing demand for the integration of DL models into safety-critical domains such as autonomous driving [3], healthcare [4].

However, despite achieving high accuracy, deep neural networks (DNNs) have exhibited quality issues in various aspects such as vulnerability to data corruptions [5], adversarial attacks [6] and data dependencies (e.g., bias [7]). Real-world incidents, such as autonomous car accidents and incorrect clinical diagnosis, have already emphasized the need for these quality assessment issues. Thus, revealing system problems and fine-tuning are necessities for model deployment in the aspect of DL safety [8].

The conventional DNN evaluation uses a train-test split; one always has corresponding labels to the test set. However, one may attempt to evaluate a DNN with new collections of test sets, which may not have labels, so that it requires additional costs for annotating them.

To mitigate this limitation, researchers have proposed test case prioritization (TCP) techniques [9 - 16], which aim to identify and prioritize instances in an unlabelled test set that could reveal the weakness of the DL model. Since TCP approaches prioritize error-inducing instances within a predefined number of annotation budgets, they effectively reduce the manual cost of labelling. In general, TCP approaches extract features of a DNN model on a training set and define patterns from the features that represent the common behaviour of the trained model. Then, they calculate the similarity between the patterns and a feature of the DNN model on an unlabelled test instance. Some structural coverage criteria are proposed to represent patterns from features of a training set, for example, neuron coverage (NC) [9] and its variants [10]-[14], surprise adequacy [15], mutation analysis [16], and so on.

NC-based TCP methods [9]-[14] are inspired by traditional white-box software tests. They extract the patterns of a training set in terms of neuron activation. From a training set, statistics (e.g., interval of neuron activation values) for all neurons or a threshold for each neuron activation are regarded as the patterns. The patterns are then applied to neuron activation of each unlabelled test input to provide the score for a sample prioritization. In general, TCP techniques give higher priority to samples with a higher score.

FD+ [14] achieves state-of-the-art performance in the NC-based TCP approach by extracting the most specified patterns based on a training set. It collects the output of all neurons over a whole training set and builds the pattern of each neuron for all classes based on exact percentile values. Thus, it requires larger time and storage for the pattern extraction and slower inference time due to its complexity.

To moderate this problem, we propose a neuron-wise threshold based on z-score. Unlike FD+ which uses exact percentile values, we approximate the percentile values based on the z-score. We refer to the proposed method as z-score neuron coverage (ZSNC). The experimental result shows that the proposed ZSNC outperforms the current state-of-the-art NC-based TCP, FD+ [14] in both effectiveness and total time consumption. Specifically, the proposed ZSNC is 7.18 times faster in pattern extraction from a training set and 2.39 times faster in total TCP inference.

The remainder of the paper is organized as follows. Section 2 introduces NC-based TCP techniques and the current state-of-the-art NC-based approach. Section 3 provides the motivation and details of the proposed ZSNC. Section 4 demonstrates the experimental results. Finally, we conclude the paper in Section 5.

II. Related works

Recently, software engineering researchers have proposed several NC-based TCP techniques. The goal of NC-based TCP technique is to effectively identify test instances that reveal errors of a trained DNNs and prioritize the error-inducing samples in predefined number of labelling budget.

A. Test case prioritization based on neuron coverage

Inspired by the white box testing of traditional software coverage testing, [9] was the first to propose NC as a testing metric for DNN. Given a test input x, they define the NC of x as the ratio of the number of *"activated"* neurons to the total number of neurons, denoted as $NC(x, \varepsilon)$. The parameter ε is a threshold that determines the conditions under which neurons in a DNN are considered *"activated"* or not. In general, ε is defined by users and is commonly set to 0. Thus, $NC(x, \varepsilon)$ represents the ratio of the number of neurons whose output values exceed ε to the total number of neurons. The authors in [9] assume that a sample with a higher $NC(x, \varepsilon)$ value is more difficult compared to the lower one. The assumption stems from the fact that a higher NC-valued sample requires the greater number of neuron explorations in a DNN.

Inspired by this work, many DL testing research focus on designing coverage metrics, such as coverage based on different ε settings [10, 11], coverage based on layer level [12], and class-wise familiarity degree [14].

B. Familiarity degree-based Neuron Coverage

The authors of [14] propose an NC-based TCP approach, FD+, that calculates the degree of familiarity based on the valuation pattern of each neuron. FD+ is motivated by the observation that the output values of the neurons in error-inducing samples deviate from the distribution of those of correctly predicted samples. Thus, FD+ extracts class-specific patterns for each neuron as the common patterns of the DNN model for a training set.

Firstly, it collects the activation value of all neurons for all instances of a training set. Based on the collection of activation output, each neuron is classified as "active" or "inactive" for all classes. When a class is given, a neuron is classified as "active" if the majority (e.g., 95%) of the collected neuron output for the given class are greater than zero, otherwise "inactive". Given N number of neurons, $N_{a,k}$ represents the set of "active" neurons and $N_{i,k}$ is the set of "inactive" neurons for class k.

In addition, they introduce different thresholds to $N_{a,k}$ and $N_{i,k}$ to classify the neruons in each set more specifically. A maximal lower bound, $l_{n,k}$, is provided to each of "active" neuron, which is the quantile value of 5% of the output distribution of the "active" neurons. A minimal upper bound, $u_{n,k}$, is provided to each of "inactive" neuron, which is the quantile value of 95% of the output distribution of the "inactive" neurons. When an unlabelled test input and its prediction class k are given, each neuron is classified into one of two classes: "triggered", "inhibited". A neuron is classified as "triggered" when the neuron is "active" for class k and its output value from a given test input is greater than or equal to $l_{n,k}$. A neuron is classified as "inhibited" when the neuron is "inactive" for class k and its output value from a given test input is smaller than or equal to $u_{n,k}$.

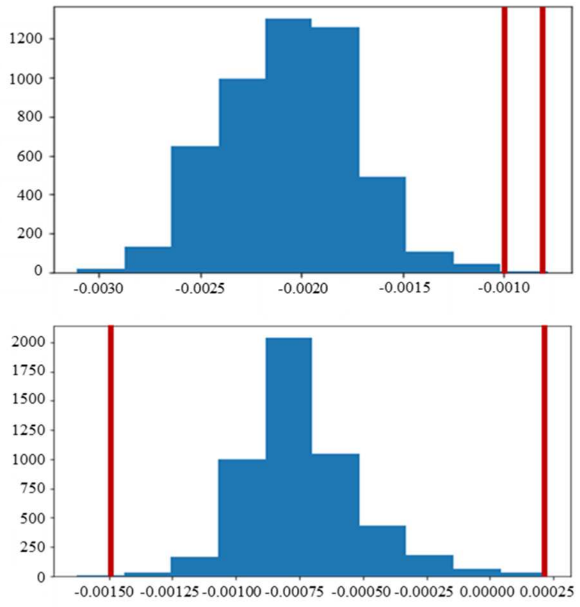


Figure 1 Feature distributions of class "airplane" for 1st, 2nd neurons in 12-th layer of VGG16, trained on CIFAR-10, respectively. The red lines represent the neuron output of samples with its ground truth label "airplane", while their predictions are not "airplane".

Finally, the degree of familiarity (FD +) of an unlabelled test instance is the ratio of "inhibited" or "triggered" neurons in the total N neurons, $(N_t + N_I)/N$, where N_t is the number of "triggered" neuron and N_I is the number of "inhibited" neuron.

However, FD+ requires a higher cost than other NC-based methods. Saving outputs of all neurons for all training instances are required to classify the activeness of each neuron for all classes and calculate the lower/upper bounds. This slows down the pattern extraction process of FD+. Also, comparing each neuron twice makes a longer inference time.

To moderate this challenge, we propose a TCP approach with z-score based interval for each neuron. The proposed method observes the ratio of neurons that do not fall within an interval for a given class. The upper and lower bounds of the interval consist of z-scores of neuron output values of a specific class.

III. Proposed method

A. Motivation

Since various activation functions have been employed, it is inappropriate to distinguish the state of each neuron based on whether the output value of the neuron exceeds 0. Furthermore, similarly to FD+ [14], we observed that the output values of a neuron for incorrect predictions deviate from the distribution of those of correctly predicted samples. Figure 1 displays examples of our observations.

A z-score is the number of standard deviations that is from the mean of a given distribution. Negative z-score indicates the value lies below the mean. Positive z-score indicates the value lies above the mean. Anomaly detection based on z-scores [18], one defines the anomalous sample as the one does not fall within the z-score based interval, is a classic approach and has been used in deep learning to detect outliers in the output of specific layers [19]. Based on the aforementioned observation and fact, we set a common pattern of each neuron for a training set as an interval based on the z-scores.

B. Z-Score based Neuron Coverage (ZSNC)

We assume that there are N neurons in a DNN classifier $f(.)$ with total C classes. We denote a set of outputs of N neurons comprising $f(.)$ for an input x as $N_{f(x)}$, and we have a total of T training instances. The total T instances are divided into C number of subsets, $T = \{T_1, T_2, \dots, T_C\}$, where each T_i is a subset of the training set based on the ground truth class i.

The pattern extraction process of FD+ [14] can be formulated as follows: 1) Collecting the $N_{f(T)}$ and separate it to C subsets based on the ground truth classes $N_{f(T)} = \{N_{f(T_1)}, ..., N_{f(T_C)}\}$. 2) Classifying each neuron into “active” or “inactive” based on $N_{f(T_i)}$. 3) Extracting values at the a% and (100-a)% percentile from each of $N_{f(T_i)}$ to determine class-specific lower and upper bounds.

However, the pattern extraction process of the proposed method does not require the first two stages of that of FD+. Also, we approximate the class-specific lower and upper bounds for each neuron based on the mean and standard deviation of neuron outputs for each class.

Given a t-th training instance x_t with corresponding label c, we iteratively calculate the mean and standard deviation of N neurons for class c as follows:

$$\mu_t^c = \mu_{t-1}^c + \tfrac{1}{t}\left(N_{f(x_t)} - \mu_{t-1}^c\right), \tag{1}$$

$$(\sigma_t^c)^2 = \frac{t-2}{t-1}(\sigma_{t-1}^c)^2 + \frac{1}{t}\left(N_{f(x_t)} - \mu_{t-1}^c\right)^2. \tag{2}$$

Here, μ_t^c is a set that consists of the mean of t-th iteration for class c and σ_t^c is a set that contains the standard deviation of t-th iteration for class c. Both μ_t^c and σ_t^c consists of N elements that each of them implies the mean and standard deviation of each neuron in t-th iteration for class c, respectively.

When the iteration is completed for all classes, C sets of mean $\{\mu^1, \dots, \mu^C\}$ and standard deviation $\{\sigma^1, \dots, \sigma^C\}$ are collected. For simplicity, we omit the number of iterations for each class. The mean and standard deviation of a n-th neuron for class i is denoted as μ_n^i and σ_n^i, respectively. Then, we set an interval for n-th neuron for a class i as $[\mu_n^i - \epsilon \cdot \sigma_n^i, \mu_n^i + \epsilon \cdot \sigma_n^i]$, where ϵ is a scaling factor and selected from the z-score table.

Finally, when an unlabelled test input and its prediction result k is given, we calculate the ratio of neurons whose output value does not fall within the range of the interval $[\mu_n^k - \epsilon \cdot \sigma_n^k$, $\mu_n^k + \epsilon \cdot \sigma_n^k]$. The proposed z-score-based neuron coverage is

$$ZSNC(x) = \frac{\left|\{n \mid n \in N,\ f_n(x) \notin [\,\mu_n^k - \epsilon \cdot \sigma_n^k\,, \mu_n^k + \epsilon \cdot \sigma_n^k]\right|}{N} \tag{3}$$

where $f_n(x)$ is the output of n-th neuron on a test input x and $|\,.\,|$ represents a cardinality of a set.

For training set pattern extraction, [14] have to store $N \times T$ values and then classify the active status and calculate the bounds of intervals. In contrast, our proposed method calculates the mean and standard deviations iteratively so that it does not require separate storage spaces. During an inference, the proposed method does not classify each neuron with ‘active’/‘inactive’ status so that it has leading to reduced inference time.

IV. Experimental Results

In this section, we describe the experimental setups and analyse the experimental results.

A. Datasets and Models

To evaluate our method, we adopt three datasets for image classification benchmark: SVHN [20], CIFAR10 and 100 [21]. Additionally, to evaluate the performance of TCP in real-world scenarios, we employ real-world corruptions [22] to CIFAR10 dataset, which is notated with ‘C’ after the dataset. We choose 19 different corruptions such as motion blur, brightness, shot noise, etc.

We tested the proposed method on different sizes of models: VGG16 [23], ResNet50 [24], WideResNet50_2 (WRes50_2) [25], MobileNetV2 (MobileV2) [26], and EfficientNetV2_S (EfficV2S) [27].

Table 1 represents details about our evaluation setups. The dataset and models, classification accuracy of the model on training and test data, and the number of training and test data are shown. For simplicity, we describe each experiment as their ID. The experiment IDs F to H represents CIFAR10 with real-world corruptions, while the trained model is same with the experiment IDs B to D.

TABLE 1 EXPERIMENTAL SETUPS FOR EVALUATION

ID	Data-Model	Train Acc. (%)	Test Acc. (%)	# Train	# Test
A	SVHN-WRes50_2	94.01	92.38	73,257	26,032
B	CIFAR10-VGG16	99.97	93.81	60,000	10,000
C	CIFAR10-MobileV2	99.51	93.38	60,000	10,000
D	CIFAR10-ResNet50	98.69	93.78	60,000	10,000
E	CIFAR100-EfficV2S	99.52	88.20	60,000	10,000
F	CIFAR10C - VGG16	-	29.56	-	190,000
G	CIFAR10C - MobileV2	-	26.91	-	190,000
H	CIFAR10C - ResNet50	-	29.71	-	190,000

TABLE 2 COMPARISON OF ATPF (%) WITH DIFFERENT NC-BASED TCP TECHNIQUES

ID	NBC [10]	NLC [12]	FD+ [14]	ZSNC
A	23.381	23.296	72.693	71.121
B	59.579	69.21	48.962	83.631
C	51.835	34.818	70.376	78.265
D	56.071	40.109	35.464	86.905
E	10.616	12.264	28.81	30.447
F	72.677	59.827	73.664	84.343
G	76.231	77.296	85.98	83.412
H	71.18	70.158	66.831	86.96
Avg.	52.696	48.372	60.348	75.511

B. Compared Methods

We compared the proposed ZSNC approach with the three existing Neuron Coverage-based TCP methods.

Neuron boundary coverage (NBC) [11] targets to observe the number of neurons that does not lies between lower and upper bound of the outputs of a neuron upon training set. The lower/upper bound is set to be the minimum/maximum values of the total neuron outputs upon the training set. The proposed ZSNC differs from NBC in that it measures class-specific boundaries for each neuron and replaces lower/upper boundaries with z-scores.

Neural coverage (NLC) [12] defines the coverage criteria as the covariance of layer output. It supposes that a sample with higher covariance between the layer outputs leads more diversity to test set. It does not requires pattern extraction from training set.

Familiarity degree (FD+) [14] is described in Section 2, one requires two steps to determine the status (inhibited, triggered) of each neuron. FD+ requires saving neuron outputs for the entire training data separately for setting the lower/upper bound thresholds to distinguish the status of neuron. In contrast, the proposed ZSNC is single-step and does not require storing neuron outputs for the entire training data. According to the setup in [14], we set maximal lower bound for active neuron as the 5%, while the minimal upper bound is 95% of the values.

We set the scaling factor for proposed method as 1.96, which implies the 95% of confidence intervals in z-score table.

C. Evaluation Metric

For evaluation of **effectiveness**, we used the Average Test Percentage of Fault (ATPF) [17] to measure the error-revealing identification performance of TCPs. ATPF is the average of the ratios of the number of erroneous instances detected within the n-th labelling budget:

$$\text{ATPF}(\%) = 100 \cdot \frac{1}{N_{fail}} \sum_{n=1}^{N_{fail}} \frac{N_{err,n}}{n} \qquad (4)$$

where N_{fail} is the total number of erroneous instances in a test set and $N_{err,n}$ is the number of error-inducing instances detected in the labelling budget, n.

To evaluate an **efficiency** of proposed method, we measure the time required for the pattern extraction from and the total inference time of NC-based TCP techniques.

D. Evaluation Result: Effectiveness

Table 2 shows the comparison of ATPF with different NC-based TCP techniques. The final row displays the average ATPF across all experiments. A higher ATPF value implies a more effective identification of error-inducing samples within the same labelling budget. The proposed ZSNC takes the first place for five among seven experiments.

Even though the NLC is the most recent approach, the performance of NLC is less than the others. It is because the initial purpose of NLC is to compare the diversity between the different sets, not to find error-inducing samples. The comparison between NBC and the proposed ZSNC shows that employing class-specific threshold for each neuron can lead the meaningful performance enhancement based on the more detailed patterns from the training set. A comparison of FD+ with the proposed ZSNC reveals the superiority of single staged TCP and z-score-based intervals over the sampling of exact quantile samples.

E. Evaluation Result: Efficiency

Table 3 displays the time for pattern extraction for each neuron from a training set, in seconds. The bottom row displays the average time across all experiments. Experiment IDs F to H are omitted because they are designed to observe inferences on real-world corruptions, using the same model with experiment IDs B to D.

NLC is marked with '-' in the table because it does not require the pattern extraction from the training set. Among the other NC-based TCP methods, NBC takes the shortest pattern extraction time because it extracts only the minimum and maximum values for each neuron from the entire training set. But, the ATPF value of NBC is always smaller than that of the proposed method and FD+.

The proposed ZSNC takes the second fastest pattern extraction as it extracts class-specific thresholds for each neuron iteratively. FD+ takes the longest time since it requires the storage for saving the activation values of all neurons on a whole training set and extracts exact percentile values as their bounds of intervals. The proposed method takes about three times longer on average than NBC but is approximately 7.19 times faster compared to the latest algorithm, FD+.

TABLE 3 COMPARISON OF TIME (SEC.) REQUIRED FOR PATTERN EXTRACTION OF TRAINING SET.

ID	NBC [10]	NLC [12]	FD+ [14]	ZSNC
A	47.61	-	1463.162	149.39
B	7.604	-	244.659	23.95
C	17.423	-	402.778	51.266
D	20.094	-	561.45	59.297
E	135.924	-	2311.569	409.675
Avg.	45.731	-	996.724	138.716

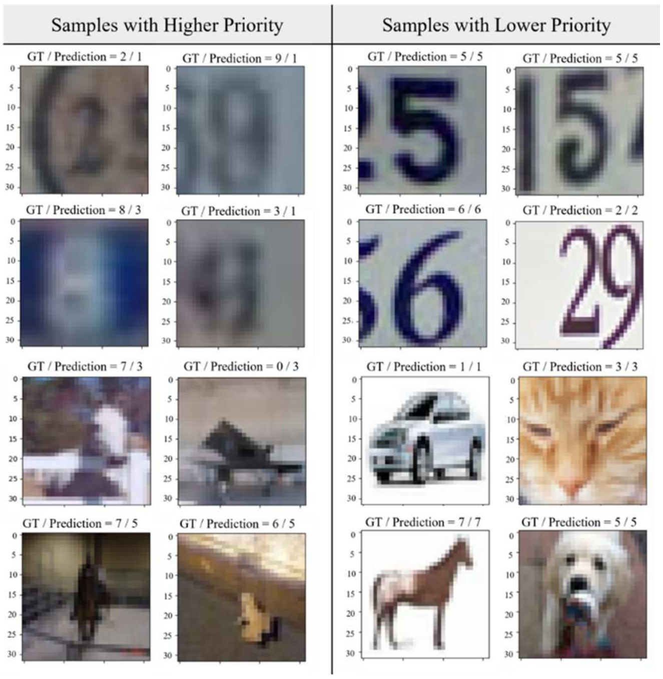


Figure 2 Sample visualizations with higher/lower priority based on proposed ZSNC. (Left) samples with high priority based on ZSNC, (Right) samples with low priority based on ZSNC.

TABLE 4 COMPARISON OF THE TOTAL INFERENCE TIME (SEC.)

ID	NBC [10]	NLC [12]	FD+ [14]	ZSNC
A	15.946	43.762	73.357	26.345
B	1.756	12.019	9.855	5.523
C	2.801	5.965	23.839	5.547
D	3.809	11.856	25.763	7.329
E	27.156	66.669	90.85	36.869
F	19.705	118.228	80.027	45.69
G	44.741	71.141	215.173	72.655
H	55.218	167.037	235.27	106.947
Avg.	21.392	62.085	94.267	38.363

Table 4 represents a comparison of total inference time of different NC-based TCP techniques for seven different classification experiments in seconds. The bottom row displays the average across all experiments.

Consistent with the results in Table 3, NBC enables the fastest inference. The proposed method achieves the second fastest total inference time, by exhibiting 1.79 times longer than NBC in average. However, the most recent algorithm [12] and the state-of-the-art in NC-based TCP [14] takes 1.58 times and 2.46 times longer than the proposed method, respectively. Furthermore, the proposed method outperforms the other NC-based TCP methods in error-inducing sample detection performance, making it the most proper option among the NC-based TCP methods.

F. *Sample Visualization*

Figure 2 shows the samples with higher/lower priority based on the proposed ZSNC, respectively. Samples that received a high priority by the proposed method tend to exhibit hard conditions such as indistinct differentiation from the background, low lighting conditions, or misalignment of objects. On the other hand, samples with lower priorities present easy conditions such as clear distinction between the background and the object, distinctive characteristics of prediction class, such as animal faces, etc.

V. Conclusions

As the importance of thorough testing before deploying DL models has increased due to higher safety standards, test case prioritization has emerged as an effective solution by prioritizing error-inducing samples for testing, thereby reducing testing costs. This paper introduces a Z-score-based outlier detection approach to existing neuron coverage-based test case prioritization. The proposed method demonstrates superior error-inducing sample detection capabilities compared to the state-of-the-art NC-based methods across various experimental settings, and it also significantly reduces both pattern extraction and total inference time, performing approximately 7.19 and 2.46 times faster, respectively.

Acknowledgment

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government Ministry of Science and ICT (MSIT) (NRF-2020R1F1A1065626), in part by the MSIT under the Information Technology Research Center (ITRC) support program (IITP- 2023-2018-0-01798) supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP), and partly supported by the BK21 FOUR Project.

References

[1] K. B. Obaid, S. Zeebaree & O. M. Ahmed, “Deep learning models based on image classification: a review*”. International Journal of Science and Business*, vol. 4(11), pp. 75-81, October 2020.

[2] D. W. Otter, J.R. Medina, & J.K. Kalita, “A survey of the usages of deep learning for natural language processing”. *IEEE Trans. on neural networks and learning systems*, vol. 32(2), pp. 604-624, April 2020.

[3] Almalioglu, Y., Turan, M., Trigoni, N., & Markham, A. “Deep learning-based robust positioning for all-weather autonomous driving.” *Nature Machine Intelligence*, vol. 4(9), pp. 749-760, September 2022.

[4] A. Esteva, K. Chou, S. Yeung, N. Naik, et al. Deep learning-enabled medical computer vision. *NPJ digital medicine*, vol. 4(1), 5, January 2021.

[5] E. Rojas, D. Pérez & E. Meneses. “Exploring the effects of silent data corruption in distributed deep learning training”. in Proc. SBAC-PAD’22, 2022, pp. 21-30)

[6] N. Akhtar, & A. Mian, “Threat of adversarial attacks on deep learning in computer vision: A survey”. *Ieee Access*, vol. 6, pp. 14410-14430, February 2018.

[7] G. Vardi, G. “On the implicit bias in deep-learning algorithms”. *Communications of the ACM*, vol. 66(6), pp. 86-93, June 2023.

[8] F. Tambon, F, G. Laberge, L. An, A. Nikanjam, P. S. N Mindom, Y. Pequignot, ... & F. Laviolette, “How to certify machine learning based safety-critical systems? A systematic literature review.” *Automated Software Engineering*, vol. 29(2), pp. 38, April 2022.

[9] K. Pei, Y. Cao, J. Yang, & S. Jana., "Deepxplore: Automated whitebox testing of deep learning systems". in Proc. SOSP'17, 2017, pp. 1-18.

[10] Y. Tian, K. Pei, K., S. Jana, & B. Ray, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars". in Proc. ICSE'18, 2018, pp. 303-314.

[11] L. Ma, F. Juefei-Xu, F. Zhang, J. Sun, M. Xue, B. Li & Y. Wang, (2018, September). "Deepgauge: Multi-granularity testing criteria for deep learning systems". in Proc. ASE'18, 2018, pp. 120-131.

[12] Y. Yuan, Q. Pang, & S. Wang, "Revisiting neuron coverage for dnn testing: A layer-wise and distribution-aware criterion". in Proc. ICSE'23, 2023, pp. 1200-1212

[13] X. Xie, L. Ma, F. Juefei-Xu, M. Xue, H. Chen, Y. Liu, ... & S. See, "Deephunter: a coverage-guided fuzz testing framework for deep neural networks". in Proc. ISSTA'19, 2019, pp. 146-157.

[14] R. Yan, Y. Chen, H. Gao, & J. Yan, "Test case prioritization with neuron valuation based pattern". *Science of Computer Programming*, vol. 215, pp. 102761, March 2022.

[15] S. Kim, & S. Yoo, "Multimodal surprise adequacy analysis of inputs for natural language processing DNN models.' in Proc. AST'21, 2021, pp. 80-89.

[16] Z. Wang, H. You, J. Chen, Y. Zhang, X. Dong, & W. Zhang, "Prioritizing test inputs for deep neural networks via mutation analysis". in Proc. ICSE'21, 2021, pp. 397-409.

[17] Y. Li, M. Li, Q. Lai, Y. Liu, & Q. Xu, "Testrank: Bringing order into unlabeled test instances for deep learning tasks". in Proc. NIPS'21, 2021 pp. 20874-20886.

[18] P. R. Bushel, H. K. Hamadeh, L. Bennett, J. Green, A. Ableson, S. Misener, ... & R. S. Paules, "Computational selection of distinct class- and subclass-specific gene expression signatures". *Journal of biomedical informatics*, vol. 35(3), pp. 160-170, June 2002.

[19] V. Aggarwal, V. Gupta, P. Singh, K. Sharma & N. Sharma, (2019, April). "Detection of spatial outlier by using improved Z-score test". in Proc. ICOEI'19, 2019, pp. 788-790.

[20] N. Yuval, W. Tao, C. Adam, B. Alessandro, W. Bo, Y. Ng. Andrew, "Reading Digits in Natural Images with Unsupervised Feature Learning" *NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011*.

[21] A. Krizhevsky, "Learning multiple layers of features from tiny images", M. Eng. thesis, University of Tront, 2009.

[22] D. Hendrycks, & T. Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". in Proc. ICLR'18, 2018

[23] K. Simonyan & A. Zisserman, " Very deep convolutional networks for large-scale image recognition." ICLR'15, 2015

[24] K. He, X. Zhang, S. Ren, J. Sun, "Identity mappings in deep residual networks". in Proc. ECCV'16, 2016, pp. 630-645.

[25] S. Zagoruyko, & N. Komodakis, "Wide Residual Networks". in Proc. BMVC'16, 2016

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks". in Proc.CVPR'18, 2018, pp. 4510-4520.

[27] M. Tan, Q. Le. "Efficientnetv2: Smaller models and faster training". in Proc. PMLR'21, 2021, pp. 10096-10106.

Hyekyoung Hwang is a Ph.D student and she received her BS degree from the Department of Electronic and Electrical Engineering, College of Information and Communication Engineering and mathematics, College of Natural Science, Sungkyunkwan University, Republic of Korea in 2018. Her research interests include image processing and deep learning, focused on explainable AI with uncertainty estimation.

Jitae Shin has been a professor from 2002 with the School of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon, South Korea. He received the B.S. degree from Seoul National University, in 1986, the M.S. degree from the Korea Advanced Institute of Science and Technology, in 1988, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, USA, in 1998 and 2001, respectively. For former industrial experiences, he worked with Korea Electric Power Corporation and the Korea Atomic Energy Research Institute from 1988 to 1996. His current research interests include image/video signal processing using deep learning, medical image processing, and video transmission over wireless/mobile communication systems.

Physics-Informed Neural Networks for solving Blood Flows

*Yao-Chung Chang, **Yu-Shan Lin, ***Jeu-Jiun Hu

*,*** Department of Computer Science and Information Engineering,
** Department of Information Science and Management Systems,
National Taitung University, 369, Sec. 2, University Rd., Taitung, Taiwan
ycc@nttu.edu.tw, ysl@nttu.edu.tw, jjhu@nttu.edu.tw

***Abstract*—The Physics-informed Neural Networks Deep Learning (PINN) framework has been introduced with the primary objective of advancing the field of blood flow simulations. PINN Deep Learning involves data-driven training for flow prediction and can incorporate the understanding of physical laws described by partial differential equations (PDEs). This paper employs the PINN Method for simulating blood flows. Multiple test cases will be computed and compared with other numerical and experimental results to validate the approach. The results demonstrate that the PINN method functions as expected, and validation against experimental and other researchers' results ensures the generation of meaningful output data and the prudent selection of parameters.**

***Keywords*—Physics-informed Neural Networks, Deep Learning, Cardiovascular System, 1D Blood Model.**

I. Introduction

The cardiovascular system, also known as the circulatory system, is composed of the heart, blood vessels (arteries, veins, capillaries), and blood. Representing it mathematically and analyzing its dynamic mechanisms is still a challenging research. Traditionally, the computation of cardiovascular system flow fields involved the solution of the three-dimensional Navier-Stokes equations and solved it numerically. However, this is time consuming and wastes a lot of computational resources.

To simplify this problem, certain assumptions can be made, reducing it to a one-dimensional problem., thus, the axisymmetric Navier-Stokes equations [1][2]. Axisymmetric equations may have a single point where the radius is zero, making numerical solutions singularity. Huang [3] applied the TVD Scheme to solve cardiovascular systems, simplifying the dynamics of cardiovascular flow fields into a one-dimensional model. This model is employed to study and comprehend the characteristics of cardiovascular flow dynamics and pressure propagation. It simplifies blood vessels into one-dimensional tube flow problems, allowing for the calculation of flow rates, blood vessel cross-sectional areas, and lengths to further understand and predict pressure distribution within blood vessels. . Consequently, scholars have started exploring the feasibility of solving the one-dimensional Navier-Stokes equations [4][5].

In recent years, the development of Artificial Intelligence (AI) has been one of the hottest topics. Physics-informed Neural Networks (PINN), proposed by scholar M. Raissi[6], are a type of neural network that uses physical equations as computational constraints for solving partial differential equations[6]. They belong to the supervised learning category. This network utilizes data-driven training, learning through loss functions described by partial differential equations. During the training process, PINN benefits from the constraints imposed by physical equations, enabling it to learn fundamental physical laws such as mass and momentum conservation. In recent years, PINN has gradually become a focal point in machine learning and computational fluid dynamics, achieving significant progress in both theory and applications [7][8][9].

To enhance the performance of computation of cardiovascular system flow fields, the systems can be treated as 1D tube like structures and PINN method is adopted for simulations. Further, we solve several preambles to demonstrate obtaining the results accuracy and efficiency.

II. METHODS

2.1 One-dimensional formulation

Blood flow through arteries treats as 1D tube flow is governed by the Navier-Stokes equations in 1D-cylinder coordinate systems:

$$\frac{\partial A}{\partial t}+\frac{\partial q}{\partial x}=0 \tag{1}$$

$$\frac{\partial q}{\partial t}+\frac{\partial}{\partial x}\left(\alpha\frac{q^2}{A}\right)+\frac{A}{\rho}\frac{\partial p}{\partial x}+\frac{f}{\rho}=0 \tag{2}$$

$$p-p_{ext}-\beta(\sqrt{A}-\sqrt{A_0})=0 \tag{3}$$

where A(x, t) represents the cross-sectional area of the blood vessel, which is a function of spatial position x and time t. q(x, t) denotes the volumetric flow rate within the blood vessel, while p(x, t) represents the average pressure of the fluid within the blood vessel. The fluid density, denoted as ρ, is assumed to have a default value of 1050 kg/m^3. Additionally, the frictional force, f(x, t), experienced per unit length within the blood vessel, can be expressed as f = βπμu, where μ = 0.0045

Pa/s and u represents the fluid velocity. For the Poiseuille flow, the fluid, β takes a value of 8 and α=1 for the uniform velocity distribution assuming. The pressure p(x, t) in the above equations can be described using the following equation:

$$p(x,t) = p_{ext}(x,t) + \Psi(x,t) \quad (4)$$

where $p_{ext}(x,t)$ is the external pressure acting on the fluid and $\Psi(x,t)$ can be represented as the follows:

$$\Psi(x,t) = K(x)\phi\big(A(x,t), A_0(x)\big) + p_0 \quad (5)$$

In the above equation, *K(x)* is a positive function that depends on the Young's modulus *E(x)* and the thickness *h(x)* of the blood vessel wall. P_0 represents the reference pressure at the blood vessel cross-sectional area A_0.

$$\phi(x,t) = \left(\frac{A(x,t)}{A_0(x)}\right)^m - \left(\frac{A(x,t)}{A_0(x)}\right)^n \quad (6)$$

where the values of m and n can be determined experimentally. For arteries, the values are (m, n) = (0.5, 0), and for veins, the values are (m, n) = (10, -1.5) [4].

2.2 Physics-informed Neural Networks

Consider a general form of partial differential equation as an example, assuming that the function u = u(t, x) and satisfies a partial differential equation in the following form:

$$u_t + N[u;\lambda] = 0,\ x \in \Omega,\ t \in [0,T] \quad (7)$$

where *N[u;λ]* represents a parameterized non-linear operator that operates on u with parameter λ. Ω denotes the Euclidean geometric space, and T is the terminal time. Given the initial state, boundary conditions, and physical parameter λ for u(t, x), solving the partial differential equation allows us to predict the values of u(t, x) at any spatiotemporal point. Numerical methods such as finite differences or finite elements can be employed for solving. First, let's define:

$$f : u_t + N[u;\lambda] \quad (8)$$

Subsequently, a neural network can be constructed to approximate the solution to the partial differential equation. Therefore, the root mean square loss function is defined as following:

$$MSE = MSE_u + MSE_f \quad (9)$$

$$MSE_u = \frac{1}{N}\sum_{i=1}^{N_u}\left|u(t_u^i, x_u^i) - u^i\right|^2 \quad (10)$$

$$MSE_f = \frac{1}{N}\sum_{i=1}^{N_u}\left|f(t_f^i, x_f^i)\right|^2 \quad (11)$$

the data-driven of the loss function, acquired through the initial and boundary condition in the training data, represents the physics model-driven portion of the loss function, which is trained in the system's machine learning process to approximate the neural network function. The network must not only minimize the error with known training data but also satisfy the constraints of the partial differential equation as much as possible. Therefore, the network structure and loss function of PINN need to be customized according to the form of the partial differential equation, while also maintaining the model's scalability.

The architecture of the PINN in this study is shown in Figure 1. After inputting temporal and spatial data, it initially employs a fully connected neural network to approximate the function, subsequently obtaining residuals for the partial differential equation and initial/boundary value constraints. These residuals are then incorporated as regularization terms in the loss function. Finally, optimization algorithms like gradient descent are used to obtain the neural network's connection weights and the physical parameters of the partial differential equation.

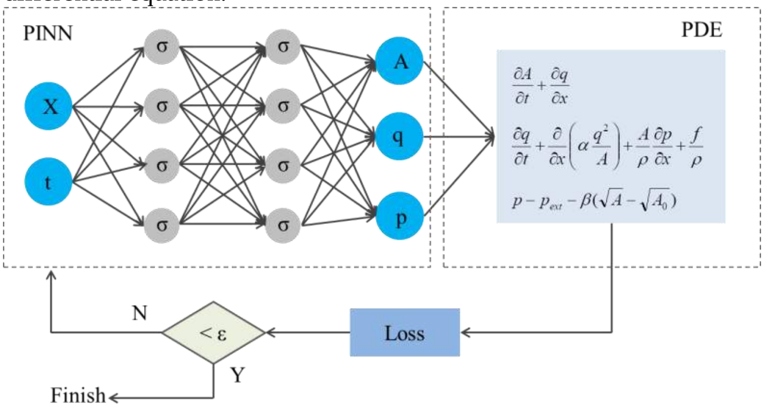

PINN
X
t
σ
A
q
p
PDE
$\frac{\partial A}{\partial t} + \frac{\partial q}{\partial x}$
$\frac{\partial q}{\partial t} + \frac{\partial}{\partial x}\left(\alpha \frac{q^2}{A}\right) + \frac{A}{\rho}\frac{\partial p}{\partial x} + \frac{f}{\rho}$
$p - p_{ext} - \beta(\sqrt{A} - \sqrt{A_0})$
N
< ε
Loss
Y
Finish

Figure 1. A schematic representation of the proposed PINN.

III.RESULTS

An arterial tree with branches：

The main objective of this test is to evaluate the capability and efficiency of the developed numerical method in computing bifurcated tube calculations. The geometry of a bifurcation blood flows shown in figure 2.

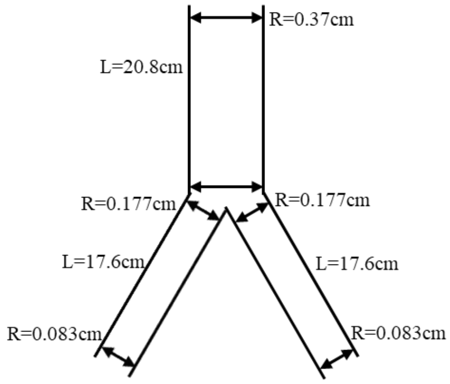


Figure 2. Geometry of a bifurcation blood flows

The input velocity profile was obtained from [1] by curve fit as shown in figure. The main vessel was composed with uniform slender cylinder and has two branch vessel connected to the end. The branch vessel has a linearly tapered cylinder with varying radius as follows:

$$r_0(z)=R_0 exp(log(R_d/R_u)(z/L)) \quad (15)$$

where $r_0(z)$ radius of a vessel in a cylinder coordinate system. The elastic property of vessel is calculated from Young’s modulus:

$$Eh/r_0 = k_1\ exp(k_2 r_0) + k_3 \quad (16)$$

Where h is vessel wall thickness, r_0 is vessel radius and the constant k_1=2×10^7 g.cm^{-1}.s^{-2}, k_2 = -22.53 cm^{-1} and k_3 = 8.65× 10^5 g.cm^{-1}.s^{-2} are adopted in simulation.

The inflow at the inlet of common carotid artery(CCA) is given by the a periodic pulse in [11] as shown in figure 3.. Time period *T = 0.917s* is the period of one pulse cycle and the kinematic viscosity is ν = 0.046 cm^{2}s^{-1}. The bifurcation occurs at a point where the outflow from the CCA is balanced with the inflow from the ICA and the external carotid artery (ECA). The geometry of the CCA, ICA and ECA branches is summarized in Table 1.

Table 1 Geometry parameter of CCA, ICA and ECA

	r_u	r_d	L
CCA	0.370 cm	0.370 cm	20.8 cm
ICA	0.177 cm	0.083 cm	17.6 cm
ECA	0.177 cm	0.083 cm	17.7 cm

The problem is addressed by utilizing neural networks with seven hidden layers, each consisting of 100 neurons, and employing a hyperbolic tangent activation function. The pressure waveform at various locations in a bifurcation vessel, obtained through the application of the PINN method to solve the 1-D equations, is illustrated in Figure 4. Specifically, the plots are at a location 6 cm from the common carotid artery (CCA), internal carotid artery (ICA), and external carotid artery (ECA). These waveforms closely resemble the corresponding ones calculated by solving the Navier-Stokes equations, considering compatible geometrical and mechanical properties, as well as identical inflow and outflow boundary conditions [12]. A noticeable increase in pressure gradient discrepancy is observed, attributed to the tapered shape of the vessel.

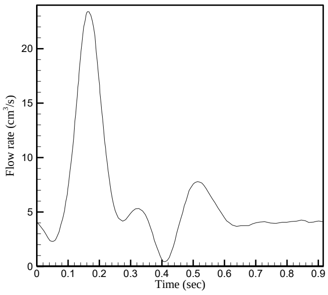
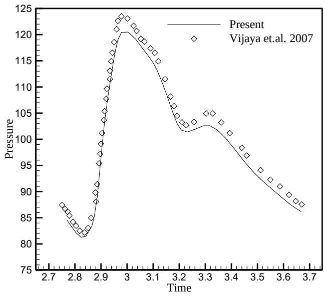
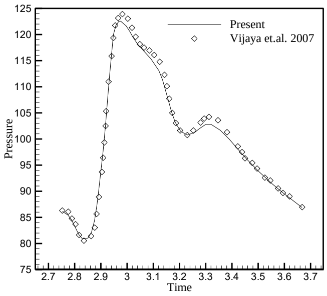


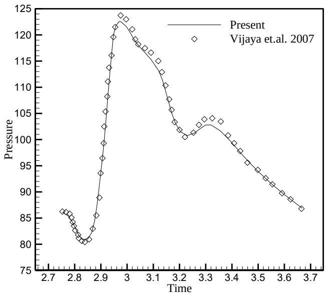
Figure 3. Inflow rate as a function of time.



(a) Common carotid artery



(b) External carotid artery



(c) Internal carotid artery

Figure 4. Pressure waveform with time.

IV. CONCLUDIONS

In this study, the PINN method is employed to simulate one-dimensional blood flows. Simulations are conducted using the hyperbolic mathematical model of one-dimensional Navier-Stokes equations in cylindrical coordinate systems. Several test cases are employed to validate the chosen approach. The proposed PINN method's results are compared to those in [12] for the common carotid artery bifurcation. The results demonstrate that the PINN method functions as expected, and validation against experimental and other researchers' results ensures the generation of meaningful output data and the judicious selection of parameters. The validation of larger arterial networks with multiple bifurcation levels has not been conducted yet, but is scheduled for future research. Demonstrations indicate that considering bifurcation pressure drops can significantly enhance accuracy. Consequently, achieving comparable accuracy in blood flow solutions for 1D models using the PINN method compared to traditional numerical methods for 2D or 3D models is feasible.

REFERENCES

[1] Kolachalama, V, et al.,“Predictive Haemodynamics in a One-Dimensional Carotid Artery Bifurcation. Part I Application to Stent Design”. In: IEEE Transactions on Biomedical Engineering, 54(5), pp. 802–812. 2007.

[2] Muller LO, Toro EF. Well-balanced high-order solver for blood flow in networks of vessel with variable properties. International Journal for Numerical Methods in Biomedical Engineering, 29:1388–1411, 2013

[3] P. G. Huang,† and L. O. Muller, Simulation of one-dimensional blood flow in networks of human vessels using a novel TVD scheme, International Journal For Numerical Methods In Biomedical Engineering Int. J. Numer. Meth. Biomed. Engng. 2015.

[4] N. Stergiopulos, D. F. Young and T. R. Rowe,"Computer simulation of arterial flow with applications to arterial and aortic stenoses", J. Biomechanics Vol. 25, No. 12, pp. 1477-1488. 1992.

[5] Lucas O. Muller, Carlos Pares, Eleuterio F. Toro, "Well-balanced high-order numerical schemes for one-dimensional blood flow in vessels with varying mechanical properties", Journal of Computational Physics 242 53–85, 2013.

[6] Maziar Raissi,et al, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations", Journal of Computational Physics 378, 686–707, 2019.

[7] Sebastián Cedillo1, et al, "Physics-Informed Neural Network water surface predictability for 1D steady-state open channel cases with different flow types and complex bed profile shapes", Advanced Model and Simulation in Engineering Sciences. 9:10, 2022https://doi.org/10.1186/s40323-022-00226-8

[8] Pao-HsiungChiu, et al, "CAN-PINN: A fast physics-informed neural network based on coupled-automatic–numerical differentiation method", Computer Methods in Applied Mechanics and Engineering Volume 395, 15 May 2022.

[9] Jun Pu, et al, "PINN-Based Method for Predicting Flow Field Distribution of the Tight Reservoir after Fracturing", Geofluids, vol. 2022, Article ID 1781388, 10 pages, 2022. https://doi.org/10.1155/2022/1781388

[10] S.J. SHERWIN, V. FRANKE, J. PEIRÓ and K. PARKER, "One-dimensional modelling of a vascular network in space-time variables", Journal of Engineering Mathematics 47: 217–250, 2003.

[11] D. W. Holdsworth, C. J. D. Norley, R. Frayne, D. A. Steinman, and B. K. Rutt, "Characterization of common carotid artery blood-flow waveforms in normal human subjects," Physiological Meas., vol. 20, pp. 219–240, 1999.

[12] V. Kolachalama et al. "Predictive Haemodynamics in a One-Dimensional Carotid Artery Bifurcation. Part I: Application to Stent Design". In: IEEETransactions on Biomedical Engineering 54.5 (2007), pp. 802-812.

Yao-Chung Chang (M'03) received the Ph.D. degree from National Dong Hwa University, Hualien, Taiwan, in 2006.

He is a Professor of the Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan. His primary research interests include intelligent communication System, IoT, and cloud computing.

Dr. Chang is a recipient of the subsidization program in universities for encouraging exceptional talent, Ministry of Science and Technology, Taiwan

Yu-Shan Lin received the Ph.D. degree from National Sun Yat-sen University, Kaohsiung, Taiwan, in 2006.

She is a Professor of the Department of Information Science and Management Systems, National Taitung University, Taitung, Taiwan. Her research interesting areas include Digital Learning, Information Technology Education, Marketing Management, Internet Marketing, and Tourism Marketing. Dr. Lin had the honor to get the Subsidy for College and University Research Rewarding from Ministry of Science and Technology (MOST).

Jeu-Jiun Hu received the Ph.D. degree from National Cheng Kung University, Tainan, Taiwan, in 2002.

He is an Assistant Professor with the Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan. His research interests include Computational Fluid Dynamics, Parallel computing, and artificial intelligence algorithm.

Implementation of IoT-based Control System for Maintenance Operation of Long-distance Air Pollution Prevention Device RTO

Hoon-Min Park[1], Hyun-Min Jung[1], Dae-Hee Lee[1], Tae-Yeung Lim[1], Myung-Ki Jang[2], Chi-Young Jang[2], Guk-Gyo Youm[3], Keun-Yong Youn[3], Dal-Hwan Yoon[4*], Min-Ki Jung[4], Min-Su Jeon[4], Hwi-Chan Oh[4], Chan-Hyouk Jeon[4], Jong-Geun Kim[4], Hyun-Ah Son[4]

[1] Emsolution Co. Ltd Suwon, Korea, [2] Devicenet Co. Ltd, Korea, [3] Npssystem Co. Ltd, Korea
[4]Department of Electronic Engineering, Semyung University, Korea
hmpark@emsolutions.co.kr, leozzang@devicekorea.net, kgyoum@npssystem.co.kr, yoondh@semyung.ac.kr

Abstract – **In this paper, a control system for maintenance and operation of regenerative thermal oxidation (RTO) and scrubbers, which are air pollution reduction devices, is developed at VOCs-generating sites. VOCs reduction technology should reduce the amount of harmful substances in advance by controlling the operating conditions and maintenance work conditions of systems deployed at long distances. At this time, based on the sensor IoT linkage, the performance is evaluated by measuring the concentration of volatile organic compounds (VOCs) emissions according to the conditions for long-distance maintenance and analyzing the change in THC concentration.**

***Keywords* – RTO, VOCs, Carbon-free, THC, Long-distance Maintenance, Sensor-IoT**

I. INTRODUCTION

Efforts are being made to reduce greenhouse gases due to global warming and to be carbon neutral (zero carbon). The global average temperature rise in response to climate change is maintained within 2°C compared to before industrialization, and 1.5°C is achieved in the long run. Korea has finalized its 37% reduction target (315 million tons) compared to its 2030 emission forecast (851 million tons), is promoting air pollution reduction technologies and research to reduce greenhouse gas emissions and minimize the impact of climate change.

There are VOCs that are discharged directly into the atmosphere or cause greenhouse gases with simple power generation and heat sources. Aromatic hydrocarbons such as benzene are strong carcinogens that cause leukemia, central nervous system disorders, and chromosomal abnormalities, and hydrogen chloride causes ozone layer destruction and global warming. The heat storage combustion system and heat storage catalyst combustion system are technologies that can effectively process VOCs processing rates above 95% [1].

VOCs reduction technology is largely divided into proactive and post-mortem measures, and the amount of harmful substances generated must be reduced in advance by adjusting operating conditions and working conditions. As a follow-up measure, VOCs reduction facilities should be operated, facilities should be improved, and various VOCs reduction facilities should be equipped. The energy costs are consumed with a low heat recovery rate of more than 95%, and the technologies that can effectively handle VOCs treatment rate of more than 95% are heat storage combustion systems and heat storage catalyst combustion systems [2]. Representative companies that treat volatile organic compounds (VOCs) using heat-storage combustion oxidation devices (RTOs) are Alliance Corp. (US), Met Pro (US), Rauschert (Germany), Condor Chemenvitech (Germany), and Taikisha (Japan), and the U.S. and Germany have the best technologies. The RTO type for VOC treatment was developed from the damper method (1st and 2nd generations) to the 3rd generation of rotary valve, and it was developed that there was no pressure change and there were few failure factors with one valve. This rotary valve method was high with a VOC removal efficiency of 95 to 99%, but since the installation area is required more than the existing method. Some leading overseas universities are working with petrochemical companies to automate the process, and process simulation and optimization algorithms such as Carnegie Mellon University, Texas A&M University, Imperial College London University, and Denmark Technical University of Denmark are actively underway [3].

Recently, strategies to reflect carbon reduction in corporate environmental, social, and governance (ESG) management are spreading to a global trend, and as the interest of all members of society, including consumers, investors, and the government, is increasing, it is emerging as a key factor in corporate survival and not a choice. Figure 1 shows the VOCs emission device and the site of the VOCs gas explosion.

Figure 1. The VOCs emission RTO and gas explosion

In this paper, a control system for maintenance and operation of Regenerative Thermal Oxidation (RTO) and scrubbers, which are air pollution reduction devices, is developed at VOCs-generating sites. VOCs reduction technology should reduce the amount of harmful substances in advance by controlling the operating conditions and maintenance work conditions of systems deployed at long distances. At this time, based on the sensor IoT linkage, the performance is evaluated by measuring the concentration of volatile organic compounds (VOCs) emissions according to the conditions for long-distance maintenance and analyzing the change in THC concentration.

II. Implementation of the Remote Control Service

The Regenerative Thermal Oxidizer (RTO) is a high-efficiency energy-saving air pollution reduction device that oxidizes volatile organic compounds (VOCs) and organic odor gases at high temperatures of 800℃ and recovers heat using ceramic heat storage materials for heat exchange. In the event of a failure due to 24-hour operation on weekdays, there are difficulties such as the exact location of the accident at the site, the treatment of VOC harmful substances, and immediate technical support services due to distance. In particular, if a complex equipment usage manual is immature, a lack of handover due to frequent turnover of field personnel, or a problem with factory operation occurs when it is located at a wide and long distance, such as a chemical complex, it cannot be immediately dealt with [4].

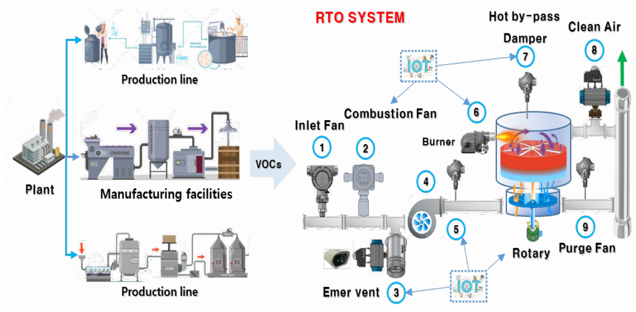


Figure 2. The connection between the production facility and the RTO equipment

Figure 2 shows the connection between the production facility and the RTO equipment, and the circle number in the RTO device is a picture of the sensor IoT attached to the location to be managed.

With the strengthening of environmental regulations on carbon reduction of plants, integrated operation and management of high-efficiency emission facilities became possible. The EU and the US EPA are systemizing using IoT to establish a smart management system for pollutants or indirect managers.

With the development of sensor IoT technology, it is possible to provide information to field managers through wired and wireless edge-IoT technologies to remote RTOs, and this device is connected to the server to analyze detected data. At this time, the data acquired through the sensor IoT is transmitted to the artificial intelligence server, and real-time analysis, judgment, prediction, and visualization are possible. Monitoring service technology, which applies AI-based learning effects for various analysis and judgment, allows managers to expand areas that cannot be managed one by one. Figure 3 shows the on-site interactive service support, big data construction, and PC monitoring system diagram.

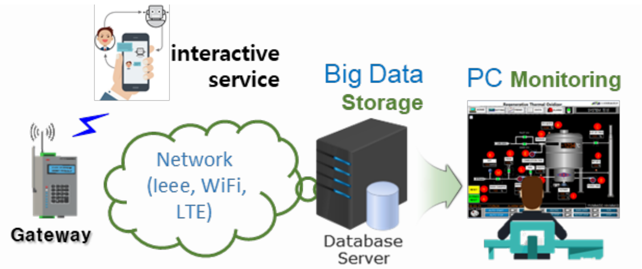


Figure 3. The on-site interactive service support, big data construction, and PC monitoring system.

Gateways can be installed depending on the distance of the site or the amount of IoT installed. At this time, it is data base on the data server through on-site low data and data logging. Figure 4 shows a sensor edge-IoT circuit. Sensors connected to Edge-IoT send data with IEEE 485, LoRa, and WiFi modules depending on wired or wireless.

The main service monitoring server program of each IoT is implemented as a middleware-based system for smooth monitoring between RTO operation status monitoring, control interface, and IoT devices installed in the field. At this time, location information, installation date, equipment replacement date, sensor replacement date, number of malfunctions, and alarm function using Flesh are added. At this time, the system-specific ID allocation, installation classification key, gateway access port are implemented, heartbeat processing per second, data network service (DNS), and the database is managed through the Graphic User Interface (GUI). Figure 4 shows the service monitoring configuration diagram of the main server.

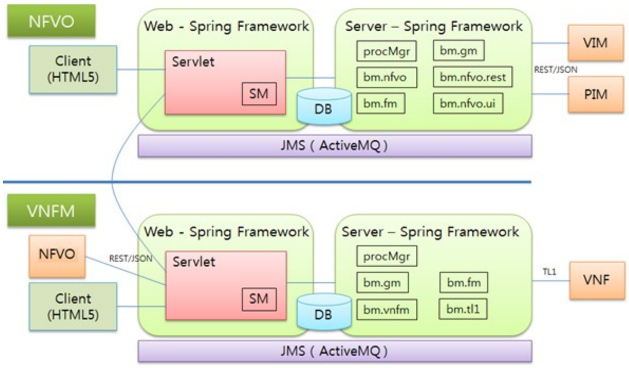


Figure 4. Service monitoring of configuration diagram

III. Building Data on Specifications

Sensor data acquisition specifications are required for detailed configuration specifications of RTO devices. There are duct size, discharge pressure, discharge temperature, and VOC emissions provided to RTO in the plant process. Figure 5 shows a detailed configuration example of an RTO device.

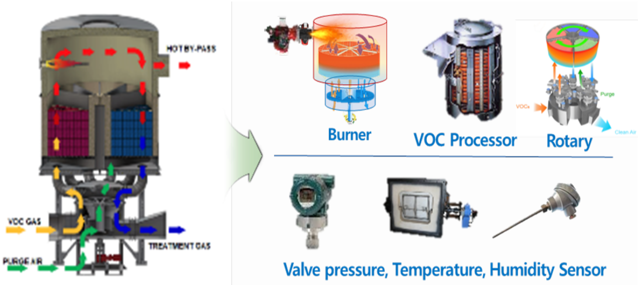


Figure 5. Configuration example of an RTO device

Table 1 shows the segmentation of RTO sensor IoT data acquisition.

TABLE 1. THE DATA SEGMENT OF SENSOR IOT

Level3 Master Components	Level3 Master Components Data	Level3 Detail	Level3 Child (Detailed specifications)	Data(Muliti Line Insert consturuction)	Algorithm (CBM/AI)	Priority (Rating)
Fan (Impeller)	Fan (Impeller)		Fan (Impeller)	Type	Temperature _Airplane	Minor
				Model number	Voltage _Motor	Critical
				Inhalation method	Voltage _Pump	Critical
				Driving method	Unbalanced Rate_Motor	Critical
				Production company	Motor talk	Critical
				Memo	Damper _ loss ratio	Major
Motor		Motor	Power (KW)			
			Blast Current (A)			
			Number of rotations (rpm)			
			Production company			

Data Segmentation

In data segmentation, Level 0 is classified as an RTO facility, Level 1 is classified as a ventilation facility, Level 2 is classified as a ventilation facility-A to segmentation A-Z, and Level 3 is classified as an exhaust-A to segmentation A-Z. Figure 6 shows the combustion chamber temperature and emission gas temperature data obtained from the sensor.

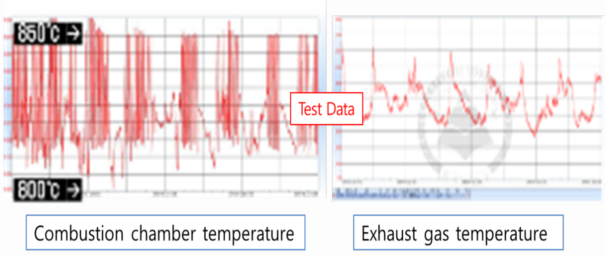


Figure 6. Combustion chamber and emission gas temperature data

The optimization of combustion efficiency from the fuel properties of RTO and the load on the burner will shows in Figure 6. There is a risk of explosion when driving in a rapid incomplete combustion section and monitor, and notify an alarm message.

Table 2 is sensor data of nine channels and shows the built sensor database.

TABLE 2. BUILT SENSOR DATABASE OF NINE CHANNELS AND THE SETTING UP DATA PACKET

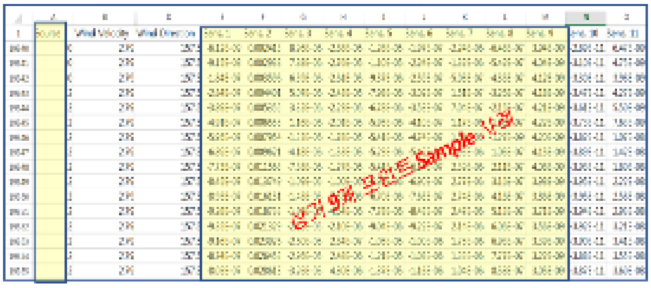


Figure 7 shows that sensor data from nine channels build a sensor database on a server. The right shows the database to be observed.

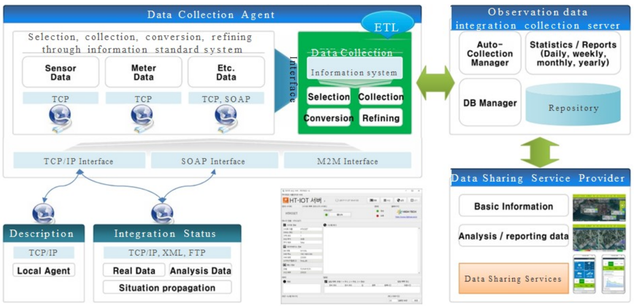


Figure 7. Sensor database on a server

The data obtained from the sensor are data generated by physical sensors such as temperature, flow rate, and pressure of the RTO. By combining them, the concept of virtual sensors is applied to software calculate new indicator values such as the status of RTO emissions.

As real-time data collection of field operations becomes possible, various artificial intelligence machine learning techniques can be applied to preserve the system's prediction. A prediction algorithm may be applied using the acquired data. Table 3 shows the configuration diagram of the monitoring operation program.

TABLE 3. CONFIGUEATION OF THE MONITORING SERVICE

Description		Observation Information management	Sensor measurement management
Status		Measure risk	Data analysis Tools
Analysis	❯	Setting	Information management
Facility Management		Equipment management	Customization
Setting		Customizable	

Figure 9 shows the visualization of temperature, gas and pressure and humidity.

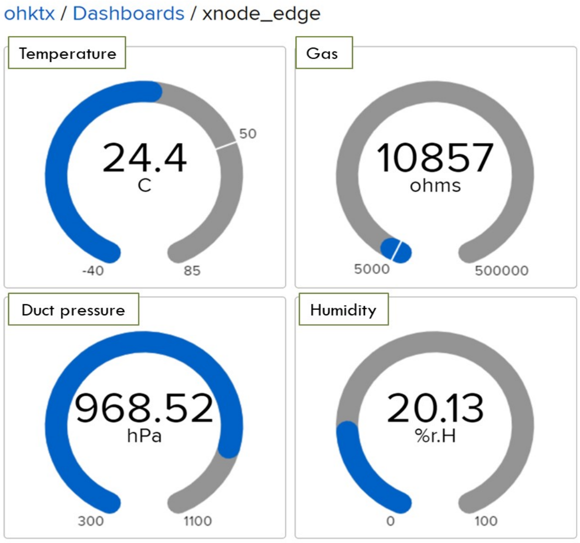


Figure 9. View of temperature, gas and pressure and humidity

After the RTO reaches a stable operating state, the test results obtained by continuously measuring for 30 minutes were as shown in Table 4. Here, the nitrogen oxide is an actual value not corrected by the standard oxygen concentration.

TABLE 4. TEST RESULTS

Sortation	Unit	Inlet	Outlet
Total hydrocarbon	ppm	6,221.27	155.55
nitrogen oxide	ppm	-	2.49
Exhaust gas temperature	℃	-	70.13

Figure 10 is a visualization diagram showing the amount of decrease in VOC.

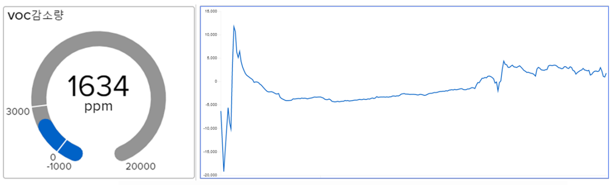


Figure 10. Showing the amount of decrease in VOC

In the RTO simulation, the average concentration of inflow gas according to the operating temperature of the combustion chamber was 6,224 ppm (THC), and the temperature of the RTO combustion chamber was 815°C. The range of change in the temperature of the combustion chamber was the lowest 808°C, and the characteristic of changing periodically up to 825°C be shown.in Figure 10.

IV. Conclusions

As a result of the experiment, the heat generated by VOC combustion was 346,766 kcal/hr, and the exhaust loss heat was 185,758 kcal/hr using the operating inflow gas temperature (49°C) and emission gas temperature (70°C). Therefore, considering only the generated heat and exhaust loss heat from VOC oxidation, it is calculated that 171,008 kcal/hr heat is used as a "regenerated heat energy" inside the RTO system. IoT/ICT technology acquires data from sensors, and machine learning algorithms are used to predict the condition of facilities and equipment before failures occur, and maintenance is planned in advancement to minimize equipment downtime. In the process of predicting, analyzing, and coping with failure causes of equipment in advance, we can expect to grow together with new materials and coating-related technologies that are more resistant to wear and corrosion.

In order to more accurately measure the condition of the facility, the resolution and precision of the sensor are required, and technologies related to the development of higher-performance sensors can be advanced. In addition, by establishing a hardware foundation, a foundation has been created for various algorithms to be developed and applied.

It is expected that the use of failure recognition and prediction of integrated solutions will minimize the stable supply of parts and the shutdown of facilities due to failure, and that VOC will improve public health welfare and respond to the Serious Accident Punishment Act.

Acknowledgment

This study is a knowledge service industry technology development project supported by the Korea Evaluation Institute of Industrial Technology (KEIT), Ministry of Trade, Industry and Energy (No. 20023721). This study was jointly conducted by the Industry-Academic Collaboration Group.

References

[1] H. M. Park, D. H. Yoon, H. M. Jeong, H. G. Min, D. H. Jeon, *"Development of porous plate scrubbers and simulation of IPA treatment efficiency to improve dust collection and deodorization,"* Safety and Culture Research, Vol.13 No.1, p.339-349, 2021

[2] D. K. Park, D. H. Jeon, "*A Numerical Study on Heat Storage System for Recovering Waste Heat from a Volatile Organic Compounds Thermal Storage Type Combustion Oxidizer (RTO)*," Korean Society of Waste Resources Circulation, 2021

[3] D. H. Jeon and S. W. Jung, *"Pre-treatment Method of Irregular Process Emissions for High-Efficiency Combustion Treatment of Volatile Organic Compounds,"* Autumn Conference of the Korea Gas Association, 2021

[4] D. H. Yoon, H. M. Park, "*VOCs Oxidation Treatment and Waste Heat Recovery Technology through the Implementation of Heat Storage Type Combustion Oxidation Device*," Journal of Korean of Safety Culture Forum, Vol.17, p.285-298, 2022

[5] D. H. Yoon, H. M. Park, B. S. Han, etc., "*Implementation of 100CMM Thermal Storage Combustion Oxidizer for VOCs Reduction*," 2022 The 22nd International Conference on Control, Automation and Systems (ICCAS 2022) BEXCO, Busan, Korea, Nov. 27~Dec. 01, 2022

2001: Specialized Bachelor in Construction Eng. Anyang College of Science. 2019: Bachelor, Business Administration, Korea Cyber University. 2021: MBA, IT Business Administration, Ajou University. 2002: Acting Section Chief. D.K Environment Co., Ltd. 2015: . Director, Kumho Environment Co., Ltd. 2016 ~: CEO, Emsolution Co., Ltd. Major Area: Environment System

2014: Bachelor, Department of Environmental System, at Korea University, Korea. 2024: MS Course, Department of Mechanical Engineering, Kyonggi University. 2016: Supervisor, Kumho Eng. Co., Ltd. 2023 Now, General Manager, Emsolution Co., Ltd.
Interests Area: Standby environments and devices, RTOs, Scrubber

2013: Bachelor, School of Civil and Environmental Engineering at Gacheon University. 2022: Deputy General Manager of Environmental Department, Kumho Environment Co., Ltd.. 2022 ~: General Manager of Environmental Department. Emsolution Co., Ltd.
Interests area: Environmental Facility Design

2022: Bachelor, Department of Electronic Engineering, Semyung University. 2023: Environmental System Researcher, Emsolution Co., Ltd.
Interests Area: Sensor IoT, Standby environments and devices, RTOs, Scrubber

2001: Bachelor, Department of Semiconductor Eng., Chungbuk National University, Korea. 2007: MBA Business Strategy, Graduate School of Business, Ajou University, Korea, 2008: Sales manager of LeCory Korea Ltd. 2008~: CEO of LAONURI Co., Ltd. 2022~: CEO of Devicenet. Co., Ltd. 2023~: CEO of IntellyScent Co., Ltd.
Interests area: Semiconductor Design, IoT Network

2006 : B.S. Electrical & Electronic Engineering. KAIST. 2008~2010: Developer, KMAC Co.,ltd. 2011~2014: CEO, M2MKOREA Co.,ltd.. 2014~: CTO, Devicenet Co.Ltd.
Major Area : Measurement & Control System
Interests Area.: USN System, IoT Network System

1996 Bachelor of Science in Electronic Computations, Kangnam University. 1997: General Manager, Ilmaek System Co., Ltd. 2008: CEO of NPS System Co., Ltd
Interests Area.: IoT Network System, Web-Software

1995 Bachelor of Science in Electronic Geoscience, Kangnam University. 1995.11~2001.04: Software Engineer, Dongyang ENP Co., Ltd. 2001.04-2004.: SI Engineer, Wolseong Information System Co., Ltd 2005.01~2009.07 SI Developer, Digital Bay System Co., Ltd. 2009.08~Now, Private Business, SI Development
Interests Area.: Web-server, IoT Network System

1994: Ph.D Graduate School of Electronic Eng., Hanyang University. 1987~1993: Professor of Electronic Eng., KMA. 2001~2003: Director, Institute of Industrial Technology, Semyung University. 2004~ 2009: CEO, Hi-win Co., Ltd. 2010~2015: CTO, Shinwoo Hi-Tech Co., Ltd. 2019~: Vice Chairman, Safety and Culture Forum. 1995.03~: Professor, Dept. of Electronic Eng., Semyung University

Department of Electronic Engineering, Semyung University in 2019
Join the frequency club in 2022.
Currently attending Semyung University
Interests area: Mobile programming & Web-serer,

Department of Electronic Engineering, Semyung University in 2019
Join the frequency club in 2022.
Currently attending Semyung University
Interests area: Mobile programming & Soft Programming. Artificial Intelligence

Department of Electronic Engineering, Semyung University in 2019
Join the frequency club in 2022.
Currently attending Semyung University
Interests area: Server programming, Artificial Intelligence

Department of Electronic Engineering, Semyung University in 2019
Join the frequency club in 2022.
Currently attending Semyung University
Interests area: Web programming & Artificial Intelligence

Department of Electronic Engineering, Semyung University in 2019
Join the frequency club in 2022.
Currently attending Semyung University
Interests area: Softwarre programming & Artificial Intelligence

Department of Electronic Engineering, Semyung University in 2019
Join the frequency club in 2022.
Currently attending Semyung University
Interests area: Web programming & Artificial Intelligence

Search and Recommendation Systems with Metadata Extensions

Woo-Hyeon Kim*, Joo-Chang Kim**

* Division of AI Computer Science and Computer Engineering, Kyonggi University, South Korea

** Contents Convergence Software Research Institute, Kyonggi University, South Korea

whkim712@kyonggi.ac.kr, kjc2232@naver.com

Abstract— **This paper proposes an AI-based video metadata extension model to overcome the limitations of video search and recommendation systems in the multimedia industry. Current video searches and recommendations utilize pre-added metadata. Metadata includes filenames, keywords, tags, genres, etc. This makes it impossible to make direct predictions about the content of a video without pre-added metadata. These platforms also analyze your previous search history, viewing history, etc. to understand your interests in order to serve you personalized videos. This may not reflect the actual content and may raise privacy concerns. In addition, recommendation systems suffer from a cold start problem, which is the lack of an initial target, as well as a bubble effect. Therefore, this study proposes a search and recommendation system by expanding metadata in videos using techniques such as shot boundary detection, speech recognition, and text mining. The proposed method selects the main objects required by the recommendation system based on the object frequency and extracts the corresponding objects from the video frame by frame. In addition, we extract the speech from the video separately, convert the speech to text to extract the script and apply text mining techniques to the extracted script to quantify it. Then, we synchronize the object frequency and the transcript to create a single contextual data. After that, we group videos and clips based on the contextual data and index them. Finally, we utilize Shot Boundary Detection to segment videos based on their content. To ensure that the generated contextual data is appropriate for the video, the proposed model compares the extracted script with the video's subtitle data to check and calibrate its accuracy. The model can then be fine-tuned by tuning and cross-validating the hyperparameter to improve its performance. These models can be incorporated into a variety of content discovery and recommendation platforms. By using expanded metadata to provide results close to a search query and recommend videos with similar content based on the video, it solves problems with traditional search, recommendation, and censorship schemes, allowing users to explore more similar videos and clips.**

Keywords— **Multimedia, Recommendation, Speech Recognition, Contextualized Data, Metadata**

I. Introduction

The multimedia industry is currently experiencing the proliferation of personal internet broadcasting and OTT platforms, creating a huge amount of new videos. Currently, general video search and recommendation uses metadata such as filenames, keywords, tags, and genres pre-added by the creator or distributor as an index [1]. Therefore, access to the video content will take a long time if there is no separate internal index. OTT platforms and video platforms that have a large amount of content may analyze a user's previous browsing history, viewing history, subscribed channels, etc. to understand their interests in order to provide them with customized videos. They compare this with metadata to recommend relevant content or analyze what other users with similar interests have watched, what's popular, etc. Based on this analysis, it selects recommended videos and serves them to you in order of priority. While most search and recommendation algorithms work well, they rely heavily on your sensitive information, such as your viewing history, search history, and internet browser cookies. This can lead to inaccurate recommendations if a user deletes their information or, depending on their level of privacy, lacks the necessary information to make a recommendation, a cold start problem that occurs in most recommendation systems. In addition, since metadata is added by the content creator or distributor, there is a possibility that information about the actual content is excluded, reducing the accuracy of search or recommendation, or being abused. Currently, similar content recommendation systems on OTT platforms often exclude key content that could be spoilers, even if the plot is included in the recommendation process. This can cause users to lose trust in the recommendation system if the actual content is different, even if it is recommended as similar simply because of a similar genre or visual atmosphere. Additionally, recommendation algorithms based on personal history can suffer from the bubble effect. The bubble effect is when only content relevant to your history appears in the recommendation algorithm, exposing users to biased content [2]. Recently, crimes have been committed to exploit this to expose children to harmful videos, raising concerns about recommendation algorithms and video censorship schemes [3]. To solve these problems, it is necessary to strengthen the censorship work and introduce stricter regulations to block and remove harmful content, which requires a lot of time and manpower.

In this study, we use AI-based content analysis to extract contextual data from real-world content. By indexing the extracted contextual data into videos and clips, we propose a model to extend metadata and improve search and recommendation systems.

II. Related Research

OTT is an acronym for Over The Top, which refers to media content such as video content, audio, and text that is typically delivered over the Internet. OTT services work by delivering content directly to consumers, bypassing traditional media distribution channels such as traditional broadcast or cable television. The main characteristics of OTT services include internet-based delivery, diverse content, subscription models, personalized recommendations, content diversity, and creator opportunities. Internet-based delivery means that OTT services deliver content to users via an internet connection, so users can watch or listen to content anytime, anywhere on a variety of devices, including smartphones, tablets, and smart TVs. Next, by diverse content, we mean that OTT services offer many different types of content, including movies, dramas, TV shows, sports broadcasts, audio podcasts, web series, and more. This allows users to choose and watch content that suits their different interests and needs. Subscription is a big feature of OTT services. Many OTT services adopt a subscription model, where users pay a set amount of money each month or year to access the service. Subscribers can use these services to watch the content they want without any ads. Next is personalized recommendations. Many OTT platforms offer personalized content recommendations by analyzing users' viewing habits. This makes it easier for users to find content that matches their tastes and interests [4]. Finally, there is content diversity and creator opportunities. Unlike traditional broadcast networks, OTT services have relatively low barriers to entry. This gives many creators the opportunity to produce and distribute their own content. Major OTT services include Netflix, Amazon Prime Video, Disney+, Hulu, Apple TV+, and Youtube Premium, which have millions of users worldwide. OTT is disrupting the media industry, and research is still ongoing.

Shot Boundary Detection is a technique for detecting scene transitions in video content. A video consists of a sequence of consecutive frames, and it detects changes in color, lighting, objects, etc. in a scene. Based on this, Shout Boundary Detection is the process of automatically segmenting video into basic units called shots. A shot is a group of consecutive video frames in a video. A shot usually consists of consecutive video frames from a single camera. There are several types of Shot Boundaries: Cut, Dissolve, Wipe, and Fade Out/In. Cut is the most common type of Shot Boundary and represents a transition from one frame to the next. Dissolve is a transition type where one scene gradually disappears and the next appears. Wipe is a transition from one scene to another, such as a horizontal or vertical wipe pattern across the screen. Fade Out/In is a type of transition where one scene gradually fades out or brightens and then disappears. Over the years, the design of the Shot Boundary Detection algorithm has evolved from simple feature comparisons to the use of rigorous probabilistic and complex models. In addition, to detect transitions with higher accuracy, orthogonal polynomials are applied to derive features in the orthogonal transform domain to detect hard transitions in video sequences [5].

Speech Recognition is a technology that recognizes human speech and converts it into text by a computer. This allows users to interact with computers through voice commands or voice input. This speech recognition technology is also known as Speech to Text (STT). STT involves voice input, digital signal processing (DSP), feature extraction, acoustic modeling, and language modeling. The user enters information by voice through a microphone, and then the voice signal is converted from analog to digital. This is followed by digital signal processing, which involves preprocessing to remove background noise, cancel echoes, etc. Then, important features are extracted from the digitized speech. This is usually done using an algorithm such as MFCC (Mel Frequency Cepstral Coefficients) [6]. The extracted features are then applied to a pre-trained model such as an HMM or DNN for acoustic modeling to recognize each word or pronunciation. Finally, linguistic modeling is performed by selecting the words that are most naturally connected into a sentence from several word candidates. STT technology is used in personal assistant services on smartphones such as Siri, Google Assistant, and Bixby, automatic translation services, and assistive devices. Other applications include customer service centers and education. STT technology is also evolving with advances in AI and machine learning. Deep learning methodologies such as LSTM (Long Short-Term Memory) and Transformer have contributed to the improvement of STT's performance. However, it is still difficult to achieve 100% accuracy due to unclear pronunciation, various accents and dialects, and noise.

Text Mining refers to the process of extracting useful information from large amounts of text data by combining Natural Language Processing and Data Mining techniques. It is used to discover and analyze patterns, trends, information, and statistical characteristics from text data [7]. The main purposes for which it is used are information retrieval, sentiment analysis, topic modeling, document classification, and information extraction. Information retrieval is concerned with finding and ranking relevant documents for a user's search query. It is one of the core technologies used by web search engines. Next, sentiment analysis is used to analyze the sentiment or opinion of a particular text to determine if it is positive, negative, or neutral. Subject modeling is a technique for identifying and grouping key subjects in large amounts of text data. This allows you to discover hidden patterns in text data. Document classification is the categorization of text into predefined categories or categories, which can be used for spam filtering, news article categorization, legal document categorization, and more. Information extraction is the process of extracting important information from text. Text Mining does this by applying machine learning and statistical analysis techniques to text data.

Automatic metadata expansion and recommendation systems have been studied before. For example, T. Tsunoda et

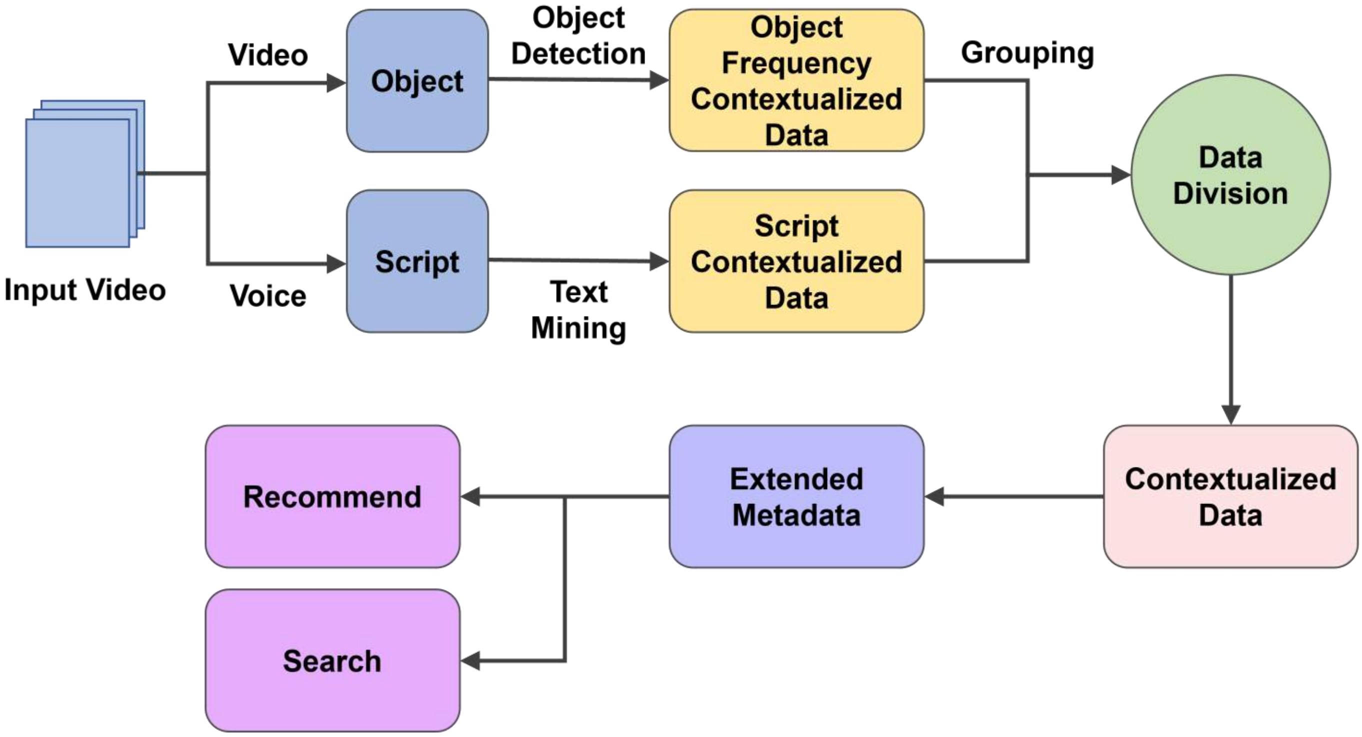


Figure 1. Search and recommendation Process based on Metadata Extensions

al. proposed "Automatic metadata expansion and indirect collaborative filtering for TV program recommendation system" [8]. They proposed two methods to improve the accuracy of TV program recommendations. The first proposed automatic metadata expansion (AME) and enrichment of TV program metadata from electronic program guide (EPG). The second is an indirect collaborative filter (ICF) to recommend non-persistent items such as TV shows based on the preferences of other members of the community. The proposed methods can generate rich data about target items. In addition, the recommendation accuracy can be improved not only by the user's preferences but also by the preferences of other users in the community. However, there are some challenges. First, AMEs can generate common profiles, such as lifestyle. This creates a dependency on users for sensitive information, which needs to be addressed. The second is that there is a lot of computation on attributes as parameters are set for each user to improve accuracy. Therefore, a way to reduce this computation is needed.

III. SEARCH AND RECOMMENDATION SYSTEMS WITH METADATA EXTENSIONS

In this study, we use a YOLO deep learning model to extract objects from videos frame by frame to develop a metadata expansion-based search and recommendation system. At the same time, we extract speech from the video separately and convert it into a script using STT technology. The extracted script data is then segmented into scenes. Then, we synchronize the object and text data to create a single contextual data. We check whether the context data is appropriate for the video and fine-tune it by making adjustments. Index the generated contextual data by grouping it with the video. Expand the metadata based on the indexed information. Include the expanded metadata in search or recommendation to improve the performance of the search and recommendation system. Figure 1. illustrates the process of generating metadata from videos.

A. Data Preprocessing

Select key objects required in a video recommendation system based on object frequency. We use the YOLO deep learning model to train a deep learning-based object recognition model on the labeled image dataset to detect and extract objects based on frames. Since the YOLO deep learning model is trained using the COCO Dataset, it can detect and classify a total of 91 objects, from 1 to 91. In addition, we extract the voice from the video separately and then use STT technology to extract the script according to the frame. To apply machine learning techniques to the extracted script in natural language form, we apply text mining techniques to quantify it.

Text mining libraries that can be used include Python's Natural Language Toolkit (NLTK), spaCy, gensim, scikit-learn, TensorFlow, PyTorch, which utilize deep learning techniques. The process of quantifying through text mining involves many different steps. For our proposed method, the text is tokenized by breaking it into small chunks and going through lemmatization to remove unimportant words, root extraction to find the basic form of words, and part-of-speech tagging to understand sentence structure and meaning.

B. Extending Metadata with Contextual Data

Object frequencies and scripts extracted from videos have the characteristics of time series data. Object frequencies are expressed as the number and type of objects observed in each frame. Scripts appear across multiple frames and are generated as contextual data from the frame where the dialog starts. The object frequencies and scripts are then synchronized. When dividing the context data into scenes, we do so based on scripts. We use iterative experiments to divide the scripts into appropriately sized sentences or paragraphs, as too small a division would clutter the data, too large a division would distort the features.

C. Grouping Video – Clips and Indexing

Since video content has a time-series nature over time, changing with events or scene transitions, Shot Boundary Detection segments videos based on content, and then analyzes and processes videos as content units.

D. Fine-tuning

In this paper, we propose a video-clip indexing model using contextual data based on intelligent content analysis. This model enables fine-tuning using contextual data extracted from videos. First, we use scripts extracted using STT technology. By comparing the extracted script with the subtitle data embedded in the video, the accuracy of the extracted script is checked and calibrated.

Adjust the hyper-parameters used to train the model and improve the model performance through cross-validation. For cross-validation, compare the extracted script with the object frequency context data obtained using the YOLO deep learning model. We then evaluate and test the developed model. This can be evaluated by putting some videos into the test and checking if the correct metadata is extracted. Figure 2. shows the fine-tuning process of the proposed model.

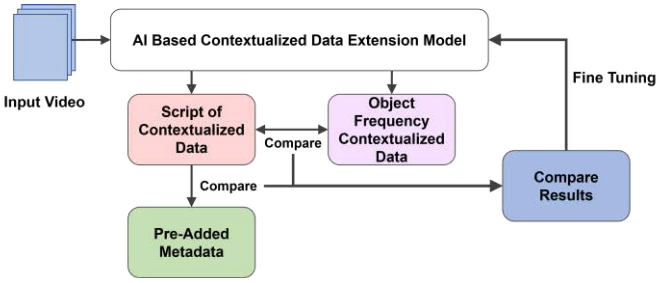

Input Video
AI Based Contextualized Data Extension Model
Fine Tuning
Script of Contextualized Data
Compare
Object Frequency Contextualized Data
Compare
Compare Results
Pre-Added Metadata

Figure 2. Process of Model Fine-Tuning

Use it to develop a proof-of-concept video-clip search program and integrate it into your environment. Continuously evolve the model through user feedback and performance monitoring, and then compare it to traditional search and recommendation methods to improve performance.

IV. CONCLUSIONS

In this paper, we proposed a model to enhance search and recommendation systems by extending metadata with contextual data extracted from actual content content through AI-based content analysis.

If this model is applied to various content search and recommendation platforms, the extended metadata can be used to provide results that are close to the search query when the user enters the search query. It can also be used to help users discover videos that contain similar content based on the video. This will provide a way to discover more similar videos and clips by allowing metadata to reflect the actual content of content that was previously excluded. This will compensate for problems with existing search, recommendation, and censorship systems, enabling more in-depth recommendations. It can also be extended to various video platforms, such as managing videos with long running times, such as CCTV, by splitting them. Therefore, it can play a role in saving time and manpower in parts of the social safety surveillance system such as CCTV control centers. However, there is a limitation in that human conversations can be transcribed by applying STT technology, but background sounds such as car sounds, or other sounds such as animal cries cannot be transcribed. Therefore, in future research, we will solve the problem of analyzing other sounds based on the proposed model.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No: RS-2023-00248899).

REFERENCES.

[1] J. C. Kim and K. Y. Chung, “Knowledge expansion of metadata using script mining analysis in multimedia recommendation,” *Multimedia Tools and Applications*, vol. 80, pp. 34679-34695, Mar. 2020.

[2] M. Ekstrand and J. Riedl, “When recommenders fail: predicting recommender failure for algorithm selection and combination,” in *Proc. RecSys'12*, 2012, p. 233-236.

[3] A. Ishikawa, E. Bollis and S. Avila, "Combating the elsagate phenomenon: Deep learning architectures for disturbing cartoons," in *Proc.IWBF'19*, 2019, p. 1-6.

[4] A. Yousaf., A. Mishra., B. Taheri and M. Kesgin, “A cross-country analysis of the determinants of customer recommendation intentions for over-the-top (OTT) platforms,” *Information & Management*, vol. 58, no. 8, pp. 103543, Oct. 2021.

[5] S. H. Abdulhussain, A. R. Ramli, B. M. Mahmmod, M. I. Saripan, S. A. R. Al-Haddad, and W. A. Jassim, “Shot boundary detection based on orthogonal polynomial,” *Multimedia Tools and Applications*, 78, pp. 20361-20382, Feb. 2019.

[6] C. Ittichaichareon., S. Suksri. and T. Yingthawornsuk, “Speech recognition using MFCC,” In *Proc. ICGSM'12*, 2012, p. 135-138.

[7] A. H. Tan, “Text mining: The state of the art and the challenges,” In Proc. KDAD'99, 1999, p. 65-70.

[8] T. Tsunoda. and M. Hoshino, “Automatic metadata expansion and indirect collaborative filtering for TV program recommendation system. Multimedia Tools and applications,” Multimedia Tools and applications, vol. 36, pp. 37-54, Jan. 2008.

Woo-Hyeon Kim (M’23) was born in Geoje-si, Gyeongsangnam-do, South Korea, July 12, 2003.
In 2022, he enrolled in the Division of AI Computer Science and Computer Engineering at Kyonggi University, majoring in Computer science.
His educational background includes:
- Bachelor’s Degree:Computer Science, Kyonggi University, South Korea, 2022

His Research interests encompass Data Mining, Big Data, and Anomaly Detection

Joo-Chang Kim has received B.S. and M.S. degrees from the School of Computer Information Engineering, Sangji University, South Korea in 2014 and 2016, respectively. He has received Ph.D. from Department of Computer Science, Kyonggi University, South Korea in 2021. He is currently a research professor in Contents Convergence Software Research Institute, Kyonggi University. Since 2021, he is currently a lecturer in the Department of Software Convergence Engineering, Inha University, South Korea. His research interests include data mining, data management, knowledge systems, machine learning, deep learning, big data, healthcare, and recommendation systems.
.

Utterance-Level Incongruity Learning Network for Multimodal Sarcasm Detection

Liujing Song[*†], Zefang Zhao[*†], Yuxiang Ma[§], Yuyang Liu[¶] and Jun Li[*‡]
[*]Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
[†]University of Chinese Academy of Sciences, Beijing, China
[§]School of Computer and Information Engineering, Henan University, Kaifeng, China
[¶]Institute of Medical Information, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
{songliujing, zhaozefang, lijun}@cnic.cn, y.x.ma@hotmail.com, liu.yuyang@imicams.ac.cn, [‡]Corresponding Author

Abstract—**With the exponential growth of user-generated online videos, multimodal sarcasm detection has recently attracted widespread attention. Despite making significant progress, there are still two main challenges: 1) previous works primarily relied on word-level feature interactions to establish relationships between inter-modality and intra-modality, which could potentially lead to the loss of fundamental emotional information. 2) they obtained the incongruity information only interacted with textual modality, which may lead to the neglect of incongruities. To address these challenges, we propose a novel utterance-level incongruity learning network (ULIL) for multimodal sarcasm detection, where the multimodal utterance-level attention (M-ULA) and incongruity learning network (ILN) are the two core modules. First, we present M-ULA to interact with utterance-level multimodal information, complementing word-level features. Furthermore, ILN selects primary modality and auxiliary modality automatically, and leverages cross-attention and self-attention to learning incongruity representations. We conduct extensive experiments on public datasets, and the results indicate that our proposed model achieves state-of-the-art performance in multimodal sarcasm detection.**

Keywords—**multimodal sarcasm detection, utterance-level attention, incongruity learning**

I. Introduction

Sarcasm is a sophisticated form of linguistic expression where the intended meaning often contrasts with the literal interpretation [1]. This implicit emotional expression poses significant challenges for natural language processing tasks such as sentiment analysis and opinion mining. Previous sarcasm detection methods have primarily focused on text modality for classification task [2], [3]. However, the proliferation of user-generated online videos presents both opportunities and challenges for multimodal sarcasm detection.

Recently, some works have conducted multimodal sarcasm detection by capturing incongruity between modalities, including textual, visual and audio. Castro et al. [4] first extracted each modality feature and then fused them through early fusion. Chauhan et al. [5] combined the inter-segment intermodal attention representation and intra-segment inter-modal attention representation for multitasking. Wu et al. [6] constructed a scare mechanism to acquire incongruity information based on word-level. Despite these advances, multimodal sentiment feature extraction remains a challenging problem. First, relying solely on word-level features is insufficient to capture the complex interplay between modalities. Second, incongruity information may occur on arbitrarily inter-modality and intra-modality. Thereby, just exploring incongruity information from textual with other modalities may cause a loss of comprehensive inconsistent information.

To address the issues mentioned above, we present a Utterance-Level Incongruity Learning Network (ULIL) for multimodal sarcasm detection, which comprises two core modules, namely multimodal utterance-level attention (M-ULA) and incongruity learning network (ILN). M-ULA is a novel multimodal information interaction and feature fusion module that can thoroughly explore complex inter-modal and intra-modal relationships. ILN is a novel incongruity representation network that performs comprehensive multimodal sarcasm detection. Specifically, ILN first dynamically determines the primary modality based on its contribution to the detection task and then hierarchically fuses the auxiliary modalities via cross-attention and self-attention mechanisms to efficiently obtain the most useful incongruity information.

In summary, the main contributions of our work are summarized as follows:

- We propose a novel utterance-level incongruity learning network for multimodal sarcasm detection. The core module is multimodal utterance-level attention (M-ULA) which relies on utterance-level to extract unimodal and multimodal features and effectively explore the complex interrelationships present in multimodal data.
- To capture the incongruity information, we design an incongruity learning network (ILN) which consists of primary modality choosing, cross attention mechanism and self-attention mechanism to explore sufficient incongruity information.
- We conduct extensive experiments and comparisons on public dataset to demonstrate the superiority of the proposed method. Furthermore, we perform ablation experiments to demonstrate the rationality and necessity of the proposed approach.

Organization The rest of the paper is organized as follows. A general exploration of textual modality and multimodal sarcasm detection methods are given in Section II. Section III describes the ULIL framework. The fundamental experimental setup is depicted in IV. Following, the extensive results and analysis are show in V. Ablation studies are discussed in Section VI. Finally, Section VII concludes the proposed model.

II. Related Work

A. Textual Modality Sarcasm Detection

Sarcasm detection in the textual modality has long been a research focal point in natural language processing. Early research relied on rule-based methods [7], [8] and feature engineering-based machine learning methods [9], [10] to represent incongruous emotional expressions. However, with the evolution of neural networks, deep learning methods like CNN [11], RNN [12], and GNN [3] have found widespread application in textual sarcasm detection. In recent years, Tay et al. [2] explored an attention-based neural model which attempts to acquire contrast and incongruity information. Xiong et al. [13] proposed a self-matching network to capture sentence incongruity by discovering word-to-word interactions. Lou and Liang et al. [3] first leveraged Graph Convolutional Network to draw a long-range incongruity pattern by modeling the affective and dependency features information. Nonetheless, manual graph construction can potentially lead to some errors. To alleviate this limitation, Wang et al. [14] adopted an iterative augmenting affective and dependency graph architecture to learn the incongruity graph structure interactively.

B. Multimodal Sarcasm Detection

With the rapid growth of online media, data modalities have become increasingly diverse, leading researchers to incorporate vision and audio modalities in addition to text modalities for multimodal sarcasm recognition. As for multimodal sarcasm detection involving textual and visual modalities, Cai et al. [1] created a public dataset and proposed a fusion model by fusing text features, image features and image attributes to address the task. Liang et al. [15] constructed a crossmodal graph for each text and region to draw the sentiment relations between textual and visual modalities directly. Qiao et al. [16] explored a local and global incongruity learning network to seek underlying consistency between textual and visual modalities. As for multimodal sarcasm detection involving textual, visual and audio tri-modalities, Castro et al. [4] argued that incorporating multimodal cues can enhance automatic sarcasm detection. Thereby they constructed a trimodal public dataset. Chauhan et al. [5] concatenated the inter-segment inter-modal attention representation and intra-segment inter-modal attention representation for multi-tasking. Wu et al. [6] adopted incongruity aware attention network which focuses on word-level to detect sarcasm by score mechanism.

III. Proposed Method

A. Problem Statement

Multimodal sarcasm detection uses valid information from different modalities to judge potential inconsistencies. In this research, three modalities sequences including text $I_t \in \mathbb{R}^{L_t}$, acoustic $I_a \in \mathbb{R}^{L_a}$ and visual $I_v \in \mathbb{R}^{L_v}$ are used as source data. Furthermore, we consider multimodal sarcasm detection as a binary classification task, where three sequences are taken as input, and sarcasm is the final prediction.

B. Overall Architecture

As illustrated in Fig. 1, our proposed utterance-level incongruity learning network for multimodal sarcasm detection method consists of three layers. For the utterance-level feature extraction layer, to capture richer effective information from both intra-modality and inter-modality, we first use the unimodal encoder and utterance-level attention to obtain text, audio and vision features respectively. Then we leverage multimodal utterance-level attention (M-ULA) to acquire interacted fusion features. For the incongruity learning network layer, we first need to determine the primary modality, then employ a cross-attention mechanism to capture the inconsistency information, and finally employ self-attention mechanism to obtain the enhanced inconsistency information. For the prediction layer, the multimodal sarcasm label is predicted through a softmax classifier.

C. Utterance-Level Feature Extractions

a) Unimodal Feature Extractions: This module includes three main functions: token-level unimodal feature extraction, alignment and utterance-level feature representation. For the textual modality, we employ a pre-trained BERT-base model to encode the input textual tokens into word embedding. For audio and vision modalities, referring to previous work, we take advantage of Long Short-Term Memory(LSTM) to extract the initial features, which are denoted as:

$$\begin{aligned} I'_t &= \mathrm{BERT}(I_t) \\ I'_a &= \mathrm{LSTM}(I_a) \\ I'_v &= \mathrm{LSTM}(I_v) \end{aligned} \tag{1}$$

where $I'_f \in \mathbb{R}^{L_f \times d_f}$, $f \in \{t, a, v\}$, L_f is the feature length and d_f represents the feature dimension.

Moreover, we deploy a 1-dimensional convolution layer and a scale layer to align the dimension of feature vectors space, which is denoted as:

$$X_f = \mathrm{Scale}(\mathrm{Conv}(I'_f)) \tag{2}$$

where $\mathrm{Scale}(\theta) = \frac{\theta}{\sqrt{\|\theta\|_2}}$, $X_f \in \mathbb{R}^{L_f \times d_f}$. It is worth noting that we standardize the dimensions of feature spaces from different modalities through convolution layer and scale, that is $d_t = d_a = d_v$.

Token-level feature interaction may lead to missing vital sentiment information. Therefore, we employ utterance-level attention(ULA) to learn the distinctive properties of unimodal modality. Nonetheless, the unimodal ULA module can only

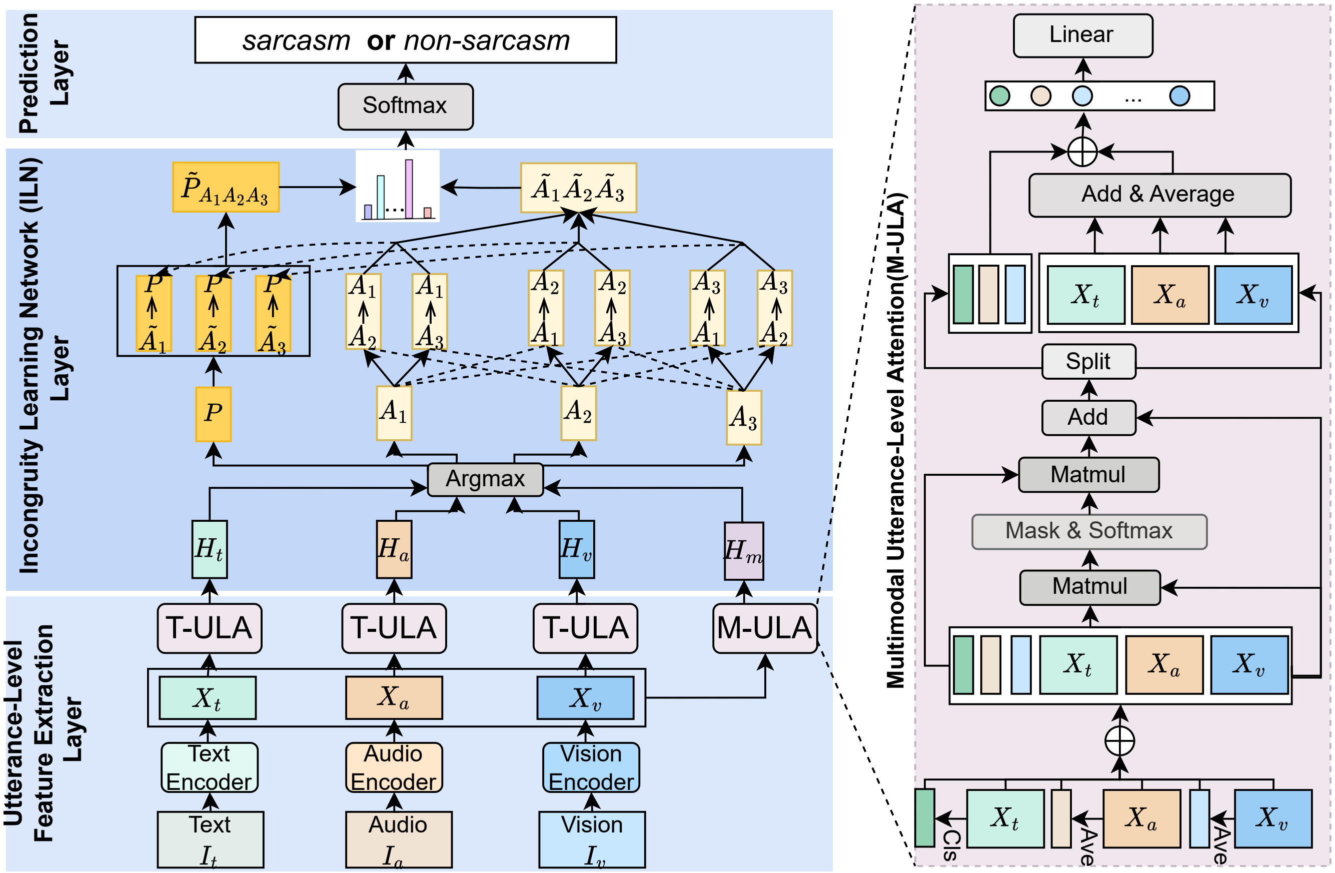


Fig. 1. The architecture of the proposed utterance-level incongruity learning network (ULIL) for multimodal sarcasm detection. At the bottom is the utterance-level feature extraction layer, which can extract unimodal features and interact with multimodal features by the Multimodal Utterance-Level Attention(M-ULA) module on the right. In the middle is the incongruity learning network layer, which can obtain incongruity information via cross-attention and self-attention.

explore internal relationships within the single modality. We extract unimodal features as input and obtain the utterance-level features representation is:

$$H_f = \mathrm{ULA}(X_f) \tag{3}$$

where $H_f \in \mathbb{R}^d$. The unimodal ULA is a simplified version of the M-ULA. For detailed information, refer to the next section. Hence, we first acquire utterance-level features U_f through token-level features X_f. Subsequently, we can obtain the updated features H_f from each unimodal ULA.

$$H_f = \mathrm{Linear}(|U'_f, \mathrm{Ave}(X'_f)|) \tag{4}$$

b) Multimodal Feature Extraction: The main task of this module is to learn complex relationships of inter-modal, interact with multimodal information and gain the fusion feature representations with multimodal ULA. We feed X_t, X_a and X_v into the M-ULA and learn the fusion multimodal feature representation:

$$H_m = \mathrm{ULA}(X_t, X_a, X_v) \tag{5}$$

The architecture of M-ULA is illustrated on the right of Fig.1. We design multimodal utterance-level attention(M-ULA) to explore complex intra-modal and inter-modal interactions. We aim to learn a more enriched multimodal fused feature representation with abundant emotional information.

We take three token-level unimodal features (X_t, X_a, X_v) as input and gain their utterance-level features represent (U_t, U_a, U_v). We employ attention computation to establish fully connected correlations among these six emotion features and finally obtain the updated features represented by fusing. For the textual modality, we utilize the first vector of X_t as U_t, corresponding to the [CLS] token in BERT. For the audio and vision modalities, we can derive their initial utterance-level features by applying an average function. Next, we concatenate these six features to obtain the fused features:

$$X_m = |U_t, U_a, U_v, X_t, X_a, X_v| \tag{6}$$

where $X_m \in \mathbb{R}^{(3+L_t+L_a+L_v)\times d}$, and $|\cdot|$ denotes concatenation function.

During the training process, the utterance-level features of the three modalities interact with different feature information based on the attention scores ($att_{i,j}$), enabling their dynamic updates. We first learn the $query(Q) = W_q X_m$, $key(K) = W_k X_m$ and $value(V) = W_v X_m$, where $Q/K/V \in \mathbb{R}^{(3+L_t+L_a+L_v)\times d}$ and W_f are the weights,

$f \in \{t, a, v\}$. Therefore, the formula for the attention matrix of multimodal features is:

$$attention_{i,j} = Q \times K^{\top} \tag{7}$$

where $attention_{i,j} \in \mathbb{R}$ and $i, j \in [1, 3 + L_t + L_a + L_v]$.

Considering the impact of padding data on feature representation, we have devised a mask matrix M. It sets the positions of actual words and utterances to zero, while the positions of padding data are set to $-\infty$. The resulting attention score matrix is derived as follows:

$$att\_score = \mathrm{Softmax}(attention_{i,j} + M) \tag{8}$$

The updated multimodal feature representations $X'_m \in \mathbb{R}^{(3+L_t+L_a+L_v)\times d}$ are obtained using the residual function:

$$X'_m = att\_score * V + X_m \tag{9}$$

Then, we split the interactive features into two sets, utterance-level features (U'_t, U'_a, U'_v) and token-level features (X'_t, X'_a, X'_v). We independently weight and sum the three learnable token-level features with weights W_t, W_a W_v, obtaining the fused token-level feature values F_token. Next, by taking the average, we obtain the global token feature D_token, which is further fused with the utterance-level features to generate fusion features D_m. Eventually, after passing through several linear layers, we obtain the multimodal features representation H_m:

$$\begin{aligned} H_m &= \mathrm{Linear}(D_m) \\ &= \mathrm{Linear}(|U'_t, U'_a, U'_v, \mathrm{Ave}(F\_token)|) \\ &= \mathrm{Linear}(|U'_t, U'_a, U'_v, D\_token|) \end{aligned} \tag{10}$$

where $H_m \in \mathbb{R}^d$, $F\_token = W_t X'_t + W_a X'_a + W_v X'_v$.

D. Incongruity Learning Network

a) Primary Modality Selecting: To assess the modality inconsistency, it is essential to first identify the primary modality. Referring to [17], we automatically assign weight values based on the varying contributions of each modality for sarcasm detection, with a higher contribution leading to a higher weight value. We update the weight values in each training batch while maintaining the sum of all trainable weight values equal to 1. We obtain the primary modality P and auxiliary modalities A_1, A_2, A_3.

b) Incongruity Learning by cross-attention: In order to obtain incongruity from primary modality and auxiliary modalities, we leverage the cross-attention(CMA) mechanism to acquire interacted auxiliary modalities:

$$\begin{aligned} \tilde{A_1} &= \mathrm{CMA}(\tilde{A_2} \to \tilde{A_1}) \oplus \mathrm{CMA}(\tilde{A_3} \to \tilde{A_1}) \\ \tilde{A_2} &= \mathrm{CMA}(\tilde{A_1} \to \tilde{A_2}) \oplus \mathrm{CMA}(\tilde{A_3} \to \tilde{A_2}) \\ \tilde{A_3} &= \mathrm{CMA}(\tilde{A_1} \to \tilde{A_3}) \oplus \mathrm{CMA}(\tilde{A_2} \to \tilde{A_3}) \end{aligned} \tag{11}$$

Subsequently, we can obtain the interacted primary modality representations:

$$\begin{aligned} \tilde{P}_{\tilde{A_1}} &= \mathrm{CMA}(\tilde{A_1} \to P) \\ \tilde{P}_{\tilde{A_2}} &= \mathrm{CMA}(\tilde{A_2} \to P) \\ \tilde{P}_{\tilde{A_3}} &= \mathrm{CMA}(\tilde{A_3} \to P) \end{aligned} \tag{12}$$

c) Incongruity Learning by self-attention: We concatenate interacted primary modality $\tilde{P}_{\tilde{A_1}}$, $\tilde{P}_{\tilde{A_2}}$ and $\tilde{P}_{\tilde{A_3}}$, and then we employ a self-attention(SA) mechanism to identify its salient components as the ultimate primary modality representation:

$$\tilde{P}_{\tilde{A_1}\tilde{A_2}\tilde{A_3}} = \mathrm{SA}(\tilde{P}_{\tilde{A_1}} \oplus \tilde{P}_{\tilde{A_2}} \oplus \tilde{P}_{\tilde{A_3}}) \tag{13}$$

In the end, we obtain the multimodal inconsistency representation by concatenating auxiliary modalities $\tilde{A}_1$, $\tilde{A}_2$, $\tilde{A}_3$ and primary modality $\tilde{P}_{\tilde{A_1}\tilde{A_2}\tilde{A_3}}$:

$$Z = W_1\tilde{A}_1 \oplus W_2\tilde{A}_2 \oplus W_3\tilde{A}_3 \oplus \tilde{P}_{\tilde{A_1}\tilde{A_2}\tilde{A_3}} \tag{14}$$

where W_1, W_2 and W_3 are the weight trainable parameters, which are learned by the modal itself to adjust the amount of auxiliary information to be extracted.

E. Multimodal Sarcasm Classification

The task of multimodal sarcasm detection is to predict the label $y \in \{\mathrm{sarcasm}, \mathrm{non\text{-}sarcasm}\}$. Therefore, the ultimate utterance-level incongruity representation is passed through a fully connected layer with a softmax activation function to produce a probability distribution y in the multimodal sarcasm decision space:

$$y = \mathrm{softmax}(W_o Z + b_o) \tag{15}$$

where W_o and b_o are trainable parameters.

IV. Experiment

A. ***Dataset***

TABLE I
DATASETS STATISTICS

	Fr	*BBT*	*TGG*	*Sa*	*total*
Sarcasm	152	140	39	14	345
Non-sarcasm	204	140	1	0	345
total	356	280	40	14	690

We evaluate our proposed model on the public multimodal sarcasm detection benchmark dataset[1] **MUStARD** published by [4]. The dataset comprises 690 utterances (345 sarcastic examples and 345 non-sarcastic examples) sourced from famous TV shows such as Friends (Fr), Big Bang Theory (BBT),The Golden Girls (TGG), and Sarcasmaholics (Sa). The detailed statistics of the experimental dataset are presented in Table I. Following [4], [6], We also evaluated our model in two experimental settings. One is the speaker-independent setup, where utterances collected from the Friends are used as testing data, and the remaining utterances are used for training data. The other is a speaker-dependent setup, where the dataset is split into five parts. In each of the five iterations, the i-th part is used as the testing, while the rest is used for training. As a result, we obtained five datasets.

[1] https://github.com/soujanyaporia/MUStARD

TABLE II
COMPARISON RESULTS OF DIFFERENT METHODS FOR UNIMODAL AND MULTIMODAL SARCASM DETECTION

Modality	**Method**	**Speaker-Dependent**			**Speaker-Independent**		
		Precision(%)	**Recall(%)**	**F1-score(%)**	**Precision(%)**	**Recall(%)**	**F1-score(%)**
T	SMSD	61.6	61.0	61.1	51.7	48.2	47.0
	MIARN	64.7	64.0	63.9	60.4	55.2	54.0
	BERT	67.5	66.9	66.8	58.2	56.7	57.0
T,A,V	RAVEN	69.1	67.5	67.1	53.8	50.4	49.7
	LSTM(A)	67.3	66.7	66.3	56.7	54.1	54.0
	MFN	70.1	69.6	69.7	66.0	62.5	62.4
	EF-Concat	71.2	70.8	70.8	64.3	63.1	63.3
	IAIE	72.1	71.6	72.0	66.0	65.5	65.6
	IWAN	75.2	74.6	74.5	71.9	71.3	70.0
T,A,V	**ULIL**	**76.9**	**76.1**	**76.5**	**73.1**	**72.6**	**72.8**

B. Baselines

To comprehensively evaluate the performance of our proposed model **(ULIL)**, we compared it with the baselines in the following modality including Text modality models and Trimodal models.

Text modality models (T):

- **SMSD** [13]: investigating a self-matching network and a low-rank bilinear pooling method to extract incongruity.
- **MIARN** [2]: adopting a multi-dimensional intra-attention recurrent network to capture textual incongruity.
- **BERT** [18]: is a pre-trained model and achieves state-of-art methods in many natural language processing tasks.

Trimodal models (T, A, V):

- **RAVEN** [19]: uses nonverbal sub-networks and gated modality-mixing network to shift word representations.
- **LSTM(A)** [20]: captures word-level multimodal features by the gated multimodal embedding LSTM with temporal attention.
- **MFN** [21]: considers and dynamically models interactions between modalities within a neural architecture over time.
- **EF-Concat** [4]: is an utterance-level multimodal sarcasm detection method by SVM classifier.
- **IAIE** [5]: exploring inter-segment inter-modal attention and intra-segment inter-modal attention to fuse the multimodal features.
- **IWAN** [6]: is an incongruity-aware attention network which focuses on word-level incongruity between three modalities by a scoring mechanism.

C. Settings

Evaluation Metrics: Following previous work [6], we report the weighted *precision*, *recall* and *F1-scores* to evaluate the performance of baselines and our model.

Implementation Details: Our proposed method is implemented by PyTorch [22] and all of the experiments in the paper are employed on the NVDIA Tesla V100 GPU. The hyperparameters are the same for the two experimental configurations. The proposed ULIL uses the Adam optimizer. The learning rate is set to 0.001. The batch size is set to 32. To mitigate overfitting, we set the dropout rate to 0.1. We set the dimension of the feature space after convolution to 300.

V. RESULTS AND ANALYSIS

The comparative evaluation experiment results of our proposed model and all baselines are shown in Table II. We can draw the following conclusions: 1) Our proposed ULIL outperforms existing baselines on all metrics, confirming the effectiveness of our model in multimodal sarcasm detection. 2) All of the trimodal sarcasm detection approaches are superior to just text modality detection methods, demonstrating the complementarity of text, image, and audio modalities, thereby enhancing the performance of sarcasm detection. 3) All models exhibit lower performance in the speaker-independent setup compared to the speaker-dependent one. This is attributed to the absence of speaker overlap between the training and testing sets in the speaker-independent configuration, making it more challenging than the speaker-dependent setting. 4) In both the speaker-dependent and speaker-independent settings, our model ULIL achieves an F1-score value that is 2% and 2.8% higher than the most recent baseline IWAN, respectively.

In summary, the experimental results validate the superiority of the proposed ULIL model for multimodal sarcasm detection. The performance improvement can be attributed to the advancement of ULIL. First, in the feature extraction layer, it not only extracts unimodal features at the utterance level but also fuses multimodal features at the utterance level. Second, it utilizes cross-attention and self-attention to extract information about inconsistency between modalities, leading to more accurate sarcasm detection.

VI. ABLATION STUDIES

To further analyze our proposed model and assess the necessity of the three modalities, we conduct ablation experiments on the dataset with different modalities removed. As text modality is the most fundamental, we separately removed audio and vision modalities and recorded the experimental results in Table III. Removing either audio or vision modality leads to a performance drop, indicating the crucial importance of these two modality features for the sarcasm detection task. Audio and vision modality features complement text features, enabling a more comprehensive capture of inconsistent features.

On the other hand, to further analyze the effectiveness of the main modules in ULIL, we removed the primary modules,

TABLE III
ABLATION STUDIES ABOUT DIFFERENT MODALITIES

Modality	Speaker-Dependent			Speaker-Independent		
	P	R	F1	P	R	F1
T,V	69.3	69.2	69.2	63.5	63.8	63.6
T,A	71.1	70.9	71.0	60.6	61.2	60.9
T,A,V	**76.9**	**76.1**	**76.5**	**73.1**	**72.6**	**72.8**

TABLE IV
EXPERIMENTAL RESULTS ABOUT MAIN MODULES

Modality	Speaker-Dependent			Speaker-Independent		
	P	R	F1	P	R	F1
w/o $\mathcal{U}$	74.8	74.1	74.4	69.6	70.3	69.9
w/o $\mathcal{I}$	75.9	75.1	75.5	70.6	70.9	70.7
ULIL	**76.9**	**76.1**	**76.5**	**73.1**	**72.6**	**72.8**

the multimodal utterance-level attention module (w/o $\mathcal{U}$) and the incongruity representation module (w/o $\mathcal{I}$). According to the results in Table IV, it is evident that the performance deteriorates when the utterance-level attention module is not used, underscoring its capability to capture crucial sentiment information. By comparing w/o $\mathcal{I}$ with ULIL, we can deduce the significance of incongruity feature representation in multimodal sarcasm detection.

VII. CONCLUTION

In this paper, we investigate a comprehensive and robust approach that employs an Utterance-level Incongruity Learning Network (ULIL) to address multimodal sarcasm detection tasks. Unlike previous works, we extract utterance-level unimodal features and multimodal features fused by a multimodal utterance-level attention. Additionally, we explore an incongruity learning network for sarcasm detection. Specifically, considering the sarcasm utterances process incongruous characteristics, we deploy a cross-attention and self-attention mechanism to acquire incongruities between primary modality and auxiliary modalities. Our proposed model has been experimentally tested on a public benchmark dataset, demonstrating a significant performance improvement compared to a series of baselines for the multimodal sarcasm detection task.

VIII. ACKNOWLEDGMENTS

This work was partially supported by the National Key Research and Development Project of China (No. 2022YFF0903203) and the Key Scientific and Technological Project of Henan Province, China (No. 222102210040).

REFERENCES

[1] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 2506–2515.

[2] Y. Tay, L. A. Tuan, S. C. Hui, and J. Su, "Reasoning with sarcasm by reading in-between," *arXiv preprint arXiv:1805.02856*, 2018.

[3] C. Lou, B. Liang, L. Gui, Y. He, Y. Dang, and R. Xu, "Affective dependency graph for sarcasm detection," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1844–1849.

[4] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an _obviously_ perfect paper)," *arXiv preprint arXiv:1906.01815*, 2019.

[5] D. S. Chauhan, S. Dhanush, A. Ekbal, and P. Bhattacharyya, "Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4351–4360.

[6] Y. Wu, Y. Zhao, X. Lu, B. Qin, Y. Wu, J. Sheng, and J. Li, "Modeling incongruity between modalities for multimodal sarcasm detection," *IEEE MultiMedia*, vol. 28, no. 2, pp. 86–95, 2021.

[7] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 704–714.

[8] A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing context incongruity for sarcasm detection," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 757–762.

[9] D. I. H. Farías, V. Patti, and P. Rosso, "Irony detection in twitter: The role of affective content," *ACM Transactions on Internet Technology (TOIT)*, vol. 16, no. 3, pp. 1–24, 2016.

[10] N. Pawar and S. Bhingarkar, "Machine learning based sarcasm detection on twitter data," in *2020 5th international conference on communication and electronics systems (ICCES)*. IEEE, 2020, pp. 957–961.

[11] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. [Online]. Available: https://aclanthology.org/D14-1181

[12] S. Hiai and K. Shimada, "Sarcasm detection using rnn with relation vector," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 15, no. 4, pp. 66–78, 2019.

[13] T. Xiong, P. Zhang, H. Zhu, and Y. Yang, "Sarcasm detection with self-matching networks and low-rank bilinear pooling," in *The world wide web conference*, 2019, pp. 2115–2124.

[14] X. Wang, Y. Dong, D. Jin, Y. Li, L. Wang, and J. Dang, "Augmenting affective dependency graph via iterative incongruity graph learning for sarcasm detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, 2023, pp. 4702–4710.

[15] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, and R. Xu, "Multi-modal sarcasm detection via cross-modal graph convolutional network," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 1767–1777.

[16] Y. Qiao, L. Jing, X. Song, X. Chen, L. Zhu, and L. Nie, "Mutual-enhanced incongruity learning network for multi-modal sarcasm detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, 2023, pp. 9507–9515.

[17] Y. Wang, Y. Li, P. Bell, and C. Lai, "Cross-attention is not enough: Incongruity-aware multimodal sentiment analysis and emotion recognition," *arXiv preprint arXiv:2305.13583*, 2023.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7216–7223.

[20] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proceedings of the 19th ACM international conference on multimodal interaction*, 2017, pp. 163–171.

[21] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

liujing Song received his Bachelor's degree from Zhengzhou University. Now she is a Ph.D. student in the University of Chinese Academy of Sciences. His research focuses on natural language processing and data mining.

Zefang Zhao received his Bachelor's degree from Taiyuan University of Technology. Now he is a Ph.D. student in the University of Chinese Academy of Sciences. His research focuses on deep learning, natural language process and sentiment analysis.

Yuxiang Ma is currently an associate professor in the School of Computer and Information Engineering at Henan University. He received the B.S. degree from Henan University in 2013, and the Ph.D. degree from the Computer Network Information Center, Chinese Academy of Sciences in 2019. His main research interests include network security, mobile computing, and privacy enhancement technologies.

Yuyang Liu received the Ph.D. degree from University of Chinese Academy of Sciences. He is currently a Research Associate Professor in Institute of Medical Information, Chinese Academy of Medical Sciences. His research interests include complex medicial knowledge networks, clinical decision support system, and various data mining and artificial intelligence applications across medical informatics.

Jun Li is a research fellow and doctoral supervisor at the Computer Network Information Center of Chinese Academy of Sciences, specially appointed researcher of Chinese Academy of Sciences. His main research interests are artificial intelligence and big data technical applications and future Internet architecture.

Anomaly Detection During Additive Processes for DLP 3D Printing

1st Hyejin S. Kim,2nd Hyonyoung Han, 3rd Ji Yeon Son
Intelligent Manufacturing Integration Lab., ETRI, Daejeon, Korea
marisan@etri.re.kr, hyonyoung.han@etri.re.kr, jyson@etri.re.kr

***Abstract*—Additive manufacturing is gaining attention in various fields such as medical applications, aerospace, defense, and complicated manufacturing industries. This is due to the advantages of additive manufacturing including reduced logistical constraints and the ability to produce customized products. However, the materials used in additive manufacturing are generally expensive and highly sensitive to changes in external conditions. For these reasons, it is crucial from a productivity standpoint to monitor the additive manufacturing process closely to detect any anomalies early on and decide whether to continue with the layering process.**

In this paper, we developed an algorithm that takes camera footage as input to determine the quality of the additive manufacturing output. We achieved an accuracy rate of 99.65%. Additionally, to simulate rare abnormal conditions, we used computer graphics to define nine different abnormal states and generated data for these conditions.

***Index Terms*—Additive manufacturing, DLP(Digital Light Processing), Anomaly Detection, Image, Anomaly Monitoring**

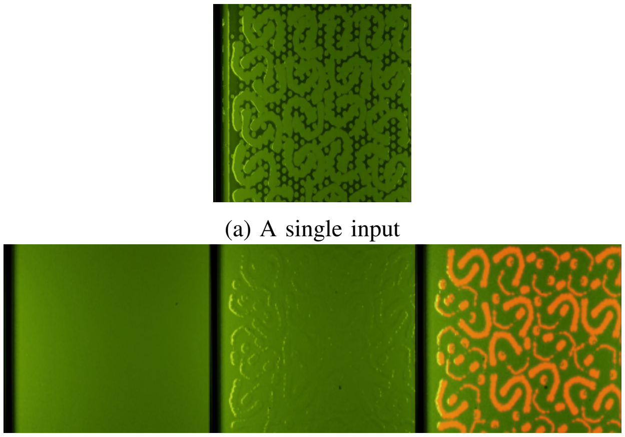
I. Introduction

Additive manufacturing has gained attention as a future digital manufacturing tool, especially when all manufacturing processes are connected via IoT. It is also considered a solution for the decline in skilled low-wage labor. One of the greatest advantages of additive manufacturing is its ability to overcome spatial constraints. That is, it allows for remote control of the manufacturing process, eliminating the need for 'supply chain' logistics and 'inventory management.' This makes it highly promising for applications in space. Additionally, it holds promise in sectors that require complex part geometries, high precision, design flexibility, and a high demand for customized production. However, additive manufacturing is sensitive to external factors such as vibration of the 3D printing equipment and temperature changes. Also, the infiltration of fine particles into the machine can cause issues in the layering process. This process has a high likelihood of defects occurring during the printing process. Not only is the process slow, but the materials are also expensive. Therefore, it is crucial to determine the quality of the output during the additive manufacturing process. This allows for decisions to be made on whether to continue the output or make corrections to the defects.

In [1], a dataset for anomaly detection in Fused Deposition Modeling (FDM) additive manufacturing has been released. There has also been research on anomaly detection based on UNet for metal powder bed additive manufacturing citemeta-anomaly. This paper proposes an anomaly detection method for the additive manufacturing process in the field of liquid material-based additive manufacturing.

II. Methods

Our research purpose is to detection anomalies during additive processing, especially for DLP additive manufacturing device. The main material for DLP is liquid type. The additive process requires three steps: the baseplate moving upward, the baseplate descending to the correct position, and exposure to UV light. Between each DLP printing step, the slave moves to remove excess residue.


(a) A single input

(b) Three consecutive inputs

Fig. 1: Comparision between single input and consecutive inputs

In general, anomalies seldom appear during manufacturing process. Therefore, supervised approach is hard to apply to solve in this setting. So, to make our data closely resemble real-world data, we used Computer Graphics to generate synthetic data. At first, we thought generating the Ground Truth for CG would be a straightforward task. In the normal state, it was a matter of creating pixel-wise mask data when making CG for abnormal points. However, working with CG turned out to be not as simple as expected. When designating areas affected by abnormal light reflections in the CG as abnormal state areas, approximately one-third of the entire image was marked as abnormal, while in reality, the abnormal areas were much smaller. Because the abnormal area was too large, it was reduced in size, and the newly created mask image took the form of occupying a smaller area than the actual abnormal region. Since the Ground Truth based on the mask image was not accurate, it was challenging to use the information about the occupied area for training. Because the abnormal area was too large, it was reduced in size, and the newly created mask image took the form of occupying a smaller area than the actual abnormal region. So, we attempted to solve this problem using the simplest classification method. Typically, when analyzing image sequences, it has been common to combine Convolution and LSTM. For early detecting anomalies, we attempt to design two networks:time-aware prediction network and CNN-based neural network. To reflect the three stages in the stacking process, we collected one image at each stage and combined them into a single input, resulting in three images as shown in Fig. 1.

III. Experimental Results

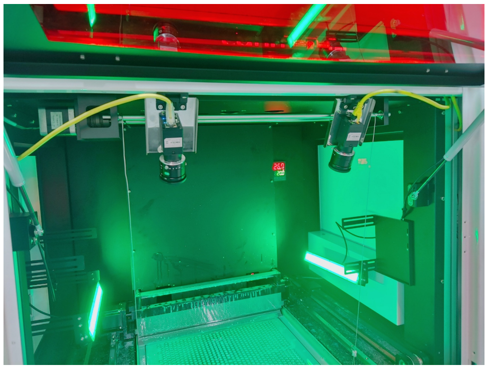
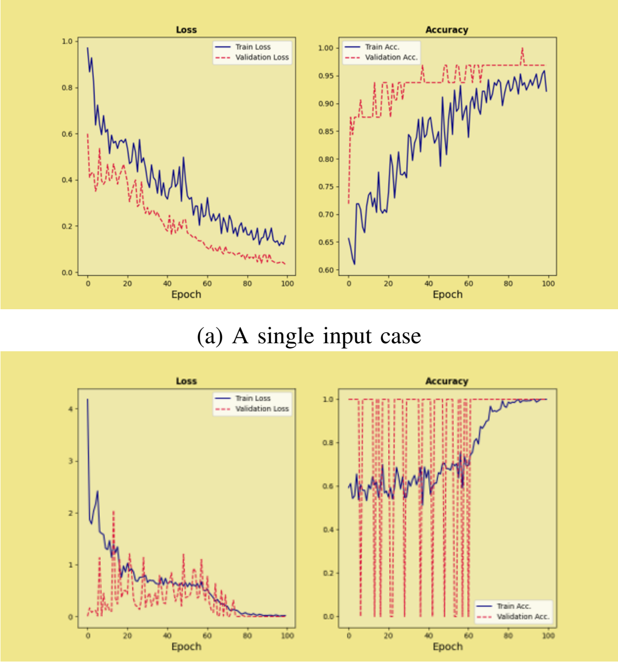


Fig. 2: The system of DLP Anomaly Monitoring System

We have installed two cameras and two green lights on the DLP 3D printing machine. We have set up a system where the camera device and the DLP device communicate over TCP/IP, allowing central PC control and the ability to receive and store image sequence data on the PC as shown in Fig. 2. The build size of this 3D printer is $400x330x500mm^3$, offering a wide printing area. For Dental models, it can print up to 60 pieces at once. To cover this entire area, two cameras and two green lights have been installed. Particularly, since the resin's color is gray, and texture changes in the image are not well captured in low light, two green lights were installed on both walls. Each camera is a machine vision camera, initially mounted vertically to precisely detect the size of defect areas. However, due to the UV light being scattered vertically within the resin, there was a significant issue of light reflection observed in the cameras. To avoid this phenomenon, the camera's position was adjusted to an inclined direction rather than vertical. Our device initially targeted ivory-colored resin. However, after the purchase of the lighting equipment, the color suddenly changed to gray resin. Due to the gray color of the resin, making the texture less visible, we used jigs to mount the lighting closer to the resin, allowing the light to shine near the resin.

We have configured the system to receive signals for slave movement and stoppage via TCP/IP, allowing the camera to capture video only when the slave comes to a halt. The captured video is then transmitted to the central PC, where a program for anomaly detection receives this video. Continuous monitoring for anomalies is performed while printing is in progress to detect any abnormal situations.



(a) A single input case

(b) Three consecutive input case

Fig. 3: Loss and Accuracy Comparision Graph

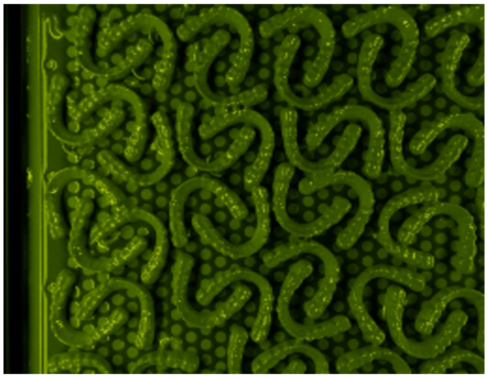


Fig. 4: An Error Sample

We construct additive manufacturing image sequence database consisting of five normal case image sequences and five abnormal cases using blender software [4]. A single sequence consists of 427 images. In the case of our proposed method, the length of input becomes 142 with a single last output image. To obtain UV output information, we included slicing images as inputs, resulting in a total of four images being used as input. However, experimental results showed similar performance even when excluding the slicing information, so we omitted the slicing information in the final proposed method.

To compare the performance of the model with three inputs and the model with one input, we compared the results in terms of loss and validation accuracy (see Fig. 3). The three concatenated input case in Fig. 3(b) obtains better performance than the single input case shown in Fig. 3 (a). Fig. 4 is an example of an error image evaluated using the proposed method.

IV. Conclusion

This proposed method has not yet been applied to real data and remains based on synthetic data. This is because the state of the DLP device for acquiring real data has been continuously changing. The initial version of the DLP equipment did not have a red-colored glass window. Therefore, initially, we generated CG data that closely resembled the acquired real data. However, upon acquiring actual data after the completion of the equipment, the red color of the glass window caused a transformation to a color closer to yellow rather than green. Furthermore, it was observed that this color was not consistent but changed continuously depending on the position of the slave, as the red glass window color was repeatedly reflected. The uniqueness of real data due to these red artifacts differs from synthetic data, and it is expected that there may be a performance degradation in real data.

V. Acknowledgement

This work was supported by internal fund/grant of Electronics and Telecommunications Research Institute(ETRI) [21YR3200, Development of autonomous recognizing and grasping technology for transferring of 3D-printed parts], Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub), the Technology Innovation Program (or Industrial Strategic Technology Development Program) (P0023760, EUV Pellicle/Mask Database Construction and Defect Detection AI Solution for EUV Semiconductor Productivity Improvement) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea), KIAT(Korea Institute for Advancement of Technology) and Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [23ZR1100, A Study of Hyper-Connected Thinking Internet Technology by autonomous connecting, controlling and evolving ways]

References

[1] Joanna Sendorek, Tomasz Szydlo, Mateusz Windak and Robert Brzoza-Woch, ”Dataset for anomalies detection in 3D printing,” arXiv.abs/2004.08817, 2020.

[2] T. Lalitha, N. K. Anushkannan, S. Shreepad, S. Sasireka, H. Anandaram and S. Razia, ”Deep Learning-based Automatic 3D Printer Anomaly Detection during the Printing Process,” 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2022, pp. 1343-1348, doi: 10.1109/ICOSEC54921.2022.9951903.

[3] Luke Scime, Derek Siddel, Seth Baird, Vincent Paquit, ”Layer-wise anomaly detection and classification for powder bed additive manufacturing processes: A machine-agnostic algorithm for real-time pixel-wise semantic segmentation, Additive Manufacturing,” vol. 36, 2020,

[4] https://www.blender.org/

Session 1C: Security & Blockchain 1

Chair: Dr. Pham Dinh Lam, Kyonggi University, Korea, ,

1 Paper ID : 20240396, 53~56

Generalized Parabola Chaotic map for Pseudorandom Random Number Generator

Dr. Nattagit Jiteurtragool,

King Mongkut's University of Technology North Bang. Thailand

2 Paper ID : 20240324, 57~62

Security Analysis of Android Applications for Hotel and Flight Booking Applications

Ms. Vatcharavaree Wongsuna, Prof. Sudsanguan Ngamsuriyaroj,

Faculty of ICT, Mahidol University. Thailand

3 Paper ID : 20240309, 63~66

The development of new system for generating training data of AI-based anomaly detection

Ms. Thi My Truong, Dr. Won Seok Choi, Mr. Jang Hyeon Jeong, Prof. Seong Gon Choi,

Chungbuk National University. Korea(South)

4 Paper ID : 20240272, 67~72

Router Penetration Testing Based on CEM Vulnerability Assessment Criteria

Ms. Tai-Ying Chiu, Mr. Bor-Yao Tseng, Mr. Bagus ATMAJA, Prof. Jiann-Liang Chen,

National Taiwan University of Science & Technology. Taiwan

5 Paper ID : 20240269, 73~78

Hybrid Clustering Mechanisms for High-Efficiency Intrusion Prevention

Ms. Pin-Shan Lin, Mr. Yi-Cheng Lai, Ms. Man-Ling Liao, Ms. Shih-Ping Chiu, Prof. Jiann-Liang Chen,

National Taiwan University of Science & Technology. Taiwan

Generalized Parabola Chaotic map for Pseudorandom Random Number Generator

Nattagit Jiteurtragool

Department of Computer and Information Sciences, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand
nattagit.j@sci.kmutnb.ac.th

Abstract— **In this paper, a generalized form of chaotic map based on nonlinear function with parabolic shape is introduced. The study involves the investigation of chaotic dynamics in terms of apparent in time-domain, and both qualitatively and quantitatively examination using bifurcation diagram and Lyapunov exponents. Furthermore, the practical application of these parabolic chaotic maps is showcased in a pseudo-random number generator, with its performance evaluated using statistical tests from the NIST SP800-22 test suite.**

***Keywords*— chaotic map, discrete-time chaotic, parabola function, pseudo random number generator, NIST**

I. Introduction

Chaotic behaviours exist in wide range of both natural and man-made phenomena [1]. Over the years, there has been a remarkable surge of interest in nonlinear systems which exhibit chaotic behaviour, with various areas of applications such as behaviour study, communication, cryptography, and system control [2-4]. The study of chaotic behaviours often involves the use of a discrete-time dynamical model known as a chaotic map.

In the meantime, countless studies of random number generator (RNG), which is a computational or hardware-based mechanism that able to produces a randomness sequence of number, has undergone thorough examination, as random numbers hold a crucial role in cryptography. This scrutiny encompasses intricate mathematical models, numerical simulations, and extensive statistical research. RNG can theoretically divided into 2 categories, true random number generator (TRNG), which mainly relied on physical phenomenal (non-deterministic), and pseudorandom number generator (PRNG), which is mathematical algorithms (deterministic). Classification of RNGs is shown in Fig.1. TRNG mostly relied on non-deterministic phenomena in the nature such as thermal noise, quantum state, electrical noise, even atmospheric noise. These phenomena will then be captured and digitized. Ideally, TRNG is expected to produce results that are unpredictable, and statistically unbiased. In contrast, PRNG, which relied on deterministic system which driven by mathematical algorithms, cannot expect a true randomness result. However, these algorithms, which beginning with an initial seed value in order to produce pseudo-random sequences, can be characterized by good statistical properties, rapid execution, repeatability, and reproducibility. All things considered, the chaotic map, which is a system that exhibit the property of extremely sensitivity to initial condition, and considered a deterministic system, is a suitable candidate as randomness sources for a pseudorandom random number generator.

Inspired by prior research [5], a generalized form of chaotic map algorithm based on nonlinear function with parabolic shape is presented in this paper. The parabola chaotic map is based on various parabola functions including the absolute value function, which is regarded as a linearized parabolic function, were employed. The chaotic characteristics of the proposed chaotic maps were explored using bifurcation diagram, and Lyapunov exponents. The findings reveal that the parabola chaotic maps demonstrate intermittently chaotic behaviour, while the linearized parabola chaotic map demonstrates robust chaos across a broader spectrum of parameter. Furthermore, to illustrate this concept practically, a pseudo random number generator utilizing parabola chaotic map was also proposed. The resulting random number output is subsequently assessed through the standard NIST SP800-22 test suite.

II. CHAOTIC MAP

A. *Previous work*

In addition to the well-known such as tent map, logistic map and gauss map, the Sigmoidal chaotic map developed in our previous work [5] which described as

$$x_{n+1} = \mp A f_{\mathrm{NL}}(B x_n) \pm C x_n \pm D \qquad (1)$$

where the parameters A, B, C, and D are real constants, x_n is a real variable, and $f_{\mathrm{NL}}(x)$ is an S-shaped nonlinear function or so-called sigmoidal. It is seen in (1) that the sigmoidal chaotic map is different from typical chaotic map since this chaotic map can offer different behaviour regarding the utilized sigmoidal function.

B. *Proposed Generalized Parabola Chaotic Map*

Motivated by the exiting sigmoidal chaotic map where the nonlinear function can be replaceable by any S-shaped nonlinear functions. This raises the question of "is it possible to replace the sigmoidal functions with other nonlinear functions in the exiting chaotic map?". Thoroughly investigations have led to a chaotic map using parabola functions, presented in this paper. The proposed parabola chaotic map can be defined by the simple mathematical model of the form

$$x_{n+1} = \mp A f_{\mathrm{NL}}(Bx_n) \pm C \tag{2}$$

where the parameters A, and B are real constants, x_n is a real variable, and the $f_{\mathrm{NL}}(x)$ is a parabola function. Moreover, it is apparent that the proposed chaotic map in (2) is a conjugation of two chaotic map which can also be expressed as

$$x_{n+1} = A f_{\mathrm{NL}}(Bx_n) - C \tag{3}$$

$$x_{n+1} = -A f_{\mathrm{NL}}(Bx_n) + C \tag{4}$$

The summarization of the chaotic maps based on (2) using various parabola functions is shown Table I. In term of mathematical, the f_1 and f_2 are polynomial functions with even degree and a Gaussian function, respectively, and f_3 is based on inverse trigonometric function such a hyperbolic cosine. Meanwhile f_4 is an absolute function which considered linearized parabolic shape function.

TABLE 1. CHAOTIC MAPS BASED ON PARABOLIC SHAPE NONLINEAR FUNCTIONS F_{NL}(X)

Parabolic Shape Functions	Descriptions	Parabola Chaotic Maps
$f_1(x) = (x)^N$	Even Degree Polynomial	$x_{n+1} = \mp A(Bx_n)^N \pm C$
$f_2(x) = e^{-x^2}$	Gaussian Function	$x_{n+1} = \mp Ae^{-(Bx_n)^2} \pm C$
$f_3(x) = cosh(x)$	Hyperbolic Cosine	$x_{n+1} = \mp A\,cosh(Bx_n) \pm C$
$f_4(x) = \lvert x \rvert$	Absolute	$x_{n+1} = \mp A\lvert Bx_n \rvert \pm C$

To exploration of chaotic behaviour both qualitatively and quantitatively, a bifurcation diagram and a Lyapunov Exponent (LE) can be used respectively. A bifurcation diagram offers insights into the possible long-term behaviours of a system, including fixed points or periodic orbits, by varying a bifurcation parameter. Conversely, the Lyapunov Exponent, a well-recognized indicator of chaos, serves a metric for assessing chaos by reflecting the average divergence between two closely related trajectories within a dynamic system [6]. The LE equation can be defined as

$$\mathrm{LE} = \lim_{n\to\infty} \frac{1}{N} \sum_{n=1}^{N} \log_2 \frac{\mathrm{d}x_{n+1}}{\mathrm{d}x_n} \tag{5}$$

where N is the number of iterations. A positive value of LE indicate chaotic behaviours of dynamical systems and the larger value is the higher degree of chaoticity.

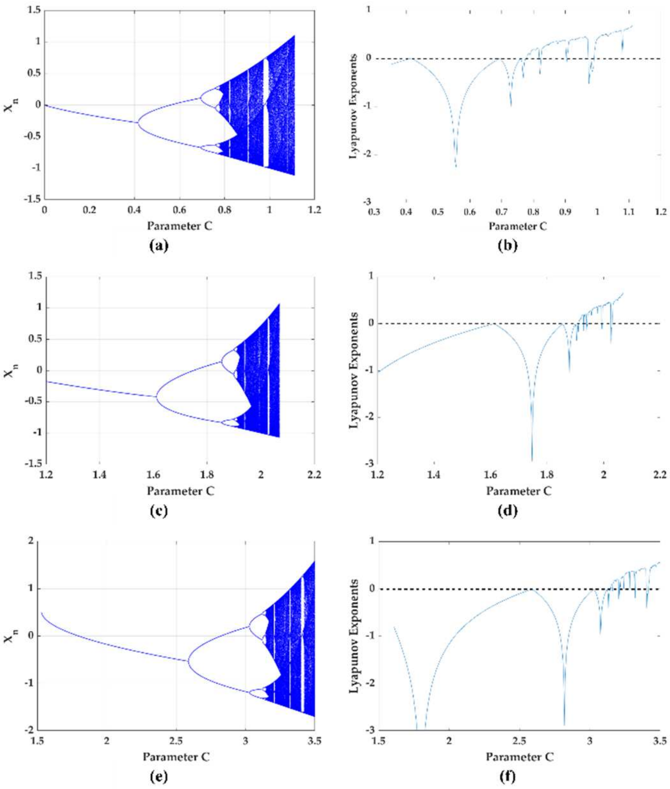


Figure 1. Bifurcation diagrams (left) and Lyapunov exponent (LE) plots (right) of the parabola chaotic maps with A = 1.8 and B = 1; (a, b) for function f_1, (c, d) for function f_2, and (e, f) for function f_3.

To evaluate the chaotic characteristics and the continuity of the proposed parabolic chaotic maps defined in Equation (3), both bifurcation diagrams and LE plots were utilized, as depicted in Fig. 1. For this investigation, parameters A and B were set to 1, while parameter C was used as a bifurcation parameter. The bifurcation diagrams in Figs. 1a, 1c, and 1e illustrate a shared chaotic behaviour among the parabolic chaotic maps corresponding to functions f_1 through f_3. It is noticeable that these bifurcation diagrams reveal periodic windows and demonstrate discontinuous chaotic behaviour, aligned with the observations in the LE plots. However, this implies that, across all three cases offer robust chaos for some portion of parameter C.

In addition to the utilization of nonlinear functions with parabolic transfer characteristics, the proposed parabolic chaotic map can also integrate a triangular function, such as an absolute value function, which can be seen as a linearized variant of the parabolic function. Figure 2 provides a visual representation of the bifurcation diagram and LE plot for the parabolic chaotic map based on the absolute value in Equation (8). In this depiction, parameter A were carefully chosen to serve as the variable responsible for generating the bifurcation, while parameters B and C remain set at 1. As observed in the corresponding LE plot, the bifurcation diagram of the linearized parabolic chaotic map consistently exhibits chaotic behaviour across the entire range of parameter A.

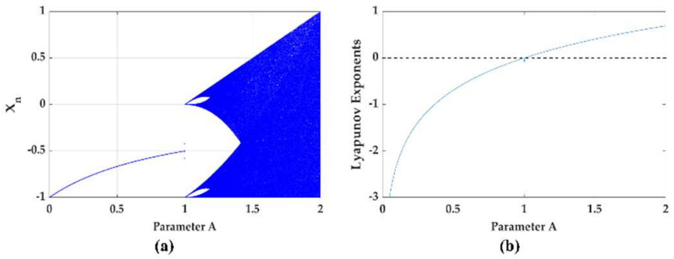


Figure 2. Bifurcation diagrams (left) and Lyapunov exponent (LE) plots (right) of absolute value function based parabola chaotic map at specified parameters B = 1 and C = 1.

III. DESIGN AND ANALYSIS OF PSEUDORANDOM RANDOM NUMBER GENERATOR

A. Random Number Generator

As mentioned, a RNG is a computational algorithm or device that generates numbers or sequences of numbers that appears random or unpredictable. While there are two primary types of RNGs one is True Random Number Generators (TRNGs) and the other is Pseudo-Random Number Generators (PRNGs). These random number generators still share common structure which consisted of several key components that work together as shown in Fig.3.

Entropy Source	Entropy Harvester	Post Processor
- Electrical Noise - Quantum Signal - Thermal Noise - Atmospheric Noise - Chaotic Signal - ….	- Periodic Sampling - Nonuniform Sampling - Comparator - D-Flip Flop - ….	- Von Neumann - XOR - LFSR - Bit Skipping - Quasi-Shift Register - ….

Figure 3. Typical structure of a random number generator.

The key element in any RNG is the entropy source, which is crucial for providing unpredictability and serves as the foundation of RNG security. The role of the entropy harvester was significant in efficiently gathering the raw signal without compromising the integrity of the entropy source. It reads the data generated by the entropy source and converts it into a series of bits, which are referred to as raw bit data. Although, the output from the entropy harvester comprises a random bit sequence, these raw bits sequence often exhibit bias and lack of uniform distribution. Therefore, a post-processor remains necessary to rectify this issue, ensuring that the output becomes uniformly distributed and addressing any statistical imperfections in the generated sequences. Even though the result from the entropy harvester is a random bit sequence, the raw bits are usually biased and not uniformly distributed for various reasons. Therefore, the post-processor is still required to make output uniformly distributed as well as improve the statistical imperfections of the generated sequences.

B. Pseudorandom Random Number Generator based on Generalized Parabola Chaotic Map

The classification of a chaotic-based random number generator as a TRNG or a PRNG is depended on its specific implementation. When a chaotic based random number generator employs unpredictable physical processes like the chaotic behaviour of electronic circuits or the unpredictability of systems such as a double pendulum to produce random numbers, it qualifies as a TRNG. This is due to its reliance on inherently unpredictable physical phenomena to generate genuine randomness. On the contrary, if the chaotic-based random number generator employs a deterministic algorithm to mimic chaotic behaviour such as chaotic map equation to establish an initial seed value, subsequently using algorithms to generate a sequence of numbers, it falls into the category of PRNG.

In order to construct PRNG based on generalized parabola chaotic map, architecture is shown in Fig.4. The proposed architecture of the PRNG consisted of the absolute value function based parabola chaotic map as an entropy source, a comparator with threshold value used as an entropy harvester to turn raw output signal from chaotic map into raw bit data.

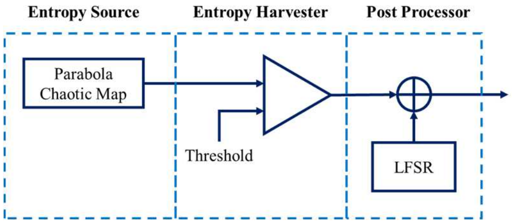


Figure 4. The proposed architechture of Pseudorandom Random Number Generator based on Generalized Parabola Chaotic map.

The selection of the comparator's threshold value is a deliberate process. Shannon's entropy is employed to assess the impact of this threshold and the resulting level of entropy, essentially measuring the system's randomness. Shannon's entropy can also be defined as follows:

$$H = -\sum_{i=0}^{1} P_i \, log_2 P_i \qquad (6)$$

The calculated entropy over the range of parameter threshold is depicted in Figure 5(a).

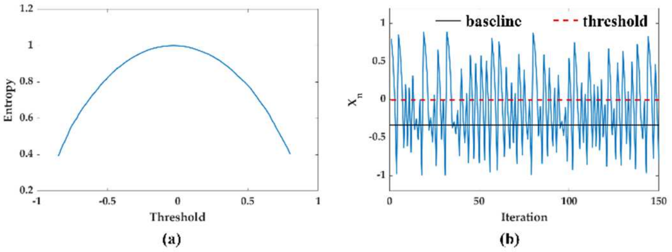


Figure 5. Typical structure of a random number generator.

Through, Chaos signals are characterized by their sensitivity to initial conditions, non-periodic behavior, and a broad range of frequency components, the plot of chaotic waveforms in time-domain, as shown in Figure 5(b), suggest that the baseline value fall between 0 and 0.5. Nevertheless, it's noticeably from the plot of entropy versus threshold indicate the value of threshold differences from the baseline.

As for the post processor, an 8-bit Fibonacci linear feedback shift register (LFSR) were selected. Fibonacci linear feedback shift register is a chain of shift registers where the inputs are linear function of their previous state. It's referred to as a "Fibonacci" LFSR as it employs a linear combination of its internal bits, similar to the mathematical Fibonacci sequence. In order to enhance the statistical properties of PRNG, the generated raw bit data then combined with the LFSR post processor through using exclusive-or (XOR) operation.

C. Randomness Evaluation

Although there are many evaluations for random number generator, the random bit output of the proposed PRNG has been examined, in terms of statistical properties, using the standard NIST SP800-22 test suite. The NIST test suite, which issued by the National Institute of Standards and Technology, is a statistical test consists of 15 tests [7]. The NIST test suite also generally accepted as a standard test suit for both TRNG and PRNG. The test can be used to analyse the number sequence by identifying any pattern of value that indicates non-randomness within the sequences. This is achieved through the assessment of probability values (P-value), a valuable metric for determining statistical property of each test. The P-value indicates a randomness of the test sequences where the P-value exceeds 0.01, the sequence can be confidently regarded as random and accepted. Conversely, if the p-value is below this threshold, the sequence is considered as non-random.

TABLE 2. NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST) STATISTICAL TEST SUITE.

Test Methods	*P*-value	Pass Ratio	Result
Frequency	0.6255	1.00	Pass
Block Frequency (m=128)	0.9307	1.00	Pass
Runs	0.5236	1.00	Pass
Longest Run	0.2717	1.00	Pass
Binary Matrix Rank	0.3534	1.00	Pass
Discrete Fourier Transform	0.9487	1.00	Pass
Non-overlapping Template	0.0734	1.00	Pass
Overlapping Template	0.9131	1.00	Pass
Universal Statistical	0.1121	1.00	Pass
Linear Complexity	0.3230	1.00	Pass
Serial (1)	0.5057	1.00	Pass
Serial (2)	0.6735	1.00	Pass
Approximate Entropy	0.3009	1.00	Pass
Cumulative Sums	0.7857	1.00	Pass
Random Excursions	0.7535	1.00	Pass
Random Excursions Variant	0.6339	1.00	Pass

The statistical properties of the proposed PRNG were evaluated using 10 Mbit data. The generated bit sequence is divided into 10 sequences with the length of 1Mbit for each block. The calculated *P*-values for ever test of the NIST is shown in Table 2. The tests result has shown that the proposed PRNG can pass all the statistical tests of the NIST.

IV. CONCLUSION

Chaotic map played an important role in various field of application such as information security and cryptography. In this paper, A generalized chaotic map based on parabola functions. Chaotic dynamics were described in terms of apparent in time-domain, and both qualitatively and quantitatively examination using bifurcation diagram and Lyapunov exponents. The PRNG which utilized the proposed parabola chaotic maps were also demonstrated and evaluation result of the random bit sequences were evaluated using the standard NIST SP800-22 test suite and pass all the tests.

REFERENCES

[1] Jensen, Roderick V, "Classical Chaos." *American Scientist*, vol. 75, no. 2, 1987.

[2] M. Rosalie, et al, "Chaos-enhanced mobility models for multilevel swarms of UAVs," *Swarm and Evolutionary Computation*, 41, 2018, p.36-48.

[3] Kaddoum, Georges, "Wireless chaos-based communication systems: A comprehensive survey," *IEEE Access*, 4, 2016, pp.2621-2648.

[4] Fradkov, Alexander L., and Robin J. Evans., "Control of chaos: Methods and applications in engineering," *Annual reviews in control*, 29(1), 2005, p. 33-56.

[5] Jiteurtragool N, Masayoshi T., and San-Um W., "Robustification of a one-dimensional generic sigmoidal chaotic map with application of true random bit generation," *Entropy*, 20(2),2018, p. 136.

[6] Abarbanel, Henry DI, Reggie Brown, and M. B. Kennel, "Lyapunov exponents in chaotic systems: their importance and their evaluation using observed data," *International Journal of Modern Physics B,* 5(09), 1991, p. 1347-1375.

[7] A. Rukhin, et al, "A statistical test suite for random and pseudorandom number generators for cryptographic applications," *National Institute of*

[8] *Standards and Technology (NIST)*, special publication 800-22, August 2008.

[9] Cicek, Ihsan, Ali Emre Pusane, and Gunhan Dundar, "A novel design method for discrete time chaos based true random number generators," *Integration*, 47(1), 2014, p. 38-47.

[10] Stipčević, Mario, and Çetin Kaya Koç, "True random number generators," *Open Problems in Mathematics and Computational Science*, Cham: Springer International Publishing, 2014.

[11] Yu, Fei, et al, "A survey on true random number generators based on chaos," *Discrete Dynamics in Nature and Society,* 2019.

Security Analysis of Android Applications for Hotel and Flight Booking Applications

Vatcharavaree Wongsuna, Assoc. Prof. Sudsanguan Ngamsuriyaroj

Faculty of Information and Communication Technology, Mahidol University, Thailand

vatcharavaree.wog@student.mahidol.ac.th, sudsanguan.nga@mahidol.ac.th

Abstract— **The tourism industry's exponential growth has currently led to an increase in hotel and flight bookings, especially through mobile applications. There are various tools and platforms to build a mobile application, and each mobile function may need specific permissions to access certain features of a mobile device, and that would make devices and applications vulnerable. In this work, we aim to conduct a security analysis of many Android applications for hotel and airline bookings. Using an open-source tool, the security-related features of 20 Android apps are extracted from the domains and divided into six groups, including security mechanisms, shared Android components, dangerous privilege calls, and three code-related groups. Those applications were then clustered according to their features, and we found that hotel applications had Android security features, such as the RSA algorithm and detecting jailbreak. Furthermore, third-party apps were more likely to share Android components with other apps on the same device. Future research could focus on analyzing Android apps for hotel or airline booking exclusively and expanding the dataset size to identify and analyze patterns effectively.**

Keywords— **Android Application Security, security Analysis**

I. Introduction

Mobile phones and other smart devices, such as tablets, have evolved from being mere communication devices to becoming more functional. They now offer features such as camera capabilities, mapping and location services, network connectivity, and the ability to record audio. Moreover, these features are used by mobile applications to serve the business needs. So, they may require accessing and compromising some sensitive features such as camera, microphone, or storage. This could have far-reaching implications in terms of privacy and security.

According to CVEdetails [1], Android was found to be the most vulnerable product in 2017. The vulnerabilities reported were mainly related to overflow, code execution, and information disclosure which can compromise sensitive information. Moreover, Pradeo's report [2] revealed that over 60% of mobile applications failed to prevent data leakage or corruption, even though they had policies and permissions in place. The study also identified 75% of vulnerabilities listed in the OWASP top ten flaws and requested more permission than the app's purpose.

Therefore, this research aims to conduct a security analysis using an automatic tool to extract security-related features from 20 well-known Android applications for hotel and flight booking. We do the static analysis technique using open-source tools and eventually perform the clustering analysis using a data mining tool such as RapidMiner [13] to identify any security patterns. However, there are some scopes of this work as outlined below.

- This research will perform only static analysis techniques to assess the application information. The analysis will involve examining the application's source code, manifest.xml file, and the security features implemented by the apps.
- Clustering analysis will be performed using the Rapid Miner data mining tool with the K-means algorithm and elbow method.
- The OWASP MSTG [4] will be used as a reference in this research, but not all aspects will be covered.
- The dataset consists of 20 Android applications from the travel category. These applications can be categorized into three types: applications that provide booking services for both hotel rooms and flights, applications developed by airline companies, and applications developed by hotels.
- Last, these 20 Android applications are not the latest versions and mostly date back to the years 2010 to 2021.

The remainder of this paper is organized as follows: Section 2 presents the related work. Section 3 describes the security analysis. Section 4 explains the Android applications. Section 5 depicts our proposed work. Section 6 presents the analysis results, and the concluding remarks are given in Section 7.

II. Related Work

There are various approaches to analysing an Android application. Each approach serves a specific purpose and provides insights into different aspects of the app. Some common approaches to analysis include security analysis static analysis, dynamic analysis, and penetration testing. As in the paper [5], they analyse and identify vulnerabilities, and malicious Android apps using vulnerability, static, and dynamic analysis approaches for both Android apps and Android platforms. They introduce an Analysis Framework for Security Analysis of Android applications and Android platforms to provide effective way and sustainable security solutions.

Furthermore, they implemented OWASP Droid Fusion [14], OWASP Mobile Top Ten, National Vulnerabilities Database (NVD) [15], and Open-Source Vulnerabilities Database

(OSVDB) [16] as the guidelines. They also decompiled the target Android app, analysed the source code, and extracted features to compare with the malware signature. In addition, they monitored and analysed the application's behaviours, data traffic, and application logs. If the application was not malicious, they will discover vulnerabilities and apply the available patches before recompiling and signing the app to perform dynamic analysis.

The paper in [6] did a privacy analysis on an Android app, as well as combined the control of the OWASP Mobile Security Project [4], Open Android Security Assessment Methodology (OASAM) [17], and other good practices to introduce a methodology that involves both static and dynamic analysis. They collect information on Android operating systems before and after installation of an application, to determine possible data leakage. They also examine the login process to identify potential vulnerabilities and intercept transactions between the app and server to verify data encryption. Once everything is completed, they perform application source code analysis and component analysis to identify hidden vulnerabilities. The results of the study reveal that the critical vulnerabilities are related to application development, which is the responsibility of developers to be aware of and protect user data.

Another research that does the privacy and security analysis is paper [7]. This work's purpose is to compare the security and privacy outlines based on the differences of the applications. They utilized the OWASP Mobile Top Ten to examine common vulnerabilities and conducted both static and network analysis. They classified threats identified by a tool named DroidSafe into categories, then mapped with OWASP. However, their work found that the Static Analysis failed to examine the functional implication of the threat's prone component, so they decided to perform a manual analysis instead.

III. Security analysis

Security analysis is a method of evaluating and assessing the security of an application to identify any vulnerabilities, weak points, or potential threats. The goal is to recommend measures for mitigating or reducing these risks. It usually uses automated tools to examine and analyse the application, and then generate a report. There are two common analysis process that can be performed: static and dynamic analysis.

Static analysis involves reviewing the application's source code to find potential security vulnerabilities and common issues, such as insecure data storage or improper input validation. This is typically carried out by an automated security scanner like MobSF [11], which decompiles the APK file to analyse the app's bytecode and resources. Checking the Manifest file and then a source code review is conducted to identify code-level vulnerabilities and generate a report.

On the contrary, dynamic analysis is performed while running an Android app. The main objective is to find any weak points in an application and verify security mechanisms that provide sufficient protection against the most prevalent type of attacks, such as authentication and authorization issues, exposing data in transit, etc. Moreover, testing of this analysis may vary depending on the mobile application type. Table 1 shows actions that may be involved in both analyses.

TABLE 1. SAMPLE ACTIONS FOR STATIC AND DYNAMIC ANALYSIS

Static analysis	Dynamic Analysis
• Information Gathering / General Information, • Decompile to java and smali (Reverse Engineering), • Permission Analysis, • Manifest Analysis, and • Application's source code review	• Capture HTTP/HTTPS Traffic, Transport Layer Testing, • Memory Analysis, • File Analysis on application data, • Server-Side Attacks • Activity Tester, • Logcat and Dumpsys • Input/output validation (cross-site scripting, SQL injection, etc.)

IV. Android application

An Android application is application software running on the Android operating system, which is a Linux-based open-source platform developed by Google. To run an app on the Android device, Android SDK tools will compile application source code with its resources into a single file format named Android Package Kit (APK), which is an archive or zip file.

A. Android Application Components

There are four main components to build an Android application: Activities, Services, Broadcast Receivers, and Content providers.

1) Activity: A screen with the user interface (UI) for a user to interact with. In addition, when creating an app, the developer must declare every activity inside the Android Manifest file. Otherwise, the activity is not authorized and allowed to be carried out by the user.

2) Service: A component that allows an application to perform long-running operations in the background, such as notifications. These services must be declared in the AndroidManifest.xml file. There are two types of services: Started services and Bound services. Started services are initiated by calling startService() and can run indefinitely until they are stopped. On the other hand, Bound services are initiated by binding to them via bindService(). They are tightly coupled to the application and can be used to provide a client-server interface between the application and the service.

3) Broadcast Receiver: A component that allows an application to receive notifications from other apps and the system. The broadcast receiver must be declared in the AndroidManifest.xml file with an associated <intent-filter> to specify the action or event that the receiver would like to react to or receive. Without this declaration, the app will not be able to listen or react to the broadcast messages. It's an important component for ensuring smooth communication between different parts of an Android system.

4) Content Provider: It manages a shared set of application data that is stored in various locations such as the file system, SQLite database, or other persistent storage location. This enables other applications to access and modify the data, provided they have been granted permission by the content provider.

B. Permissions and Protection Level

Android provides predefined permission for Android developers to protect user's privacy. If an app needs to access sensitive data or potentially dangerous system features like location or storage, the application must first request permission from the user. The user must then approve the request before the app can proceed. This is an important measure to ensure that users have control over their own data and can make informed decisions about how it is accessed and used.

In addition, Android permissions are categorized into different levels of protection, including normal, dangerous, and signature protection levels.

1) Normal Permission level: This refers to the basic services that Android applications use, including Bluetooth, Internet, and Vibration. When an Android app declares its permissions in a manifest file, they are typically classified as PROTECTION_NORMAL. At install time, the operating system automatically grants these permissions, and the user cannot revoke them.

***2) Dangerous Permission level and Permission group*:** This refers to features that are protected by Android. In case an app requests access to one or all API calls under a permission group, the system will automatically grant access to all APIs under that group. For example, an app declares access to the "coarse location" in the AndroidManifest file, it can access all API calls under the location permission group. In addition, if the app calls any API under the Location permission group again in the future, the system will automatically grant access without asking for permission from the user. So, this is essential to be mindful of the permissions an app requests and understand the risks associated with granting access to potentially sensitive data.

***3) Signature Permission*:** There are certain permissions that are granted at the installation time. These permissions are signed with the same certificate as the app itself, to ensure that they are legitimate and safe to use.

C. Google Primary Security Services for Android

Google also provides some security services to protect users and legitimate apps from malicious applications such as Verify Apps, SafetyNet, SafetyNet Attestation, and Android Device Manager. These services are not part of the Android Open-Source Project, but they are integrated into several Android devices shown in Table 2.

TABLE 2. GOOGLE PRIMARY SECURITY SERVICES

Google Play Security Services	Description
Verify Apps	It warns users or automatically blocks attempted installation of any harmful apps. This also continually scans applications on device.
SafetyNet	It provides a set of services and APIs that help protect an application against security threats, including device tampering, bad URLs, potentially harmful apps, and fake users.
SafetyNet Attestation	Third-party API to determine whether the device is CTS compatible. Attestation can also assist identify the Android app communicating with app server.
Android Device Manager	A web app and Android app to locate lost or stolen devices.

V. PROPOSED WORK

Figure 1 demonstrates an overview of the process of our proposed work. We selected Android apps widely used in Thailand for hotel and airline booking and conducted a security analysis by using an automated scanner named MobSF [11] to analyze and extract security-related data from the target application. In addition, we manually mapped results with MSTG [4] for some parts. Then we categorized the features into six groups and clustered to see some patterns. Finally, we manually calculated the centroid table and found the best K-value using the Elbow method. Table 3 shows some steps we performed during the static analysis.

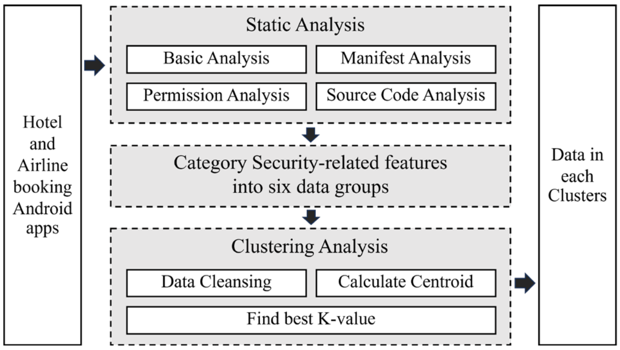


Figure 1. System overview diagram

A. Static Analysis

Four subprocesses in the static analysis are Basic analysis, Manifest analysis, Permission analysis and Source Code analysis. The results are obtained from MobSF analysis report.

TABLE 3. DETAILS OF TOOLS AND STEPS TO PERFORM STATIC ANALYSIS

Basic Analysis		
Tool	**How to perform**	**Output Information**
MobSF	collect information from MobSF analysis report under Information section.	**General Information** - Application Name - Application Version - Target SDK version - Min SDK version support - Category - Package Name

APKTool v.2.4.1	Decompile an APK file, then search for platform folder	**Security Features** - Root Detection Capability
MobSF	Decompile an APK file, then search for platform folder	**Security Features** -Signature Verification Algorithm - Signature Algorithm Status
MobSF	Check root detection capability by searching word "Root Detection" inside report.	- SafetyNet API Implemented
Manifest Analysis		
Tool	**How to perform**	**Output Information**
MobSF	Collect information. on MobSF analysis report under info. and Manifest analysis section.	**Android Components that are not protected.** - Exported Activity - Export Services - Exported Broadcast Receiver - Exported Content Provide
MobSF	check if the application allows the attacker to perform data Backup or not.	- Check if "android:allowBackup=true"
Permission Analysis		
Tool	**How to perform**	**Output Information**
MobSF	Collect info under permission analysis section	- List of Normal Permissions - List of Protected Permission
Source Code Analysis		
Tool	**How to perform**	**Output Information**
MobSF	We collect the information from the MobSF analysis report under the Code Analysis section.	- Type of code issues - Severity - Threats
Manual	Search the vulnerable code in the java files.	- Vulnerable code
Manual	Map result with OWASP Mobile Top Ten and the MSTG Checklist	- Summary of code issues map to MSTG and OWASP top 10 2016. - CVSS scores

B. Defined Parameters

This section will talk about six data groups that we classified.

1) ***Data Group 1:*** it's about Android security features used by an application. Regarding the MSTG Android Anti-Reversing Defenses [8], we will focus on root detection, SafetyNet implemented, and file integrity checks.

2) ***Data Group 2:*** This group focuses on "exported" Android components (activity, service, or content) that can be accessed by other apps.

3) ***Data Group 3:*** This group focuses on app permissions that are considered dangerous. To evaluate whether the permissions requested are related to the app's purpose or not.

4) ***Data Group 4:*** This group focus on analysing Android Manifest file and check for the ADB full back up support, copies data to the clipboard or not, and check whether app requests root privileges.

5) ***Data Group 5:*** it's about others source code issues with the CVSSv2 (Common Vulnerability Scoring System).

6) ***Data Group 6:*** This group is an others source code issues related to WebView and SQL. This group focuses on other source code issues that did not have CVSS scores.

C. Clustering of output parameters

In this work, we use the K-mean algorithm with the centroid models. We select Euclidean Distance to calculate the distance between two points and find the best K value using Elbow Curve [√ Σ (Ai-Bi)2] methods. After getting the results of the centroid table from the RapidMiner tool, we calculated using k-value from 2 to 6 clusters. Then we use this value to calculate the Within-Cluster Sum of Square or WCSS, which is the sum of the squared distance between each point and the centroid in the cluster.

VI. ANALYSIS RESULTS

This section will discuss the results of our analysis. Table 4 shows the result we extracted during static analysis phase before clustering. After doing the data preparation, import to the RapidMiner tool, and calculate the WCSS scores using Elbow curve method. We found that the best k-value is 4 clusters.

A. Output Analysis for Each Data Group

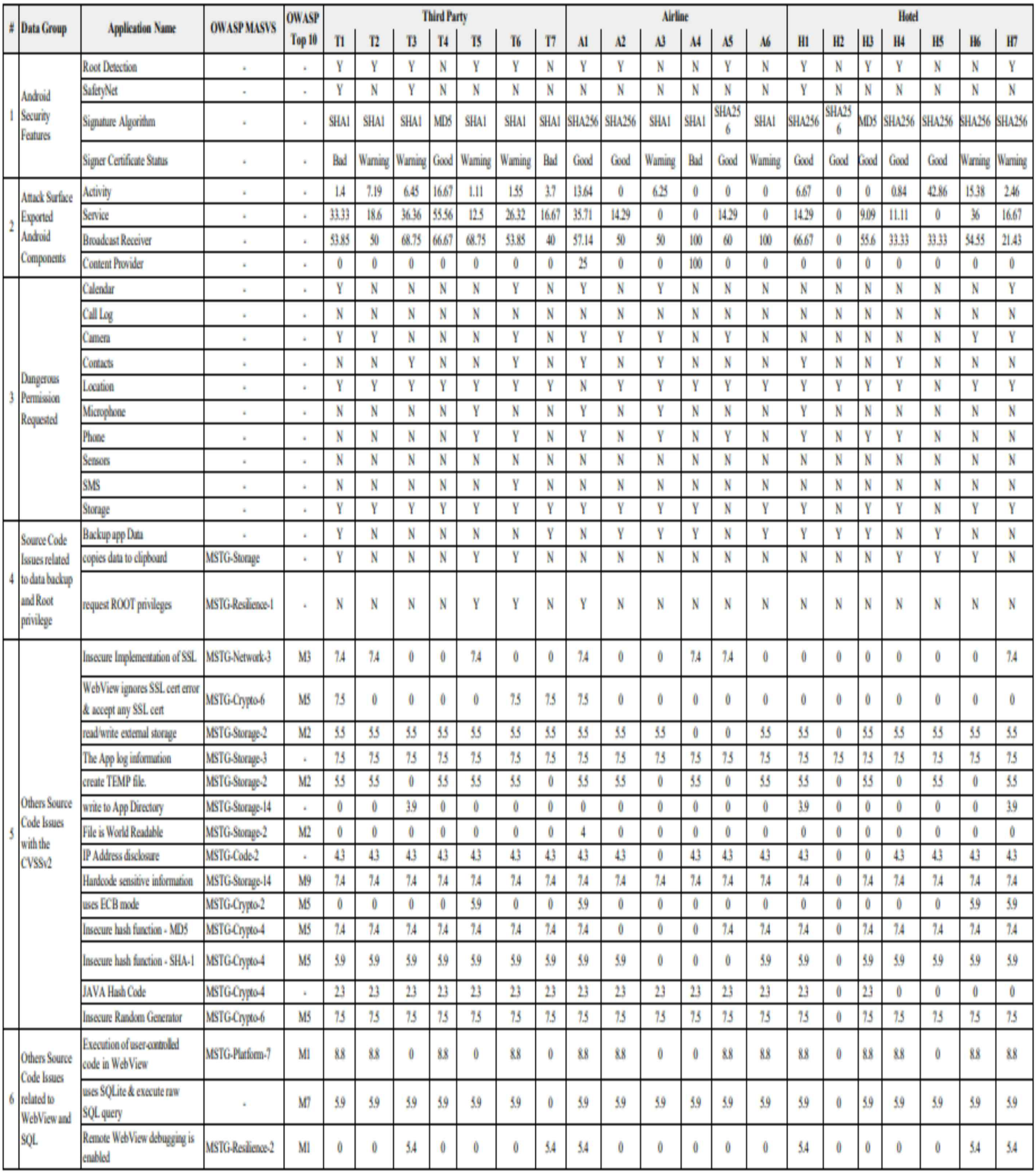
To see some pattern or relationship in which the application may be similar or completely different, we are assigning a colour to each cluster and then summarize them based on the parameters or attributes, which can group just some of them.

1) ***Data Group 1 (Android Security Features):*** Third-party apps who's capable to book both hotels and airlines. Likely to implement similar Android security features as shown in cluster 0, while applications developed by airlines or hotels are classified in cluster 1. And once we are looking more closely at the parameters or attributes, seems like those applications in Cluster 1 are implementing Android security features more than apps in Cluster 3, either. in terms of algorithms for signing applications and detecting jailbreak devices. Moreover, the application in cluster 3 implements a low number of Android security features also the application was signed with "SHA1 with RSA" whose SHA1 hash algorithm is known to have collision issues.

2) ***Data Group 2 (Attack Surface Exported Android Components):*** In this group, the distribution of applications from different types which may result from platforms or tools for development. Causing the results to come out so different that no conclusions can be drawn. In addition, the results show that applications in cluster 2 are more likely to share Android components (Activities, Services, Content Providers, and Broadcast Receivers) with other apps in the same device.

3) ***Data Group 3 (Dangerous Permission Requested):*** In this group, there is no application that can be identified to a particular cluster. In addition, from the results, the app parameter shows that most applications request access to storage and location. We also see that apps in cluster 1 request more dangerous permissions than applications in other clusters,

TABLE 4. STATIC ANALYSIS RESULTS FOR 6 DATA GROUPS



#	Data Group	Application Name	OWASP MASVS	OWASP Top 10	Third Party							Airline						Hotel						
					T1	T2	T3	T4	T5	T6	T7	A1	A2	A3	A4	A5	A6	H1	H2	H3	H4	H5	H6	H7
1	Android Security Features	Root Detection	-	-	Y	Y	Y	N	Y	Y	N	Y	Y	N	N	Y	N	Y	N	Y	Y	N	N	Y
		SafetyNet	-	-	Y	N	Y	N	N	N	N	N	N	N	N	N	N	Y	N	N	N	N	N	N
		Signature Algorithm	-	-	SHA1	SHA1	SHA1	MD5	SHA1	SHA1	SHA1	SHA256	SHA256	SHA1	SHA1	SHA25 6	SHA1	SHA256	SHA25 6	MD5	SHA256	SHA256	SHA256	SHA256
		Signer Certificate Status	-	-	Bad	Warning	Warning	Good	Warning	Warning	Bad	Good	Good	Warning	Bad	Good	Warning	Good	Good	Good	Good	Good	Warning	Warning
2	Attack Surface Exported Android Components	Activity	-	-	1.4	7.19	6.45	16.67	1.11	1.55	3.7	13.64	0	6.25	0	0	0	6.67	0	0	0.84	42.86	15.38	2.46
		Service	-	-	33.33	18.6	36.36	55.56	12.5	26.32	16.67	35.71	14.29	0	0	14.29	0	14.29	0	9.09	11.11	0	36	16.67
		Broadcast Receiver	-	-	53.85	50	68.75	66.67	68.75	53.85	40	57.14	50	50	100	60	100	66.67	0	55.6	33.33	33.33	54.55	21.43
		Content Provider	-	-	0	0	0	0	0	0	0	25	0	0	100	0	0	0	0	0	0	0	0	0
3	Dangerous Permission Requested	Calendar	-	-	Y	N	N	N	N	Y	N	Y	N	Y	N	N	N	N	N	N	N	N	N	Y
		Call Log	-	-	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
		Camera	-	-	Y	Y	N	N	N	Y	N	Y	Y	Y	N	Y	N	N	N	N	N	N	Y	Y
		Contacts	-	-	N	N	Y	N	N	Y	N	Y	N	Y	N	N	N	Y	N	N	Y	N	N	N
		Location	-	-	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y
		Microphone	-	-	N	N	N	N	Y	N	N	Y	N	Y	N	N	N	Y	N	N	N	N	N	N
		Phone	-	-	N	N	N	N	Y	Y	N	Y	N	Y	N	Y	N	Y	N	Y	Y	N	N	N
		Sensors	-	-	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
		SMS	-	-	N	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N
		Storage	-	-	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	N	Y	Y	N	Y	Y
4	Source Code Issues related to data backup and Root privilege	Backup app Data	-	-	Y	N	N	N	N	N	Y	N	Y	Y	Y	N	Y	Y	Y	Y	N	Y	N	N
		copies data to clipboard	MSTG-Storage	-	Y	N	N	N	Y	Y	N	N	N	N	N	N	N	N	N	N	Y	Y	Y	N
		request ROOT privileges	MSTG-Resilience-1	-	N	N	N	N	Y	Y	N	Y	N	N	N	N	N	N	N	N	N	N	N	N
5	Others Source Code Issues with the CVSSv2	Insecure Implementation of SSL	MSTG-Network-3	M3	7.4	7.4	0	0	7.4	0	0	7.4	0	0	7.4	7.4	0	0	0	0	0	0	0	7.4
		WebView ignores SSL cert error & accept any SSL cert	MSTG-Crypto-6	M5	7.5	0	0	0	0	7.5	7.5	7.5	0	0	0	0	0	0	0	0	0	0	0	0
		read/write external storage	MSTG-Storage-2	M2	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	0	0	5.5	5.5	0	5.5	5.5	5.5	5.5	5.5
		The App log information	MSTG-Storage-3	-	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5
		create TEMP file.	MSTG-Storage-2	M2	5.5	5.5	0	5.5	5.5	5.5	0	5.5	5.5	0	5.5	0	5.5	5.5	0	5.5	0	5.5	0	5.5
		write to App Directory	MSTG-Storage-14	-	0	0	3.9	0	0	0	0	0	0	0	0	0	0	3.9	0	0	0	0	0	3.9
		File is World Readable	MSTG-Storage-2	M2	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
		IP Address disclosure	MSTG-Code-2	-	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3	0	4.3	4.3	4.3	4.3	0	0	4.3	4.3	4.3	4.3
		Hardcode sensitive information	MSTG-Storage-14	M9	7.4	7.4	7.4	7.4	7.4	7.4	7.4	7.4	7.4	7.4	7.4	7.4	7.4	7.4	0	7.4	7.4	7.4	7.4	7.4
		uses ECB mode	MSTG-Crypto-2	M5	0	0	0	0	5.9	0	0	5.9	0	0	0	0	0	0	0	0	0	0	5.9	5.9
		Insecure hash function - MD5	MSTG-Crypto-4	M5	7.4	7.4	7.4	7.4	7.4	7.4	7.4	7.4	0	0	0	7.4	7.4	7.4	0	7.4	7.4	7.4	7.4	7.4
		Insecure hash function - SHA-1	MSTG-Crypto-4	M5	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	0	0	0	5.9	5.9	0	5.9	5.9	5.9	5.9	5.9
		JAVA Hash Code	MSTG-Crypto-4	-	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	0	2.3	0	0	0	0
		Insecure Random Generator	MSTG-Crypto-6	M5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	0	7.5	7.5	7.5	7.5	7.5
6	Others Source Code Issues related to WebView and SQL	Execution of user-controlled code in WebView	MSTG-Platform-7	M1	8.8	8.8	0	8.8	0	8.8	0	8.8	8.8	0	0	8.8	8.8	8.8	0	8.8	8.8	0	8.8	8.8
		uses SQLite & execute raw SQL query	-	M7	5.9	5.9	5.9	5.9	5.9	5.9	0	5.9	5.9	5.9	5.9	5.9	5.9	5.9	0	5.9	5.9	5.9	5.9	5.9
		Remote WebView debugging is enabled	MSTG-Resilience-2	M1	0	0	5.4	0	0	0	5.4	5.4	0	0	0	0	0	5.4	0	0	0	0	5.4	5.4

requesting access to contact, camera, microphone, phone, or even SMS.

4) ***Data Group 4 (Source Code Issues related to data backup and Root privilege):*** In this group, the results shows that those apps in cluster 3 do not have the ability to backup app data / copy data and request higher permissions. It is considered good in terms of security. In addition, applications in Cluster1 can backup app data, copy data to the clipboard, and request root privileges. Mean these apps may have higher risks than others.

5) ***Data Group 5 (Others Source Code Issues with the CVSSv2):*** In this group, we can conclude into 3 groups. The first group is Cluster 0 and Cluster 1 which are different from SSL and ECB modes. Second is cluster 2, which is different from all other groups. The third group is Cluster 3 who's like Cluster 0 except hash algorithm, SHA-1 / md5, and Java hash.

6) ***Data Group 6 (Others Source Code Issues related to WebView and SQL):*** In this group, the applications in Clusters 0 and 1 are different in the "remote web view" attribute. From the security point of view, the apps in such clusters are more vulnerable than the apps in other clusters.

VII. Conclusions

In this paper, we perform the security analysis of 20 android applications for hotel and airline reservation. We divided the security features of such applications into 6 groups, and we do the clustering on them. Eventually, we found the results listed below.

- For data group 1, applications in cluster 1 implemented android security features than apps and have a good status of signer certificate except H7. While apps in cluster 3 implement a low number of Android security features also the app was signed with "SHA1 with RSA".
- For data group 2, applications in cluster 2 are more likely to share Android components with other apps in the same device. Moreover, apps in cluster 1, are more likely to allow other applications to query or modify the data.
- For data group 3, apps in cluster 1 request more dangerous permissions than applications in other clusters, requesting access to contact, camera, microphone, phone, or even SMS. While applications in clusters 0 request permission only they needed.
- For data group 4, only applications in clusters 1 requested root privileges. While applications in cluster 2 and some of applications in cluster 0 allow to backup applications data.
- For data group 5, applications in cluster 2 have a low number of source code issues such as apps log information, create TEMP file, etc.
- Last, data group 6, applications in cluster 1 have the code issues related to remote WebView, execute raw SQL query and execution of user-controlled code in WebView. While applications in other clusters have some of them.

In addition to the data analysis presented in this work, we encountered some limitations. Firstly, the number of applications used was small and resulted in a limited range of clustering such that only a few patterns could be identified. Secondly, the versions of the applications used were not the latest version; thus, some security features may have been updated, and the analysis may not fully reflect the current status. Thirdly, the number of clusters was determined based on the Elbow Curve graph, and we chose four clusters for all data groups, but this may not be suitable for all groups. To achieve better meaningful results, future research may consider analysing the same type of Android applications or the same group, such as apps exclusively for hotel bookings or airlines, so that we can facilitate detailed comparisons. In addition, expanding the dataset size and using the latest version of the applications would enhance the ability to identify and analyse patterns effectively. Finally, each data group should have a different number of clusters in the analysis.

Acknowledgment

This research work was partially supported by the Faculty of Information and Communications Technology, Mahidol University.

References

[1] "Top 50 Products by Total Number of Distinct Vulnerabilities in 2017". [Online]. Available: https://www.cvedetails.com/.

[2] Pradeo, Mobile Applications Threats Review for S1 2017 pages 4-10. 2017.

[3] Google, Android Security 2017 Year in Review Final Report pages 15, 2017.

[4] OWASP Mobile Application Security. [Online]. Available: https://owasp.org/www-project-mobile-app-security/.

[5] Umasankar, Analysis of Latest Vulnerabilities in Android, 2017.

[6] Alejandro Argudo, Gabriel López, Franklin Sánchez, Privacy Vulnerability Analysis for Android Applications, 2017.

[7] Ashish Rajendra Sai, Jim Buckley, Andrew Le Gear, Privacy and Security analysis of cryptocurrency mobile applications, 2019.

[8] "Permissions on Android". [Online]. Available: https://developer.android.com/guide/topics/permissions/

[9] Sen Chen, Yuxin Zhang, Lingling Fan, Jiaming Li, Yang Liu, AUSERA: Automated Security Vulnerability Detection for Android Apps, 2022.

[10] Pasquale Stirparo, Ioannis Kounelis, The mobileak project: Forensics methodology for mobile application privacy assessment, 2013

[11] "Mobile Security Framework (MobSF)". [Online]. Available: https://github.com/MobSF/Mobile-Security-Framework-MobSF.

[12] "Android Testing Cheat Sheet". [Online]. Available: https://owasp.org/index.php/Android_Testing_Cheat_Sheet

[13] RapidMiner. [Online]. Available: https://rapidminer.com/

[14] OWASP Droid Fusion. [Online]. Available: https://wiki.owasp.org/index.php/OWASP_Droid_Fusion

[15] OWASP Mobile Top Ten. [Online]. Available: https://owasp.org/www-project-mobile-top-10/

[16] Open-Source Vulnerabilities Database (OSVDB). [Online]. Available: https://osv.dev/

[17] Open Android Security Assessment Methodology (OASAM). [Online]. Available: https://github.com/b66l/OASAM

The development of a new system for generating training data of AI-based anomaly detection

Thi My Truong *, Won Seok Choi *, Jeong Jang Hyeon **, Seong Gon Choi *

* College of Information and Communication Engineering, Chungbuk National University, Chungdae-ro 1, Seowon-gu

** JJ SOLUTION INC, Chungbuk National University, Cheongju-si, Chungcheongbuk-do, South Korea

mytruong@cbnu.ac.kr, wschoi@cbnu.ac.kr, jjsol210120@gmail.com, sgchoi@cbnu.ac.kr

Abstract— **This paper proposes a method and system for generating training data to support AI based anomaly detection. The use of AI in abnormal behavior detection systems is becoming increasingly popular, with active research on AI-based anomaly detection methods using machine learning. In general, existing research relies on open datasets provided by various laboratories like Swat, WaDI, SMAP and MSL for testing and validation purposes. Since the types of normal and malicious packets depend on the specific network to which they are applied, verifying AI-based anomaly detection methods using an open dataset may yield different results than when applied in real-world scenarios. In other words, open datasets captured from specific networks may not be suitable for applying AI-based abnormal detection methods to other networks. In addition, AI-based datasets may be insufficient for learning, leading to the use of simulated attacks. Open datasets are difficult to provide sufficient data for training and often contain malicious packets using simulated attack packets. Since malicious attacks are always transformed into new forms and developed in types, it is necessary to prepare a database for new malicious attacks and to learn about them. Therefore, one of the major challenges in developing effective anomaly detection systems is acquiring an appropriate dataset. To address this issue, we propose a system for extracting training data by collecting packets from the actual network to apply AI-based abnormal detection. Our proposed system offers the advantage of accurately reflecting the network's packet characteristics by gathering data from live networks for AI-based abnormality detection and dataset creation. Furthermore, as it incorporates a dataset for the latest malicious attacks within the network, it enables more practical anomaly detection compared to the use of existing datasets. We simulated and tested the proposed system at the laboratory level to confirm its behavior.**

***Keywords*— anomaly detection, artificial intelligence (AI), dataset, training data, cybersecurity**

I. Introduction

Anomaly detection, a fundamental component of network security, plays a pivotal role in identifying abnormal patterns and potential threats within a given environment. In this context, anomaly detection refers to the process of discerning deviations from expected or typical behavior in a system or dataset [1]. These deviations, often referred to as anomalies, can indicate a wide range of irregularities, including network intrusions, cybersecurity threats, equipment malfunctions, or fraudulent activities.

The integration of artificial intelligence (AI) and machine learning techniques has revolutionized the field of anomaly detection, allowing for more highly advanced and adaptable solutions. AI-based anomaly detection methods leverage the power of algorithms and statistical models to learn and recognize patterns in data, enabling them to automatically detect anomalies that might be difficult to be identified through traditional approaches.

While AI-based anomaly detection methods have made significant progress, they are only as effective as the data they are trained on. Existing research in this domain has often relied on open datasets offered by various laboratories and research institutions, such as Swat, WaDI, SMAP and MSL. These datasets have been used for testing and validating anomaly detection algorithms. However, a critical limitation arises when applying AI-based anomaly detection methods in the real-world network environments. The challenge lies in the diversity among networks, each with its distinct attributes and traffic patterns. Therefore, a method that performs well on one network may yield different results when applied to another, making the transition from research to practical implementation complex.

Moreover, to train AI models efficiently, the datasets must be diverse, comprising of both normal and malicious packets. However, the aforementioned open datasets do not produce a sufficiently diverse and comprehensive set of data for training. The landscape of malicious attacks is constantly evolving, giving rise to new forms and types of threats. To build effective anomaly detection systems, it is essential to have access to data that accurately reflects the latest network conditions and includes the most recent malicious attacks.

This paper addresses these challenges and proposes a novel method and system for generating training data specifically to support AI-based anomaly detection. Our approach is grounded in the collection of real-world network traffic data, offering a distinct advantage in accurately reflecting the unique characteristics of the network under consideration. Furthermore, our system is designed to incorporate data related to the latest malicious attacks within the network, ensuring that AI-based anomaly detection methods are well-equipped to handle the dynamic nature of cybersecurity threats.

In the subsequent sections of this paper, we will present in detail our method for extracting training data from live networks, the mechanisms for dataset creation, and the results of simulating and testing our proposed system at the laboratory level.

II. RELATED WORK

A. Existing Anomaly detection research

Machine learning based approach in anomaly detection is favored by researchers for flexibility and applicability to any network structure in comparison to the traditional detection [2]. Machine learning has been applied to detect abnormal behavior in various types of networks, using a variety of model types, including supervised learning, unsupervised learning, reinforcement learning, and deep learning models.

Chen et al. [3] analyzed network traffic and employed machine learning methods to identify abnormal behavior and detect malicious apps. They used imbalanced classification methods, including the Synthetic Minority Oversampling Technique (SMOTE) combined with Support Vector Machine (SVM), SVM Cost-Sensitive (SVMCS), and C4.5 Cost-Sensitive (C4.5CS) methods. While this approach performed well with highly imbalanced training data, its performance became unstable when the dataset's imbalance ratio was under 1000.

Hamamoto et al. [4] utilized Genetic Algorithms to analyze the network and subsequently employed a Fuzzy Logic scheme to determine whether an instance represents an anomaly. This method exhibits high performance in Denial of Service (DoS) and Distributed Denial of Service (DDoS) attack detection, but it is associated with a high false-negative rate.

Alauthman et al. [5] employ the output of a supervised learning model as the state in the reinforcement learning model, both the SL and RL model are improving through this interaction. This method has a good accuracy rate when the input data is reduced in the model, leading to reduced training time. However, it has been validated in MATLAB using three datasets and has not been implemented in a real network.

Wei et al. [6] employ Convolutional Neural Networks (CNN) to learn spatial features in the data and use Recurrent Neural Networks (RNN) with long-short term memory to learn temporal features. Subsequently, the original datasets DARPA1998 and ISCX2012 undergo preprocessing. The advantage is the improved performance achieved, this approach has only been validated using a fixed dataset, which can be considered a limitation.

All these studies have encountered data limitations, such as unbalanced data, insufficient data, and a lack of validation in real network environments.

B. Available open datasets

SWAT (Secure Water Treatment) [7] simulates the operations of a real-world industrial water treatment plant. SWaT was run and data were collected over an 11-day period. The first 7 days were dedicated to normal data collection without any attacks or errors, while the remaining 4 days involved 36 attacks created by the research team. This dataset includes physical properties relevant to the plant and the water treatment process, as well as network traffic within the testbed. It comprises 51 channels, encompassing sensors such as flow meters, level transmitters, conductivity analyzers, and actuators like motorized valves and pumps.

The WADI dataset [8] was collected from the WADI testbed, which is an extension of the SWAT testbed. This dataset comprises data from 1233 sensors and actuators and was collected over a 16-day period. Of these 16 days, 14 days were dedicated to normal data collection, while the remaining 2 days involved 15 attacks.

The SMAP and MSL datasets [9] consist of expert-labeled telemetry anomaly data from NASA's Soil Moisture Active Passive (SMAP) satellite and Mars Science Laboratory (MSL) rover. The number of SMAP variables is 1375, while the that of MSL features is 1485, making them significantly more extensive compared to single-entity datasets. To effectively handle this increased complexity, the data has been systematically organized into 55 distinct entities for SMAP and 27 entities for MSL. This categorization facilitates a structured methodology for the analysis and detection of anomalies.

These open datasets only reflect the characteristics of the specific networks to which they are applied. While SWAT and WADI datasets pertain to networks in a simulated water plant, SMAP and MSL datasets deal with telemetry data.

III. PROPOSED METHOD

The core of our system revolves around a three-step process described in Figure 1. First, we collected log information from security devices in public institutions. Then, we analyzed the collected IP addresses, comparing them to Threat Intelligence (TI) information to check for potential threats. When a threat is detected, the system is automatically applied to the relevant policy to agencies with API connection, while the rest without API connection receive email notifications about the threat.

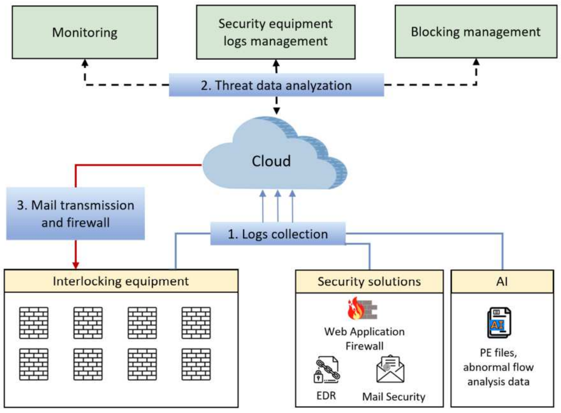


Figure 1. Overall System Architecture

The data collecting process is described in Figure 2. In each target network, packets are transmitted from the switch to the data collecting servers using packet control techniques such as mirroring or inline. The output of the data collecting servers consists of log files or extracted files with unique network characteristics. These files are subsequently transmitted to a virtual machine via the Internet. Before being sent to the virtual machine, a firewall is configured to receive only the traffic directed to a specific port number and the IP address of the collection sensor associated with the relevant agency. Finally, the extracted files have been successfully transmitted by sending an acknowledgment message.

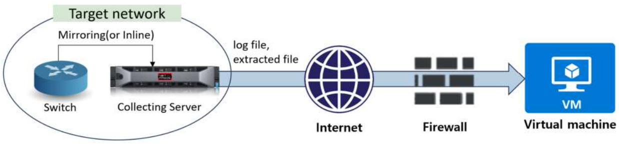


Figure 2. Data collecting process

The core of our system is the data collecting process that takes place at the collecting server as Figure 3. Packets are transmitted from the Packet Collecting module to the Metadata extracting module, where the network traffic is analyzed. This process extracts application protocols and information of flow, including IP addresses, port numbers, packet counters, and additional details corresponding to the protocol type. The output of Metadata extracting module, in the form of a string (raw data), is transmitted through a pipeline and stored in a temporary file in minutes. During the data processing, this temporary file is read, and the timestamp is converted to UNIX time format. After removing any abnormal lines, such as those lacking proper newline delimiters, the information is saved to a final file, and the temporary file is deleted. Statistical information regarding the extracted protocols is updated on a minute-by-minute basis.

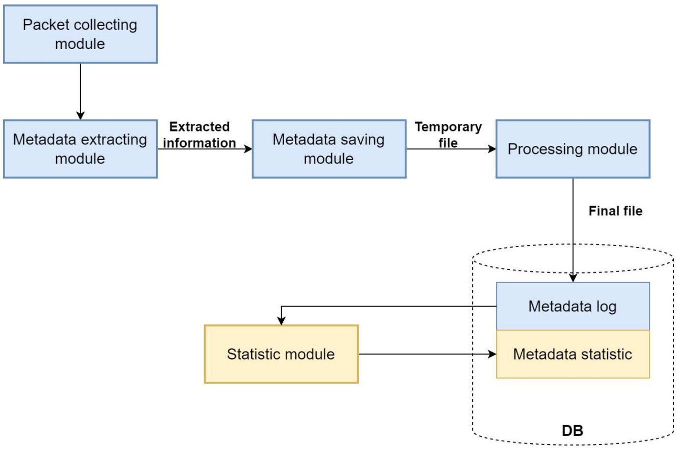


Figure 3. Collecting server

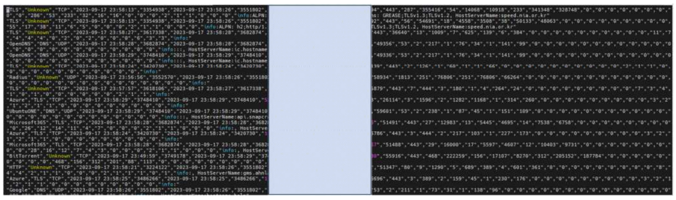
Each metadata record is comprised of 44 features, described in Table 1. Features 1-3 explain the primary protocol, sub-protocol, and Layer 4 protocol. Features 4-7 pertain to flow start and end times in seconds and sub-seconds. Information about the IP addresses and port numbers of both the source and destination is provided in features 8-11. Features 12-19 provide information regarding the number of packets, packet size, and actual valid data throughput in both directions (source to destination and vice versa, destination to source). Features 20-41 represent the count of packets with specific flags, such as TCP Congestion Window Reduced, ECN-Echo, Urgent, Acknowledge, Push, Reset, Synchronized Sequence Numbers, and the Finish flag, enabled within the flow, in each direction from source to destination and from destination to source. Additional flow-related information is provided in the final feature.

TABLE 1. METADATA OF THE DATASET

Number	Field name	Number	Field name
1	proto_app	23	sum_ece_cnt
2	proto_master	24	s2d_ece_cnt
3	proto_l4	25	d2s_ece_cnt
4	epoch_t_first	26	sum_urg_cnt
5	micro_s_first	27	s2d_urg_cnt
6	epoch_t_last	28	d2s_urg_cnt
7	micro_s_last	29	sum_ack_cnt
8	src_ip	30	s2d_ack_cnt
9	dst_ip	31	d2s_ack_cnt
10	src_port	32	sum_psh_cnt
11	dst_port	33	s2d_psh_cnt
12	sum_packets	34	d2s_psh_cnt
13	sum_bytes	35	sum_rst_cnt
14	s2d_packets	36	s2d_rst_cnt
15	s2d_bytes	37	d2s_rst_cnt
16	s2d_goodput	38	sum_syn_cnt
17	d2s_packets	39	s2d_syn_cnt
18	d2s_bytes	40	d2s_syn_cnt
19	d2s_goodput	41	sum_fin_cnt
20	sum_cwr_cnt	42	s2d_fin_cnt
21	s2d_cwr_cnt	43	d2s_fin_cnt
22	d2s_cwr_cnt	44	info

IV. Experiment Result

Real data collection is illustrated in Figure 4. The system detects a variety of application protocols, including TLS, OpenDNS, Radius, Azure, Microsoft 365, BitTorrent, HTTP, Google, DNS, etc. IP address and port number information has been masked for security purposes. Actual traffic information is retained for the training of AI-based anomaly detection models.



```
"TLS","Unknown","TCP","2023-09-17 23:58:13","3354938","2023-09-17 23:58:26","3551802", [redacted] 94","443","287","355416","54","14068","10918","233","341348","328748","0","0","0","0","0","0","0","
0","0","286","53","233","32","16","16","0","0","0","2","1","1","0","0","0","info:ALPN: [redacted] s: GREASE;TLSv1.3;TLSv1.2, HostServerName:speed.nia.or.kr"
"TLS","Unknown","TCP","2023-09-17 23:58:13","3354938","2023-09-17 23:58:26","3551802", [redacted] 92","443","56","54691","18","4558","3508","38","50133","48063","0","0","0","0","0","0","0","0","0",
"55","17","38","11","6","5","0","0","0","2","1","1","0","0","0","info:ALPN: h2;http/1. [redacted] TLSv1.3;TLSv1.2, HostServerName:speed.nia.or.kr"
"TLS","Unknown","TCP","2023-09-17 23:58:27","3617338","2023-09-17 23:58:28","3682874", [redacted] "443","36640","13","1009","7","625","139","6","384","0","0","0","0","0","0","0","0","0","0","11","7
","4","4","4","0","2","0","2","0","0","0","6","3","3","info:"
"OpenDNS","DNS","UDP","2023-09-17 23:58:28","3682874","2023-09-17 23:58:28","3682874", [redacted] ,"49356","53","2","217","1","76","34","1","141","99","0","0","0","0","0","0","0","0","0","0","0","0
","0","0","0","0","0","0","0","0","0","0","0","0","info:::, HostServerName:ic.hostname
"OpenDNS","DNS","UDP","2023-09-17 23:58:29","3748410","2023-09-17 23:58:29","3748410", [redacted] ,"49336","53","2","217","1","76","34","1","141","99","0","0","0","0","0","0","0","0","0","0","0","0
","0","0","0","0","0","0","0","0","0","0","0","0","info:::, HostServerName:ic.hostname
"TLS","Unknown","TCP","2023-09-17 23:58:24","3420730","2023-09-17 23:58:24","3420730", [redacted] 139","443","2","126","1","60","1","1","66","0","0","0","0","0","0","0","0","0","0","2","1","1","0",
"0","0","0","0","0","0","0","0","0","0","0","info:"
"Radius","Unknown","UDP","2023-09-17 23:56:56","3552570","2023-09-17 23:58:26","355180 [redacted] "58934","1813","251","76806","251","76806","66264","0","0","0","0","0","0","0","0","0","0","0","0",
"0","0","0","0","0","0","0","0","0","0","0","0","0","0","0","info:"
"TLS","Unknown","TCP","2023-09-17 23:57:57","3618106","2023-09-17 23:58:27","3617338", [redacted] 6879","443","7","444","3","180","1","4","264","24","0","0","0","0","0","0","0","0","0","7","3","4",
"1","0","1","0","0","0","0","0","0","2","1","1","info:"
"Azure","TLS","TCP","2023-09-17 23:58:29","3748410","2023-09-17 23:58:29","3748410","5 [redacted] 3","26114","3","1596","2","1282","1168","1","314","260","0","0","0","0","0","0","0","0","0","3","2"
,"1","2","1","1","0","0","0","0","0","0","0","0","0","info:"
"UbuntuONE","DNS","UDP","2023-09-17 23:58:29","3748410","2023-09-17 23:58:29","3748410 [redacted] ,"59601","53","2","238","1","87","45","1","151","109","0","0","0","0","0","0","0","0","0","0","0","
0","0","0","0","0","0","0","0","0","0","0","0","0","info:::, HostServerName:api.snapcr
"Microsoft365","TLS","TCP","2023-09-17 23:58:28","3682874","2023-09-17 23:58:28","3682 [redacted] 5","51491","443","27","12983","13","5445","4695","14","7538","6758","0","0","0","0","0","0","0","0"
,"0","26","12","14","11","4","7","0","0","0","2","1","1","0","0","0","info:, HostServe
"Azure","TLS","TCP","2023-09-17 23:58:24","3420730","2023-09-17 23:58:24","3420730","1 [redacted] 5786","443","3","444","2","217","103","1","227","173","0","0","0","0","0","0","0","0","0","3","2","
1","2","1","1","0","0","0","0","0","0","0","0","0","info:"
"Microsoft365","TLS","TCP","2023-09-17 23:58:28","3682874","2023-09-17 23:58:28","3682 [redacted] 17","51488","443","29","16000","17","5597","4607","12","10403","9731","0","0","0","0","0","0","0","
0","0","28","16","12","7","4","3","0","0","0","2","1","1","0","0","0","info:, HostServ
"BitTorrent","Unknown","TCP","2023-09-17 23:49:59","3749178","2023-09-17 23:58:29","37 [redacted] 80","55916","443","468","222259","156","17107","8270","312","205152","187784","0","0","0","0","0","
0","0","0","0","468","156","312","201","88","113","0","0","0","0","0","0","0","0","0","0",
"HTTP","Unknown","TCP","2023-09-17 23:58:21","3224122","2023-09-17 23:58:26","3551802" [redacted] ,"51347","80","9","1290","5","689","389","4","601","361","0","0","0","0","0","0","0","0","0","8","
4","4","2","1","1","0","0","0","2","1","1","1","0","1","info:, HostServerName:gms.ahnl
"Azure","TLS","TCP","2023-09-17 23:58:25","3486266","2023-09-17 23:58:25","3486266","1 [redacted] 5696","443","3","389","2","159","45","1","230","176","0","0","0","0","0","0","0","0","0","3","2","1
","2","1","1","0","0","0","0","0","0","0","0","0","info:"
"Google","DNS","UDP","2023-09-17 23:58:26","3551802","2023-09-17 23:58:26","3551802"," [redacted] "53","2","211","1","73","31","1","138","96","0","0","0","0","0","0","0","0","0","0","0","0","0","0"
```

Figure 4. Real Data Collection (Private IP Addresses)

V. CONCLUSIONS

While AI-based anomaly detection is gaining widespread attention and application, the reliance on open datasets for research and testing has certain limitations. Existing open datasets, collected from specific networks, may not be directly applicable to other network environments due to variations in normal and malicious packet behaviors. As a solution, we have proposed a system that collects packets directly from live networks, produced a more accurate representation of the network's unique characteristics. This approach to data collection not only enhances the performance of AI-based anomaly detection but also contributes to the ongoing development of more adaptable systems.

ACKNOWLEDGMENT

This work was partly supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2024-2020-0-01462, 33%) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). This work was partly supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2020R1A6A1A12047945, 34%). This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C1006167, 33%).

*corresponding author is S.G. Choi(choisg@cbnu.ac.kr).

REFERENCES

[1] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.

[2] Wang, S., Balarezo, J. F., Kandeepan, S., Al-Hourani, A., Chavez, K. G., & Rubinstein, B. (2021). Machine learning in network anomaly detection: A survey. IEEE Access, 9, 152379-152396.

[3] Chen, Z., Yan, Q., Han, H., Wang, S., Peng, L., Wang, L., & Yang, B. (2018). Machine learning based mobile malware detection using highly imbalanced network traffic. Information Sciences, 433, 346-364.

[4] Hamamoto, A. H., Carvalho, L. F., Sampaio, L. D. H., Abrão, T., & Proença Jr, M. L. (2018). Network anomaly detection system using genetic algorithm and fuzzy logic. Expert Systems with Applications, 92, 390-402.

[5] Alauthman, M., Aslam, N., Al-Kasassbeh, M., Khan, S., Al-Qerem, A., & Choo, K. K. R. (2020). An efficient reinforcement learning-based Botnet detection approach. Journal of Network and Computer Applications, 150, 102479.

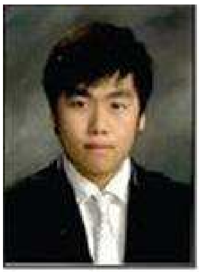
[6] Wei, G., & Wang, Z. (2021). Adoption and realization of deep learning in network traffic anomaly detection device design. Soft Computing, 25(2), 1147-1158.

[7] Mathur, A. P., & Tippenhauer, N. O. (2016, April). SWaT: A water treatment testbed for research and training on ICS security. In 2016 international workshop on cyber-physical systems for smart water networks (CySWater) (pp. 31-36). IEEE.

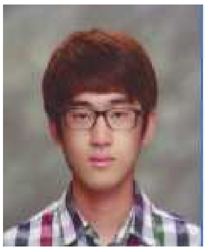
[8] Ahmed, C. M., Palleti, V. R., & Mathur, A. P. (2017, April). WADI: a water distribution testbed for research in the design of secure cyber physical systems. In Proceedings of the 3rd international workshop on cyber-physical systems for smart water networks (pp. 25-28).

[9] Hundman, K., Constantinou, V., Laporte, C., Colwell, I., & Soderstrom, T. (2018, July). Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 387-395).

Thi My Truong received B.S. degree in College of Information & Communication Engineering from Chungbuk National University in 2023. She is currently pursuing the Master degree in Radio Communication Engineering, Chungbuk National University. Her research interests include cybersecurity, Blockchain, AI.



Won Seok Choi received B.S. and Ph.D. degree in the College of Electrical and Computer Engineering, Chungbuk National University, Korea in 2008 and 2014 respectively. He is currently researcher in Research institute of Computer and Information Communication, Chungbuk National University. His research interests include Vehicle network, Energy saving network, SDN, NFV and NGN.



Jang Hyeon Jeong received B.S. and M.S. degree in the College of Electrical & Computer Engineering, Chungbuk National University, Korea in 2019 and 2021. His research interests include Network Security, Smart Grid. He is currently researcher in Xabyss Inc and CEO in JJsolution Inc. His research interest is network security.

Seong Gon Choi received B.S. degree in Electronics Engineering from Kyungpook National University in 1990, and M.S. and Ph.D. degree from KAIST in Korea in 1999 and 2004, respectively. He is currently a professor in College of Electrical & Computer Engineering, Chungbuk National University. His research interests include V2X, AI, smart grid, IoT, mobile communication, high-speed network architecture and protocol.

Router Penetration Testing Based on CEM Vulnerability Assessment Criteria

Tai-Ying Chiu*, Bor-Yao Tseng*, Bagus Tri ATMAJA*,

Jiann-Liang Chen*, Jian-Chang Hsu**

* Department of Electrical Engineering, NTUST (National Taiwan University of Science and Technology)

** ITRI (Industrial Technology Research Institute), Taiwan

M11107508@gapps.ntust.edu.tw, M11107501@mail.ntust.edu.tw, D11207806@mail.nutst.edu.tw, lchen@mail.ntust.edu.tw, hcc@taics.org.tw

Abstract— **As a crucial component of modern network infrastructure, routers' performance can be severely impacted if they fall victim to hacker attacks, affecting all network devices connected to them. Router manufacturers should ensure sufficient cybersecurity measures for their products, such as undergoing information security certification, to prevent customers from encountering unforeseen vulnerabilities when using their routers. In reality, most routers on the market have not been certified for information security, primarily because the certification process is time-consuming and costly. One of the main reasons for this is that security testing during the certification process requires different security validation methods based on the product's functionalities. This study adheres to the Common Methodology for Information Technology Security Evaluation (CEM) testing methodology under the international information security certification standard "Common Criteria." It conducts penetration testing on common router functionalities. It integrates the necessary testing tools, processes, and results to provide a reference for future researchers, reducing testing complexity and enhancing product security assurance.**

Keywords— **Common Criteria, Router, Common Methodology for Information Technology Security Evaluation(CEM), Penetration Testing, Vulnerability Assessment**

I. Introduction

As information technology advances, concerns about information security grow. Routers, central to data transmission in modern networks, are vital. Securing routers is crucial because malicious actors can exploit their vulnerabilities, causing data leaks, network disruptions, and potential attacks. Hence, router security assessment is a vital product indicator.

Since August 1999, an ISO standard called Common Criteria (ISO/IEC 15408) has existed for IT product security certification. However, CC certification is expensive and complex, demanding substantial resources. Consequently, most routers on the market often lack CC or other security certifications.

The complexity and cost of CC certification result from the need to account for diverse product functionalities and designs. This lack of standardization prolongs testing. This paper adheres to CC standards and references CEM documentation. It employs CEM's testing methods and assessment report requirements to document router product security tests. The goal is to compile testing tools and procedures as a reference for future researchers to streamline router testing and reduce duration.

II. Related work

This chapter introduces global infosec standards, relevant security concepts, and research literature on router product security testing methods.

A. Common Criteria

CC (ISO/IEC 15408) is a global standard for evaluating and certifying information security products and systems. It offers a standardized framework for assessing security solutions, ensuring they meet specific requirements. CC has evaluation levels from EAL1 to EAL7. It comprises three main parts: Introduction and General Model [1], Security Functional Requirements [2], and Security Assurance Requirements [3].

B. Common Methodology for Information Technology Security Evaluation

The Common Methodology for Information Technology Security Evaluation (CEM) [4] ensures consistent and effective information security assessments, particularly in the Common Criteria (CC) certification framework. CEM offers specific guidance for assessors, including assessment procedures, testing methods, and report writing to align with CC requirements. It promotes standardization and interoperability in global information security assessments. Our research will follow the AVA_VAN chapter requirements in CEM, focusing on procedures and criteria for analyzing vulnerabilities and weaknesses to assess the target system or product's security.

C. Penetration Testing

Security assessments often employ penetration testing, a method simulated by attackers to uncover vulnerabilities in computer systems, applications, or networks [5]-[10]. Baluni et al. [8] note its time-consuming complexity, involving tailored security validation methods. Türpe et al. [11] stress the vital role of the testing environment to prevent harm, data leaks, disruptions, and legal issues. Zakaria et al. [12] highlight the need for standardized report formats, as reports may differ across companies.

[5]-[9] introduce practical penetration testing tools like Nmap, Metasploit, and Wireshark, and mention Kali Linux in [10]. Kali Linux, a specialized Debian-based OS, is designed for testing and evaluating infosec tools. It offers a range of network testing, vulnerability scanning, and password cracking tools, empowering security professionals to enhance system security. Kali Linux comes with pre-installed or installable versions of these tools, making it an ideal testing platform for our research.

D. Related Research on Router Vulnerabilities

Karamanos Emmanouil [13] delves into router security, illustrating router-targeted attacks. Niemietz Marcus et al. [14] performed XSS and UI redressing attacks on routers from ten manufacturers, showcasing rapid fingerprinting attacks and suggesting countermeasures. Fuzz testing by F. Li et al. [15] unveiled Ping of Death and Denial of Service flaws in Cisco routers. Jin-bing Hou et al. [16] summarized firmware vulnerabilities in embedded systems, covering multiple router models and offering an overview of vulnerability detection.

III. METHOD

In the documents we tested, based on the previously mentioned CEM, and with Zyxel company's firewall product ATP 100 as our testing target. In the Vulnerability Assessment document standard, the testing phase is divided into 11 steps, and we will conduct the evaluation of the Target of Evaluation (TOE) in accordance with the testing specifications and its assessment report.

During the assessment phase, the testing is divided into 11 sections, including requirements, evaluator findings, rationale, and conclusions.

A. AVA_VAN. 1-1

First is the requirement of AVA_VAN.1-1, which demands that the evaluator, when inspecting the TOE, must confirm the consistency of the test configuration with the evaluation configuration specified in the Security Target (ST). Therefore, it is necessary to examine its features and firmware version. At the time of our testing, the device had a firmware version of V5.37(ABPS.0), which is newer than the V5.31 indicated in the User's guide. Furthermore, a comparison between the User's guide and the TOE's backend management revealed feature consistency, and in terms of the physical configuration of the device, it aligns with the specifications provided in the ST. Therefore, the conclusion drawn at this stage is that the test configuration of the Zyxel ATP 100 is consistent with the evaluation configuration in the ST.

B. AVA_VAN. 1-2

The requirement at this stage is for the evaluator to verify that the TOE is correctly installed and ensure that it is in a known state. In this assessment, the evaluator follows the steps outlined in the User's guide, as follows:

1. Plug the device into the power cord and ensure that it is receiving power.
2. Turn on the device.
3. Use RJ-45 to connect the wall outlet to the WAN interface of the ATP-100.
4. Use RJ-45 to connect to the computer from the LAN port of the ATP 100.
5. Configure the device from the web user interface (192.168.1.1).
6. Follow the "Initial Setup".
7. Configure the WAN Network address
   IP Address: 140.118.xxx.xxx
   Subnet Mask: 255.255.255.0
   Gateway: 140.118.xxx.xxx
8. Upgrade the firmware to the newest available version by following these steps
   Maintenance -> File Manager -> Firmware Management -> Upgrade
9. The device is ready, and it is working properly.

After the installation is completed, a comprehensive inspection of the TOE is carried out to ensure that each function operates correctly, indicating the successful installation and initial configuration.

C. AVA_VAN. 1-3

Next is AVA_VAN.1-3, where the requirement is for the evaluator to search in publicly available information to identify potential vulnerabilities in the TOE. In this stage, we utilized various publicly available sources such as Zyxel's official security advisory page and the MITRE Corporation's Common Vulnerabilities and Exposures (CVE) database. Here are some of the CVE vulnerabilities we found related to the Zyxel ATP series: CVE-2020-29299, CVE-2021-35029, CVE-2022-0342, CVE-2022-0734, CVE-2022-0910, CVE-2022-2030, and more. Additionally, we identified some common patterns within these vulnerabilities, such as potential vulnerabilities in the device's web interface leading to Cross-Site Scripting (XSS), risks associated with device functionality during authentication, and potential exploitation via Directory Traversal to access undisclosed confidential data in the TOE. Since the impact and exploitation of each vulnerability require significant expertise, this information is considered essential for evaluating the TOE. In this stage, we present the results of identifying past vulnerabilities in the ATP series as part of completing AVA_VAN.1-3.

D. AVA_VAN. 1-4

The requirement of AVA_VAN.1-4 is for the evaluator to document the identified potential vulnerabilities, which are candidates for testing and applicable to the TOE within its operational environment. In our actual examination, we can locate the section titled "Security Problem Definition" in the TOE's Security Target document and analyze potential vulnerabilities that may exist therein. We found vulnerabilities that may exist in the TOE documented in the file, which are considered test candidates, as shown in Table 1. Once the stage is complete, and testing of vulnerability issues can proceed to the next step.

E. AVA_VAN. 1-5

In the requirement standard AVA_VAN.1-5, the evaluator must take the potential vulnerabilities identified through research and design penetration tests. Before commencing testing, it is necessary to prepare the tools required during the testing process, such as NMAP, Wireshark, LOIC, and others, as detailed testing items correspond to the tools mentioned in Table 1.

F. AVA_VAN. 1-6

Next is AVA_VAN.1-6, where the requirement is for the evaluator to thoroughly document the penetration testing based on the list of potential vulnerabilities, ensuring that the testing can be replicated. Additionally, the documentation should include the prior preparation of testing tools and the anticipated test outcomes. Following the above guidelines, during our evaluation, we conducted tests on the TOE using the predetermined candidates outlined in AVA_VAN.1-4, utilizing the equipment and configurations established in AVA_VAN.1-5. The necessary tools and preparations for the testing candidates are listed in Table 3, and the expected outcomes are detailed in Table 2.

G. AVA_VAN. 1-7

The requirement of AVA_VAN.1-7 is for the evaluator to initiate penetration testing, and our testing findings are summarized in Table 1.

TABLE 1. SUMMARY OF PENETRATION TESTING

Test Candidate	**Test Candidate**	**Tests Performed**
T.WEAK_CRYPTOGRAPHY	Wireshark	Packets use a reliable protocol, but TOE can't identify QUIC certificates.
T.NETWORK_ACCESS	IP block list, Scapy, Wireshark	List the target IP and use it to log in to Zyxel. Besides, we use Scapy to send a packet and Wireshark to check it.
Buffer overflow attack	NMAP	No buffer overflow vulnerability was found in the test report.
OS command injection	NMAP, Burp Suite	No command injection went through the router.
T.UNAUTHORIZED_ADMINISTRATOR_ACCESS	Window CMD, ZYXEL Backstage, Google Authenticator	Enable two-factor in UI settings; configure Email and Google authenticator. Verified if the device asking extra code with two-factor enabled. The device sends and verifies a code for admin page access.
T.NETWORK_DISCLOSURE	Window CMD, ZYXEL Backstage	Set up the target IP in the block list, and then test whether the router can ping it and receive packets.
T.WEAK_AUTHENTICATION_ENDPOINTS	Wireshark, VPN setup of Zyxel ATP 100	Using IKE for IPsec VPN setup. Wireshark to inspect the encryption used for the packet.
T.UPDATE_COMPROMISE-upgrade test	HxD	Using HxD, modify the firmware update file, and update the modified file to verify if any issues.
T.UNTRUSTED_COMMUNICATION_CHANNELS	Wireshark	Utilize Wireshark to verify TOE's management interface protocol and reliable certificates.
T.NETWORK_DOS	Hping3, LOIC, NMAP, Botnet Filter	Configure a computer as the target under a router and another with an external IP to launch a DoS attack.
T. DATA_INTEGRITY	Sandbox	Malware downloaded and then check logging and Sandbox service.
T.MALICIOUS_TRAFFIC	LOIC	Using LOIC to simulate DoS.

H. AVA_VAN. 1-8

Following AVA_VAN.1-7, the next is AVA_VAN.1-8, where the requirement is for the evaluator to document the actual results of the penetration testing. In our documentation of the actual results, we will compare them with the expected outcomes previously defined in AVA_VAN.1-7, as detailed in Table 2.

TABLE 2. COMPARISON OF EXPECTED AND ACTUAL RESULTS FOR PENETRATION TESTING

Test Candidate	**Expected result**	**Actual Result**
T.WEAK_CRYPTOGRAPHY	SSL Inspection: Encrypts, decrypts, and scans reliable protocol packets; blocks unreliable ones.	SSL Inspection doesn't support QUIC, allowing its packets to pass uninspected.
T.NETWORK_ACCESS	It can block connections, packets, and access permissions.	Blocks computers from connecting but allows target login.
Buffer overflow attack	If NMAP finds a weakness then receive a warning of “ProFTPD server TELNET IAC stack overflow.”	The NMAP report didn’t include Buffer overflow vulnerability.
OS command injection	If the URL command injection exploit succeeds, router config files are visible. If NMAP shellshock testing detects vulnerability, a warning will received.	URL command injection failed, the router redirects to the login page. The NMAP report didn’t include command injection vulnerability.

T.UNAUTHORIZED_ADMINISTRATOR_ACCESS	Receive a two-factor authentication message through SMS, E-mail, or Google Authenticator, while logging in.	Successful two-factor with Google Authenticator; TOE sends and validates codes.
T.NETWORK_DISCLOSURE	Shouldn't receive packet from IP block list, through both Router CLI and PC cmd.	In the router's CLI, ping to the blocked IP works, but the PC's CMD doesn't receive packets from the blocked IP.
T.WEAK_AUTHENTICATION_ENDPOINTS	The encryption is by IKE and used for packet transmission during the intermediate stages.	The payload is encrypted and authenticated by IKE.
T.UPDATE_ COMPROMISE-upgrade test	The TOE detects that the update file is insecure, resulting in a failed update and displaying a warning.	The TOE detects that the update file is insecure, resulting in a failed update and displaying a warning.
T.UNTRUSTED_COMMUNICATION_CHANNELS	Each packet of TOE's configuration interface uses a highly reliable protocol and certificate.	TOE employs TSLv1.3, yet the browser distrusts the Zyxel-signed certificate.
T.NETWORK_DOS	TOE detects DOS attacks, and blocks malicious traffic. Script opens two server connections, and lacks final CRLF. After 10 seconds, the second connection sends an extra header. If the second times out after the first, the header extends timeout, exposing Slowloris DoS vulnerability.	TOE effectively shields protected devices from external IP DoS attacks. The Botnet Filter which can defend zombie devices, didn't show up on the device. Nmap report reveals no DoS vulnerability in open port scan.
T. DATA_INTEGRITY	Detect malware files and intercept them at the same time.	Malware downloads: No detection by Logging and Sandbox services.
T.MALICIOUS_TRAFFIC	Using a DoS attack to test whether the TOE will crash.	After being attacked, the TOE will crash.

I. AVA_VAN. 1-9

Next is AVA_VAN.1-9, where the requirement is for the evaluator to document the actual results of penetration testing, providing an overview of the testing methodology, configurations, depth, and outcomes. After conducting penetration tests tailored for the TOE, we provide detailed descriptions of the testing results, as outlined in Table 3.

TABLE 3. THE DETAILED RESULTS OF PENETRATION TESTING

Test Candidate	Expected result	TSFI	Actual Result
T.WEAK_CRYPTOGRAPHY	Wireshark: Check the protocol version of each package.	GUI	SSL Inspection doesn't support QUIC, allowing its packets to pass uninspected.
T.NETWORK_ACCESS	IP block list, Scapy, Wireshark	GUI	Blocks computers from connecting but allows target login.
Buffer overflow attack	Run NMAP ftp-vuln-cve2010-4221 testing	CLI	The NMAP report didn't include Buffer overflow vulnerability.
OS command injection	Adding URL command (Burp Suite) Run nmap -sV -p- --script http-shellshock <target>	CLI	URL command injection failed, the router redirects to the login page. The NMAP report didn't include command injection vulnerability.
T.UNAUTHORIZED_ADMINISTRATOR_ACCESS	Window CMD, ZYXEL Backstage, Google Authenticator	CLI/GUI	Successful two-factor with Google Authenticator; TOE sends and validates codes.
T.NETWORK_DISCLOSURE	Window CMD, ZYXEL Backstage	CLI	In the router's CLI, ping to the blocked IP works, but the PC's CMD doesn't receive packets from the blocked IP.
T.WEAK_AUTHENTICATION_ENDPOINTS	Wireshark, VPN setup of Zyxel ATP 100	GUI	The payload is encrypted and authenticated by IKE.
T.UPDATE_ COMPROMISE-upgrade test	HxD: Modify firmware upgrade file.	GUI	The TOE detects that the update file is insecure, resulting in a failed update and displaying a warning.
T.UNTRUSTED_COMMUNICATION_CHANNELS	Wireshark: Verify TOE's config interface protocol version.	GUI	TOE employs TSLv1.3, yet the browser distrusts the Zyxel-signed certificate.
T.NETWORK_DOS	Hping3, LOIC: Conduct a DOS attack on the device protected by TOE. NMAP, Botnet Filter	CLI	TOE effectively shields protected devices from external IP DoS attacks. The Botnet Filter which can defend zombie devices, didn't show up on the device. Nmap report reveals no DoS vulnerability

			in open port scan.
T. DATA_INTEGRITY	Sandbox	GUI	Malware downloads: No detection by Logging and Sandbox services.
T.MALICIOUS_TRAFFIC	LOIC	GUI	After being attacked, the TOE will crash.

J. *AVA_VAN. 1-10*

In AVA_VAN.1-10, the requirement is for the evaluator to examine the results of all penetration tests to ensure that the TOE can withstand attackers with basic attack potential in its operational environment. Here, we have reviewed all the test results and calculated the attack potential for each factor. Tables 4, 5, and 6 display the Attack Potential Calculation (APC) values we have computed for the TOE.

TABLE 4. ASSESSMENT OF APC VALUES FOR EACH TEST (1)

Test Candidate	**Buffer overflow attack**	**OS command injection**	**T.UNAUTHORIZED_ADMINISTRATOR_ACCESS**	**T.NETWORK_DISCLOSURE**
Factor	Value	Value	Value	Value
Elapsed Time	7	4	4	10
Expertise	6	3	3	8
Knowledge of TOE	4	4	3	4
Window of Opportunity	10	4	4	4
Equipment	4	4	9	4
Total	31	19	23	30

TABLE 5. ASSESSMENT OF APC VALUES FOR EACH TEST (2)

Test Candidate	**T.NETWORK_DOS**	**T. DATA_INTEGRITY**	**T.WEAK_CRYPTOGRAPHY**	**T.UPDATE_COMPROMISE-upgrade test**
Factor	Value	Value	Value	Value
Elapsed Time	1	1	7	10
Expertise	3	0	6	6
Knowledge of TOE	3	0	3	3
Window of Opportunity	4	0	4	10
Equipment	4	0	4	4
Total	15	1	24	33

TABLE 6. ASSESSMENT OF APC VALUES FOR EACH TEST (3)

Test Candidate	**T.UNTRUSTED_COMMUNICATION_CHANNELS**	**T.NETWORK_ACCESS**	**T.WEAK_AUTHENTICATION_ENDPOINTS**	**T.MALICIOUS_TRAFFIC**
Factor	Value	Value	Value	Value
Elapsed Time	19	7	10	1
Expertise	6	6	6	3
Knowledge of TOE	7	3	7	0
Window of Opportunity	10	10	4	1
Equipment	7	4	7	0
Total	51	30	34	5

K. *AVA_VAN. 1-11*

Finally, in AVA_VAN.1-11, the requirement is for the evaluator to report all exploitable vulnerabilities and any remaining vulnerabilities, along with detailed explanations of their sources. After a series of tests examining public vulnerabilities and security threats, the last stage involves summarizing the results. Out of the 12 testing items, the TOE failed in the following categories: T.MALICIOUS_TRAFFIC, T.UNTRUSTED_COMMUNICATION_CHANNELS, and T. DATA_INTEGRITY. Detailed information regarding these failures will be provided in Chapter 4, while all other testing candidates pass successfully.

IV. EVALUATION

In the penetration testing conducted on the router, we followed the Vulnerability Assessment standards outlined in the CEM, planning and documenting the testing process and results. Throughout this process, we conducted thorough testing of the TOE. However, we discovered that several testing items did not fully align with the functionalities mentioned in its User's guide. Below are our findings.

Regarding T.UNTRUSTED_COMMUNICATION_CHANNELS, this test aimed to confirm the TOE's ability to use highly reliable protocols and certificates. While the TOE uses the highly reliable TLSv1.3 protocol, browsers did not trust the certificates signed by Zyxel, resulting in a test failure. In T.DATA_INTEGRITY, the test was designed to evaluate the TOE's Sandbox functionality, which isolates unknown files and detects new malware to enhance network security. However, during testing, the TOE failed to detect and block malicious files, leading to a test failure. The handbook mentions that the Sandbox should be able to identify unknown threats, but this was not achieved in this test. Lastly, in T.MALICIOUS_TRAFFIC, we simulated a DoS attack to evaluate the TOE's firewall functionality for blocking malicious traffic and DoS attacks. The test results showed that the TOE encountered issues during a DoS attack and did not provide the expected protection, resulting in a test failure.

V. CONCLUSION

In today's technologically advanced era, there has been a significant increase in general awareness of information security in daily life. Concurrently, to ensure that users can use information and communication products such as routers with greater confidence, various countries have introduced a certification framework for the security of IT products known as Common Criteria. This article is based on the certification framework mentioned above, utilizing the Common Methodology for Information Technology Security Evaluation. It is a specialized method for evaluating information security and we implement penetration testing for routers within this framework. This includes recording the development of testing tools, procedures, and results as required by AVA_VAN in the CEM. Additionally, certain vulnerabilities and shortcomings in the TOE are identified.

The current information security issues are of utmost significance, and international certification standards are a major focus. This article takes routers as an example and is written based on the testing methodology outlined in the CEM. It aims to provide a reference for future researchers, reduce testing complexity, and enhance product security assurance.

References

[1] "Common Criteria for Information Technology Security Evaluation – Part 1: Introduction and general model," November 2022.

[2] "Common Criteria for Information Technology Security Evaluation – Part 2: Security functional components," November 2022.

[3] "Common Criteria for Information Technology Security Evaluation – Part 3: Security assurance components," November 2022.

[4] "Common Methodology for Information Technology Security Evaluation – Evaluation methodology," November 2022.

[5] H. M. Z. A. Shebli and B. D. Beheshti, "A study on penetration testing process and tools," 2018 IEEE Long Island Systems, Applications and Technology Conference (LISAT), Farmingdale, pp. 1-7, NY, USA, 2018.

[6] Robert Shimonski, "Conducting a Penetration Test," in Penetration Testing For Dummies , Wiley, pp.129-145, 2020.

[7] P. Anand and A. Shankar Singh, "Penetration Testing Security Tools: A Comparison," 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), pp. 182-184, MORADABAD, India, 2021.

[8] S. Baluni, S. Dutt, P. Dabral, S. Maji, A. Kumar and A. Chaudhary, "Penetration Testing on Virtual Machines," 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1-6, Noida, India, 2022.

[9] M. Patel and H. R. Patel, "Analytical Study of Penetration Testing for Wireless Infrastructure Security," 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), pp. 131-134, Chennai, India, 2019.

[10] D. Sweigert, M. M. Chowdhury and N. Rifat, "Exploit Security Vulnerabilities by Penetration Testing," 2022 IEEE International Conference on Electro Information Technology (eIT), pp. 527-532, Mankato, MN, USA, 2022.

[11] S. Türpe and J. Eichler, "Testing Production Systems Safely: Common Precautions in Penetration Testing," 2009 Testing: Academic and Industrial Conference - Practice and Research Techniques, pp. 205-209, Windsor, UK, 2009.

[12] N. Zakaria, P. A. Phin, N. Mohmad, S. A. Ismail, M. N. Kama and O. Yusop, "A Review of Standardization for Penetration Testing Reports and Documents," 2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS), pp. 1-5, Johor Bahru, Malaysia, 2019.

[13] Karamanos Emmanouil, "Investigation of home router security," 2010.

[14] Niemietz Marcus and Jörg Schwenk, "Owning your home network: Router security revisited," arXiv preprint arXiv:1506.04112, 2015.

[15] F. Li, L. Zhang and D. Chen, "Vulnerability mining of Cisco router based on fuzzing," The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014), pp. 649-653, Shanghai, China, 2014.

[16] Jin-bing Hou, Tong Li and Cheng Chang, "Research for Vulnerability Detection of Embedded System Firmware," Procedia Computer

Tai-Ying Chiu was born in Taiwan, in 1999. She received the B.S. degree, in 2022. She is currently pursuing the M.S. degree in electrical engineering with the National Taiwan University of Science and Technology, Taipei. Her main research interests include artificial intelligence, and the Internet of Things (IoT).

Bor-Yao Tseng was born in Taiwan, in 1999. He received the B.S. degree, in 2022. He is currently pursuing the M.S. degree in electrical engineering with the National Taiwan University of Science and Technology, Taipei. His main research interests include artificial intelligence, and the Internet of Things (IoT).

Bagus Tri Atmaja, born on December 4th, 1996, is currently a doctoral candidate in Electronic Engineering at the National Taiwan University of Science and Technology (NTUST). He is the third son of his family from East Java, Indonesia. Bagus completed his Master's Degree at NTUST in 2023, following a Bachelor's Degree from the Institut Teknologi Sepuluh Nopember (ITS), Indonesia, in 2019.

Jiann-Liang Chen Prof. Chen was born in Taiwan on December 15, 1963. He received the Ph.D. degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan in 1989. Since August 1997, he has been with the Department of Computer Science and Information Engineering of National Dong Hwa University, where he is a professor and Vice Dean of Science and Engineering College. Prof. Chen joins the Department of Electrical Engineering, National Taiwan University of Science and Technology, as a Distinguished professor and Dean now. His current research interests are directed at cellular mobility management, cybersecurity, personal communication systems and Internet of Things (IoT). The second paragraph uses the pronoun of the person (he or she) and not the author's last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (city, state: publisher name, year) similar to a reference. Current and previous research interests end the paragraph.

Hybrid Clustering Mechanisms for High-Efficiency Intrusion Prevention

Pin-Shan Lin, Yi-Cheng Lai, Man-Ling Liao, Shih-Ping Chiu and Jiann-Liang Chen

Department of Electrical Engineering, National Taiwan University of Science and Technology
Taipei, Taiwan

m11107510@mail.ntust.edu.tw, m11007505@mail.ntust.edu.tw, m11007502@mail.ntust.edu.tw, spchiu@mail.ntust.edu.tw, lchen@mail.ntust.edu.tw

Abstract— **With the advancement of information and communication technology, cyberattack techniques have evolved into increasingly complex trends. Malicious network traffic attacks have become one of the information security problems for all organizations. This study is aimed to combat malicious network traffic attacks by actively collecting commands from attackers using honeypots. It involves pre-processing the raw network traffic data, employing a K-means algorithm to group the payloads, and label payloads using the MITRE ATT&CK framework. To improve the accuracy of the generated snort rules, the system utilizes Locality-Sensitive Hashing (LSH) method for secondary clustering, combined with snort rule generation, to form a comprehensive intrusion prevention system. In addition, to speed up the experimental process, this study adapted a script for this system to simulate an attacker's attack automatically. Through experimentation, it can be observed that hybrid clustering techniques such as K-means and LSH mechanisms can yield a defensive effectiveness of up to 93% for malicious payloads. This result proves the system's ability to identify and prevent different packet attacks effectively.**

Keywords—**K-means Algorithm, MITRE ATT&CK, Snort, Locality Sensitive Hashing (LSH), Malicious Packet**

I. Introduction

As technology continues to advance, the popularity of the Internet continues to expand, profoundly affecting people around the globe. This trend has changed how people live their daily lives and reshaped the nature of business, information dissemination, and social interaction. This phenomenon has led most people to adopt the Internet as the primary medium for shopping, information retrieval, and social interaction. The Internet contains a wide variety of private and public information, making it a complex environment full of opportunities and challenges.

The widespread adoption of the global network has been accompanied by a startling surge in cyberattacks in recent years. The Internet has become a lucrative tool or medium for various illicit purposes. According to the latest data, in 2022, the annual growth rate of global cyberattack incidents reached 38% [1], with organizations facing an average of 1,168 attacks per week. These attacks come from hackers targeting financial institutions, government agencies, healthcare organizations, and military organizations. Furthermore, according to the IBM Data Breach Report [2], the average cost of global data breaches has reached $4.45 million by 2023, a 15% increase from three years ago. According to the Trend Micro 2023 Mid-Year Roundup Full Report, cybercriminal groups are continually developing new tools and techniques to enhance attack rates while reducing the chances of their attacks being detected. For instance, APT34, in its latest campaign, employed Command and Control (C&C) communication concealed within legitimate Simple Mail Transfer Protocol (SMTP) email traffic. Another group, Earth Longzhi, also utilized a novel technique known as stack rumbling in its attacks [3]. The reports mentioned above highlight the escalating diversity in attackers' tactics. To mitigate harm and prevent economic or data losses, government entities, and private enterprises must adopt innovative approaches to effectively detect and defend against evolving hacker attacks.

This study employs a hybrid approach that combines machine learning algorithms and hashing algorithms to conduct further analysis of network packets and generate firewall rules for detecting malicious packets and issuing system warnings. The primary objective of this method is to enhance the performance of the Intrusion Prevention System (IPS) so that it can more effectively respond to various types of network packet attacks and achieve better prevention capabilities. The techniques mentioned above will be described in detail in the next section.

II. Related Work

This section is an introduction to the techniques used in this research. Related work can be categorized into three parts: literature on malicious packet detection, analysis of the application of LSH and associated methods, and an overview of snort and its usage.

The diversity of cyberattacks has made malicious network traffic detection a widely researched area. Iñigo Perona et al. [4] combined payload-based and packet-header-based techniques for pre-processing raw packets. They analysed raw packets using standard compression algorithms like GZip, n-gram models, and the byte-frequency idea. Hongyu Liu et al. [5] proposed a novel payload classification method based on neural networks for detecting network attacks. This method utilizes deep learning models to analyse payloads and employs

PL-CNN and PL-RNN to automatically extract features without feature engineering, finally achieving an effective classification result. Bontupalli et al. [6] proposed memristor-crossbar-based architectures for high-speed packet classification and instruction detection systems. This system combines software and hardware, enabling rapid and accurate detection of malicious packet attacks.

LSH is an algorithm that reduces the data query range while preserving data information. It is an important tool for fast similarity matching in various applications, such as searching for similar documents, images, and audio [7], [8]. Its applications include different fields, including image deduplication and recommendation systems. To make the snort rule more adaptable to different attack packets, this study uses LSH to calculate the similarity based on the payload content and performs secondary clustering.

Snort is an open-source Intrusion Detection System (IDS) and Intrusion Prevention System (IPS) widely evaluated in the network security domain with excellent real-time traffic analysis and matching rules. Using snort makes it possible to effectively monitor abnormal behaviour, potential attack patterns, and the transmission of malicious packets in network traffic [9]. Fahmida Alam Rafa et al. [10] developed a Cloud-based Intrusion Detection System to deal with different cloud-based attack techniques, which can be applied in both traditional and virtual networks within a cloud environment

This paper combines the above three techniques based on the system proposed in [11], [12]. It offers a malicious payload prevention system, which can detect the payloads with pre-written rules into the snort rules and prevent malicious packets without extracting the packet features.

III. Proposed Method

This study proposes a malicious packet-based detection and prevention system. This system uses a honeypot to collect threatening packet information and commands actively, further analyses them, and generates corresponding preventive measures and rules. As shown in Figure 1, there are five main steps: A. Pre-processing the data collected from the honeypot. B. Grouping similar payloads together into each cluster. C. Labeling payloads with MITRE ATT&CK. D. Calculating payload similarity and clustering using Locality-Sensitive Hashing. E. Generating snort rules through regular expressions.

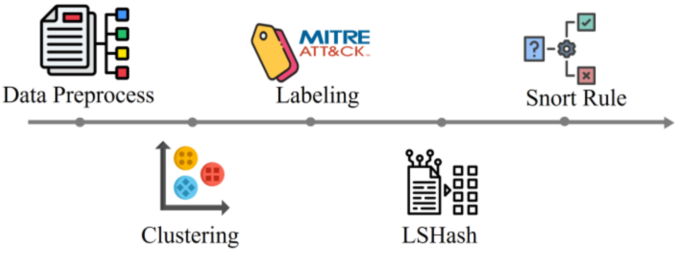


Figure 1. Proposed System Architecture

A. Pre-processing Data

This system uses the Cowrie Honeypot as a trap to lure potential attackers into the system. Cowrie is a highly interactive honeypot designed to capture and log sessions initiated by attackers using SSH and Telnet protocols. This tool is primarily used to record messages sent by attackers or actions related to attackers, enabling information security experts to study attack behaviours, attack patterns, and strengthen defensive measures. In this system, we pre-process the data once the honeypot receives packets sent by potential attackers. The initially received data is in JSON format, and to analyse these packet details more effectively, this study segments the packets based on timestamp and reorganizes them into a tabular dataset. Each packet in the table contains the following information: session ID, packet ID, packet payload, and timestamp. These actions simplify the original complex data structure and make the data content more organized and readable.

B. Payload Clustering

The data collected by Cowrie Honeypot is pre-processed to ensure optimal utilization of the data, including checking and cleaning of redundant data. Subsequently, to enhance data consistency and analyzability, payload data is transformed into vector form using natural language processing techniques such as the BERT encoder. To provide more helpful information for snort rule generation in the future and to help understand the distribution of malicious packets sent by attackers, the transformed payloads are then clustered using the K-means algorithm.

K-means, as an unsupervised learning algorithm, is utilized to identify data clusters with similar behaviours or distributions without labels. It is commonly applied in tasks like image compression and document classification. In the K-means algorithm, clustering is based on distance calculations between data points to determine cluster centroids. The distance between each point and the centroid is computed, as shown in Figure 2.

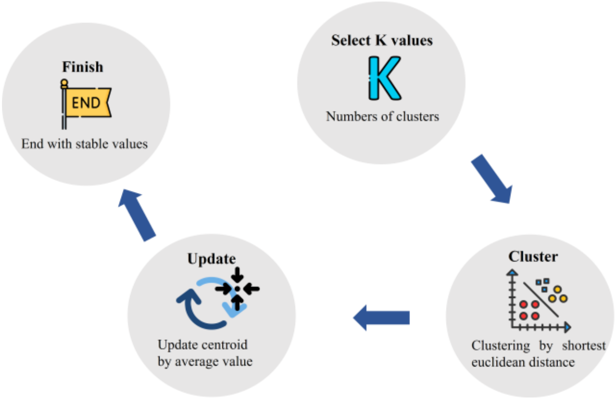


Figure 2. K-means Algorithm

C. Labeling with MITRE ATT&CK

MITRE ATT&CK is a concept of ATT&CK (Adversarial Tactics, Techniques & Common Knowledge) developed by MITRE, a non-profit organization, to help organizations understand and analyse cyber attacks using a common language and framework. The core focus of the MITRE ATT&CK framework lies in the TTP components, which are tactics, techniques, and procedures. TTPs enable cybersecurity researchers to understand the attacker's tactics, allowing for further in-depth analysis of attack methods.

Following K-means clustering, data is labeled based on the payload content using MITRE ATT&CK's Tactic, Technique, and Subtechnique. Tactic reflects the adversary's objective in executing a particular action and is used for categorizing and understanding attack behaviour, as shown in Table 1. Technique describes the methods or techniques used by the attacker in achieving the tactical goal, including the related behaviours, the tools used, the impact of the attack, and other details, which are helpful to understanding the attacker's method of operation, as shown in Table 2. Subtechnique provides a more specific description of the methods attackers use when implementing a Technique. Each Technique may involve multiple Subtechniques, including different tools, attack approaches, or subtle variations, as illustrated in Table 3.

TABLE 1. THE CURRENT MITRE ATT&CK TACTIC

Tactic	Decription
TA0043 Reconnaissance	Attacker collect information about the target network or system to understand the potential target of the attack.
TA0042 Resource Development	The tools, vulnerabilities, and resources that an attacker intends to use in an attack, including developing malware or gathering effective attack resources.
TA0001 Initial Access	Attackers gain first access to a target system through a variety of approaches such as vulnerabilities, social engineering
TA0002 Execution	Attackers execute malicious code or commands on a target system in order to achieve his or her attack objectives

TABLE 2. SOME TECHNIQUES IN TACTIC: EXECUTION

Tactic: Execution(TA0002)	
Technique	**Decription**
T1053 Scheduled Task/Job	Adversaries use task scheduling for malicious code execution, requiring admin privileges.
T1204 User Execution	Adversaries use social engineering to trick users into executing malicious code via files or links.
T1059 Command and Scripting Interpreter	Adversaries may exploit command and script interpreters for executing commands, scripts, or binaries.
T1072 Software Deployment Tools	Adversaries utilize third-party software suites present within an enterprise network.

TABLE 3. SUBTECHNIQUES IN TECHNIQUE: COMMAND AND SCRIPTING INTERPRETER

Technique: Command and Scripting Interpreter (T1059)		
ID	**Name**	**Decription**
T1059.001	PowerShell	Adversaries exploit PowerShell for the purpose of execution.
T1059.002	AppleScript	Adversaries exploit AppleScript for the purpose of execution.
T1059.003	Windows Command Shell	Adversaries exploit Windows Command Shell for the purpose of execution.
T1059.004	Unix Shell	Adversaries exploit Unix Shell for the purpose of execution.
T1059.005	Visual Basic	Adversaries exploit Visual Basic for the purpose of execution.
T1059.006	Python	Adversaries exploit Python for the purpose of execution.
T1059.007	JavaScript	Adversaries exploit JavaScript for the purpose of execution.
T1059.008	Network Device CLI	Adversaries exploit Network Device CLI for execution.
T1059.009	Cloud API	Adversaries exploit Cloud API for the purpose of execution.

D. Locality-sensitive Hashing

LSH is one of the most common techniques in the Approximate Nearest Neighbor Algorithm (ANN). Its core concept revolves around applying a hash function to two neighboring data points in space, such that after hashing, these two data points remain close to each other in the area. LSH transforms data vectors into hash values while preserving information about their similarity. This approach reduces the search space, as illustrated in Figure 3. The algorithm mainly identifies similar data points from a large dataset. It finds practical applications in tasks such as text similarity detection and web search.

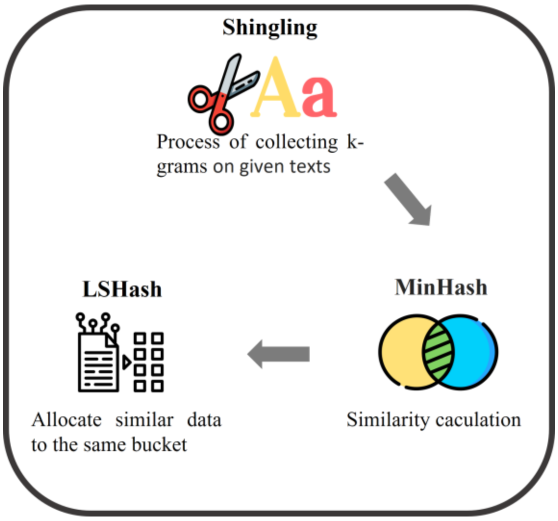

Shingling
Aa
Process of collecting k-grams on given texts
LSHash
MinHash
Allocate similar data to the same bucket
Similarity caculation

Figure 3. Locality-sensitive Hashing Algorithm Flow

In this phase, further processing is applied to the data that was previously clustered using K-means and labeled with

MITRE ATT&CK. LSH is utilized for clustering, grouping similar payloads into the same clusters to generate more clusters. These clusters will be used in the future to create snort rules, with each rule targeting the data from different clusters. The goal of this step is to precisely address specific attack patterns and behaviours within different clusters, thereby effectively generating more accurate snort rules. The typical processing involves the following steps:

1) Shingling: Shingling is a method based on k-grams, where k-grams are composed of a series of consecutive tokens. Specifically, these tokens can be words or symbols, depending on the context of the text. The ultimate goal of Shingling is to encode each document by utilizing the previously collected k-grams, thereby achieving a representation and analysis of textual data. This encoding process is a crucial step in data processing, as it allows for capturing patterns and features within documents, serving as the foundation for subsequent analysis and search tasks.

2) Min-Hashing: Due to the computational cost of calculating similarities between two one-hot encoded vectors, min-Hashing is used to convert one-hot encoded vectors into smaller-dimensional but denser vectors while retaining their similarity information. Min-Hashing is a hash function that rearranges the input vector and returns the index of the first value equal to 1. To obtain a denser vector of length n, referred to as a signature, n min-Hash functions are used to obtain n min-Hash values.

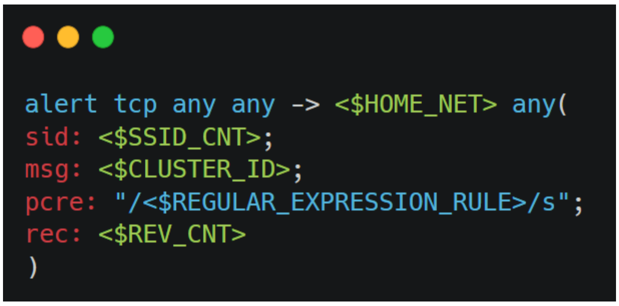
3) LSH: After applying Min-Hashing, we obtain signatures that retain similarity information. However, these dense signatures are still high-dimensional, and it is challenging to compare them efficiently. To address this issue, we use LSH to solve the above problem. LSH utilizes a signature matrix divided into b parts, each containing r rows known as subsignatures. Each part has r rows called subsignatures, which are then processed by a hash function. For instance, there are subsignature A and subsignature B, which need to compare their similarity. At this time, they are divided into b parts, and their respective b subsignatures are extracted and processed by the hash function. The final step involves determining whether the subsignatures of subsignature A and subsignature B are distributed into the same bucket based on hash values. Suppose the hash values of subsignature A and subsignature B indicate that they belong to the same bucket with a sufficient number of shared hash values. In that case, it can be concluded that subsignature A and sub-signature B are similar.

E. Snort Detection Rules

This study uses Snort as the primary tool for malicious packet detection and analysis. Depending on various objectives, rules for Snort can be crafted and optimized accordingly [14]. Through the previously mentioned steps, an analysis of payload similarity was conducted, resulting in a total of 42 clusters. Next, this system will utilize snort based on Perl Compatible Regular Expressions (PCRE) to provide a more structured approach to assist the well-clustered groups in generating more standardized representations, as depicted in Figure 4. The meanings of each parameter are detailed in Table 4.

TABLE 4. DETAILS OF SNORT RULE

Snort Rule	Decription
sid	The sid keyword is used to be the representation of unique snort rule.
msg	The "msg" rule option specifies the message for log/alert, with packet dump or an alert.
pcre	The "pcre" allows user to design the Regular Expression to be the rules.
rev	The "rev" keyword identifies Snort rule's revision number uniquely.



```
alert tcp any any -> <$HOME_NET> any(
sid: <$SSID_CNT>;
msg: <$CLUSTER_ID>;
pcre: "/<$REGULAR_EXPRESSION_RULE>/s";
rec: <$REV_CNT>
)
```

Figure 4. Snort Rule with PCRE

This system employs an algorithm to generate snort rules based on a Regular Expression generator constructed using a Genetic Algorithm [15]. The concept of the Genetic Algorithm is to find a regular expression among a group of similar strings or texts that best matches the characteristics of all strings within that group. This optimal regular expression is then written into PCRE, enabling snort to detect and efficiently identify malicious behaviours associated with that cluster.

IV. Experiment

A. Test and Verify

It is necessary to establish a corresponding testing environment and test the results with appropriate tools to validate whether the snort rule generator produces snort rules that can be detected effectively. The testing environment used here is shown in Table 5. The testing process involves two primary roles: the sender and receiver. The snort firewall is set up on the receiver side, with the Snort rules generated by the Genetic Algorithm being written into the configuration file. This configuration specifies the network segments and ports to be protected. Once the initial configuration is completed, testing can be started.

TABLE 5. TESTING ENVIRONMENT

Hardware	Use in this study
Virtual Machine	VMware Work Station
OS System	Kali-linux2023.3

The testing process for this research is illustrated in Figure 5. In this procedure, the sender transmits packets to the receiver, where the receiver uses the snort-generated firewall to inspect these packets. It takes appropriate actions based on preconfigured rules for the payload. The sender utilizes a Python script written with scapy to generate attack packets and send them to the receiver, simulating the behaviour of an attacker. To achieve process automation, the above process has been scripted, enabling the automated execution of packet sending. This script is designed to improve efficiency and reduce human intervention, make the experiment process more automated and increase repeatability, ensure the consistency and accuracy of the experiment, and reduce possible human errors.

The payloads used in the testing scripts of this study were collected by the Cowrie Honeypot in 2022. After data pre-processing, which involved removing duplicates and unusable data, 193 distinct payloads remained. Examples of payload content simulated during the attacks are shown in Table 6.

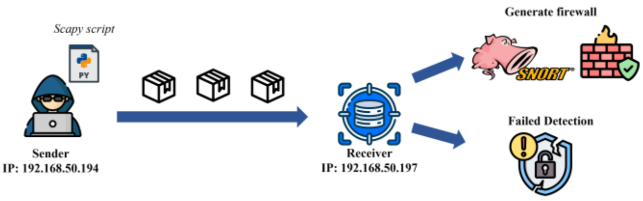


Figure 5. Testing and Verifying

TABLE 6. SOME EXAMPLE OF TESTING PAYLOAD

Subtechnique	Payload
T1059.004	$ nohup MipsLinuxTF &
T1059.004	echo "./syntem&">>/etc/rc.local
T1120.000	cat /proc/cpuinfo | grep name | wc -l

B. Performance Analysis

In this work, the honeypot packets were collected from various Internet Service Providers (ISPs). The honeypot-generated JSON files were processed by data pre-processing, K-means, and labeled with MITRE ATT&CK, LSH. Based on the similarity of payload contents, two clustering methods were employed to group similar payloads to generate more compliant snort rules. The distribution of clusters is presented in Table 7. It is evident from the table that the sixth cluster represents the largest portion of attack types collected during this period, comprising 63.45% of all attack clusters. Its Subtechnique corresponds to MITRE ATT&CK framework T1120.000: Peripheral Device Discovery, where attackers attempt to gather information about physical devices.

In this study, 193 different packets were used for experimentation, both before and after applying LSH. The main purpose is to evaluate whether the secondary clustering of snort rules improved the defensive rate of malicious packets. The experimental results are presented in Table 8. Following the experiments mentioned above, it can be found that the defensive rate of malicious packets increased from 91% to 93%, which shows that adding LSH as one of the clustering techniques improves the detection effect of snort abnormal packets.

TABLE 7. DETAILS OF SNORT RULE

Cluster ID	Packet number	Percentage(%)
6	24454	63.45919
19	4192	10.87842
18	3438	8.92176
30	1663	4.31556
24	1143	2.96613
37	928	2.40820
35	735	1.90736
3	579	1.50253
23	241	0.62541
25	237	0.61503
26	214	0.55534
20	128	0.33217
38	113	0.29324
0	78	0.20241
27	69	0.17906
28	62	0.16089
29	61	0.15830
1	32	0.08304
2	24	0.06228
11	20	0.05190
4	20	0.05190
5	18	0.04671
15	17	0.04412
17	12	0.03114
14	11	0.02855
33	9	0.02336
8	8	0.02076
31	4	0.01038
10	4	0.01038
41	3	0.00779
32	2	0.00519
16	2	0.00519
34	2	0.00519
12	2	0.00519
36	2	0.00519
22	2	0.00519
21	1	0.00260
13	1	0.00260
9	1	0.00260
7	1	0.00260
39	1	0.00260
40	1	0.00260

TABLE 8. RESULT ANALYSIS

	K-means	K-means + LSH
Numbers of cluster	27	42
Dst IP: port	192.168.50.197:100	192.168.50.197:80
Time Execution(s)	13.991	14.069
#Success Packet Detection	176	180
Packet Detection Rate	91%	93%

V. Conclusion

This study proposes a real-time system to prevent malicious payloads collected by honeypots in network traffic. While prior research primarily focused on analyzing packet headers to detect malicious packets, this study demonstrates the effectiveness of payload analysis for malicious packets. After pre-processing the raw packets, similar packets are clustered based on payload similarity using the K-means algorithm. MITRE ATT&CK labels are applied to evaluate the results of the initial clustering. To enhance snort firewall's defensive rate of packets, the system adopts a secondary clustering technique. The LSH method is utilized to re-aggregate clusters with higher similarity from the previous step, enabling snort to generate more effective snort rules for different payloads. The system, as mentioned, undergoes experiments simulating attacks where packets are continuously sent by the attacker. The defensive rate of packets reaches 93%. This study makes a valuable contribution to preventing malicious packets. Secondary clustering and payload analysis are feasible research directions for detecting malicious packets. Future research could explore alternative clustering methods, such as fuzzy string analysis, to improve the detection rate of malicious packets.

References

[1] Check Point Research Team, Check Point Research Reports a 38% Increase in 2022 Global Cyberattacks, [Online]. Available: https://blog.checkpoint.com/2023/01/05/38-increase-in-2022-global-cyberattacks/, 2023.

[2] IBM, Cost of a Data Breach Report 2023, [Online]. Available: https://www.ibm.com/reports/data-breach, 2023.

[3] Trend Micro, Preventing Risk - Trend Micro 2023 Cybersecurity Report,[Online]. Available:https://www.trendmicro.com/zh_tw/security-intelligence/threat-report/2023-midyear-security-roundup.html, 2023.

[4] I. Perona, I. Gurrutxaga, O. Arbelaitz, J.I. Martin, J. Muguerza, and J.M. Perez, "Service-independent payload analysis to improve intrusion detection in network traffic," *Proceedings of the 7th Australasian Data Mining Conference*, vol. 87, pp. 171–178, 2008.

[5] H. Liu, B. Lang, M. Liu, and H. Yan, "CNN and RNN based payload classification methods for attack detection," *Knowledge-Based Systems*, vol. 163, pp. 332–341, 2019.

[6] V. Bontupalli, C. Yakopcic, R. Hasan, and T. M. Taha, ''Efficient memristor-based architecture for intrusion detection and high-speed packet classification,'' *ACM J. Emerg. Technol. Comput. Syst.*, vol. 14, no. 4, pp. 1–27, Dec. 2018.

[7] K. Ozdem and M. A. Akcayol, "Locality Sensitive Hashing Based Clustering for Large Scale Documents," *Proceedings of 6th International Conference on Mathematics and Artificial Intelligence*, pp. 137-142, 2021.

[8] A. Bahri, K. E. Moutaoikil, and I. Badi. "A combined CNN and LSH for fast plant species classification," *Proceedings of the 4th International Conference on Big Data and Internet of Things*, pp. 1–6, 2019.

[9] M. Roesch, "Snort: Lightweight intrusion detection for networks," *In Lisa*, vol. 99, no. 1, pp. 229-238, 1999.

[10] F. Rafa, Z. Rahman, M. M. Mishu, M. Hasan, R. Rahman and D. Nandi, "Detecting Intrusion in Cloud using Snort: An Application towards Cyber-Security," *Proceedings of the 2nd International Conference on Computing Advancements*, pp. 199-206, 2022.

[11] Y. C. Lai, C. L. Yu, M. L. Liao, Y. S. Lin, Y. C. Chang, and J. L. Chen, "An Intelligence Defense System with SNORT Rules," *Proceedings of the 25th International Conference on Advanced Communication Technology*, pp. 249-254, 2023.

[12] M. L. Liao, C. L. Yu, Y. C. Lai, S. P. Chiu, and J. L. Chen, "An Intelligent Cyber Threat Classification System," *Proceedings of 25th International Conference on Advanced Communication Technology*, pp. 189-194, 2023.

[13] S. Abdulrezzak and F. A. Sabir, "Enhancing Intrusion Prevention in Snort System," *Proceedings of the 15th International Conference on Developments in eSystems Engineering*, pp. 88-93, 2023.

[14] P.J. Chuang, P.C. Chao, H.C. Lu, "Regular Expression Generator by using Genetic Algorithm," Available: https://github.com/maojui/Regex-Generator, 2020.

Pin-Shan Lin was born in Hsinchu, Taiwan, in 2000. She received her B.S. degree in 2022. She is pursuing an M.S. degree in electrical engineering from the National Taiwan University of Science and Technology, Taipei. Her main research interests include data analysis, machine learning, and cyber threat intelligence.

Yi-Cheng Lai was born in Taipei, Taiwan, in 1998. He received his B.S. degree in 2021. He is pursuing an M.S. degree in electrical engineering from the National Taiwan University of Science and Technology, Taipei. His main research interests include cyber security, vulnerability research, and defense.

Man-Ling Liao was born in Tainan, Taiwan, in 1998. She received her B.S. degree in 2021. She is pursuing an M.S. degree in electrical engineering from the National Taiwan University of Science and Technology, Taipei. Her main research interests include data analysis, machine learning, and cyber threat intelligence.

Shih-Ping Chiu was born in Taipei, Taiwan, in 1984. She received the B.S. degree in 2007. She is a research assistant in electrical engineering from the National Taiwan University of Science and Technology, Taipei. Her main research interests include data analysis and the Internet of Things (IoT)

JIANN-LIANG CHEN (Senior Member, IEEE) was born in Taiwan in December 1963. He received a Ph.D. in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1989. Since August 1997, he has been with the Department of Computer Science and Information Engineering, National Dong Hwa University, where he is a Professor and the Vice Dean of the Science and Engineering College. He joins the Department of Electrical Engineering, National Taiwan University of Scienceand Technology, as a Distinguished Professor. His current research interests include cellular mobility management, cyber security, personal communication systems, and the Internet of Things (IoT).

Session 2A: Wireless Communication 2

Chair: Prof. Ming An Chung , National Taipei University of Technology, Taiwan

1 Paper ID: 20240287, 79~83

Enhancing Inter-Satellite Data Relay in Dynamic Space Communication

Prof. REFIK CAGLAR KIZILIRMAK, Mr. Israel Ehile, Mr. Bekzat kabdrashev, Mr. Sergey Khvan,

Nazarbayev University. Kazakhstan

2 Paper ID: 20240473, 84~88

A Study on the evaluation of the ICT development indexes and some results

Ms. Narantuya Erkhembaatar, Prof. Otgonbayar Bataa,

SICT, MUST. Mongolia

3 Paper ID: 20240009, 89~93

Decoding Convolutional Hadamard Codes and Turbo Hadamard Codes using Recurrent Neural Networks

Dr. Sheng Jiang, Prof. Francis C.M. Lau,

The Hong Kong Polytechnic University. Hong Kong

4 Paper ID: 20240070, 94~99

QPSO-based Beamforming in Dual RIS-assisted Uplink Anti-jamming Communication System

Ms. Di Zhou, Prof. Zhiquan Bai, Ms. Jinqiu Zhao, Mr. Zeyu Liu, Prof. Dejie Ma, Prof. KyungSup Kwak,

Shandong University. China

5 Paper ID: 20240022, 100~103

A compact dual-band metamaterial absorber using square split rings for C-band and X-band sensors applications

Mr. Ramesh Amugothu, Prof. Vakula Damara,

NITW. India

Enhancing Inter-Satellite Data Relay in Dynamic Space Communication

Refik Caglar Kizilirmak, Israel Ehile Ehile, Bekzat Kabdrashev, Sergey Khvan

Dept. of Electrical and Computer Engineering, Nazarbayev University, Astana, Kazakhstan

***Abstract*—In this paper, we explore the design and operations of an inter-satellite data relay network, which includes Low Earth Orbit (LEO), Medium Earth Orbit (MEO), and Geostationary Orbit (GEO) satellites. Due to the different orbiting speeds and dynamic behaviors of these satellite constellations, we consider Delay-Tolerant Networking (DTN) between them and implement contact graph routing (CGR) at the layer-3 and Licklider Transmission Protocol (LTP) at layer-4. We base our analysis on a realistic space scenario where satellite locations and movements over time are extracted from a real-time satellite database. We demonstrate end-to-end latency and goodput of a file transfer through the satellite network under consideration.**

***Index Terms*—Delay Tolerant Networks (DTN), Contact Graph Routing (CGR), Licklider Transmission Protocol (LTP)**

I. Introduction

Traditional transport and network layer technologies are not suitable for dynamic space networks, especially for interplanetary networks (IPN) [1], prompting the search for novel architectures and protocols. These novel approaches must address various space network challenges, such as intermittent connectivity among nodes, vast distances, and significant delays. In 2007, an Internet Research Task Force (IRTF) working group proposed a delay-tolerant networking (DTN) architecture as the backbone of space networks, restructuring the layered architecture of traditional networks. They introduced the Bundle Protocol (BP) with store-and-forward features to manage intermittent connectivity and, at layer-4, replaced TCP with the Licklider Transmission Protocol (LTP) to reduce acknowledgment traffic overhead between nodes. Subsequently, the Consultative Committee for Space Data Systems (CCSDS) introduced IPN Contact Graph Routing (CGR) at layer-3 to handle dynamic topology and find the shortest path in space [2]. Recently, the standardization efforts for IPN shifted from IRTF to the Internet Engineering Task Force (IETF).

The fundamental components of DTN architecture in space (CGR, LTP, BP) have been the focus of extensive research. They have also received continuous support through the implementation of ION (Interplanetary overlay network) [3], developed by NASA-JPL. Many studies have explored these technologies individually rather than examining the DTN system as a cohesive whole. In studies conducted in [4] [5], the focus was on investigating BP and LTP for modeling and performance analysis of space networks, with limited attention to their interactions with CGR. Similarly, [6] [7] [8] delved into the study of CGR in isolation, with minimal consideration of its interactions with upper layers. There have also been sporadic efforts that comprehensively analyzed the interplay between the various multilayer technologies within DTN in space [9]. In this paper, we study and demonstrate end-to-end latency and goodput of a file transfer through a satellite network while considering multilayer interactions of DTNs.

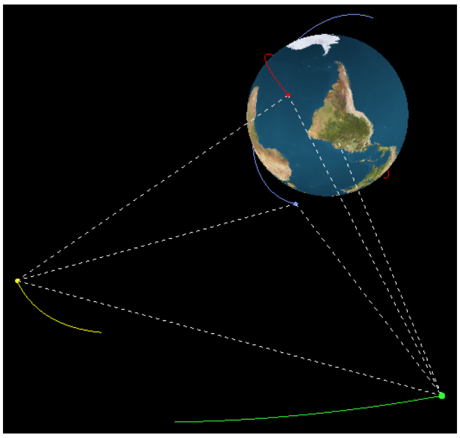


Fig. 1. Satellite constellation under consideration. Intermittent connectivity and dynamic topology can be seen at https://youtu.be/c7G9_8GdAX4

Fig. 1 illustrates the satellite constellation that we consider, consisting of two Low Earth Orbit (LEO) satellites, one Medium Earth Orbit (MEO) satellite, and one Geostationary Orbit (GEO) satellite. We obtained real satellite locations from http://celestrak.org/ and visualized them using a software tool that we developed. Please note that these satellites serve distinct operational objectives; however, we configured a DTN utilizing them for our research study. LEO satellites orbit the Earth, traversing various Earth regions to capture sensor data. MEO satellites play a vital role in data relay, receiving data from LEO satellites and forwarding it to GEO satellites. Positioned over specific Earth regions, GEO satellites gather data from LEO satellites and transmit it to ground stations. This configuration is beneficial in cases where data collection in a particular Earth region is distant from the ground station, and data needs to be relayed.

In this scenario, we employ DTN protocols to ensure reliable data transfer between the satellites which is a challenging task with traditional protocols due to the intermittent connectivity. As LEO satellites pass through the coverage region of MEO satellite, they establish connections and transmit the stored sensor data. MEO satellite implements store-and-forward relaying, buffering the received data until they come within the coverage area of GEO satellite. Finally, the GEO satellite forwards the data received from the MEO satellite to the ground station through its continuous communication link.

This paper is organized as follows. In Section II, we briefly provide discussion of use of DTNs in space communication. In Section III, we revisit LTP. In Section IV, we present our performance analysis results, namely for latency and goodput, for the scenario under consideration. In Sectoin V, we conclude our paper.

II. Delay Tolerant Networking for Deep Space Communication

The delay tolerant network research group (DTNRG) suggested a structure and framework design in response to the demand for interoperable communication in harsh and efficiency-defying conditions where constant point-to-point connectivity is not possible [10]. According to [11] DTN architecture is a viable technique for Interplanetary Internet (IPN) networks. To overcome the difficulty of constantly delivering messages in intermittent connections, the DTN architecture employs a store-and-forward strategy that comprises storing messages in the buffer of DTN routers for long periods of time. This architecture employs a new layer known as the ”bundle layer” to provide communications [12]. According to [9] the end-to-end path is divided into many DTN hops by adding the bundle protocol (BP) to endpoints and certain intermediate nodes on the route. The term ”bundles” refers to data packets transmitted at the bundle layer. Since bundles are devoid of any size restrictions, they have the potential to be substantially bigger than standard IP packets, mainly when they include vast amounts of data, such as an image. The bundle layer offers store-and-forward relay for end-to-end bundle transfer.

Instead of IP, which links heterogeneous networks by assigning an IP address to each node at the network layer, a DTN employs ”naming, layering, encapsulation, and permanent storage” [13]. The same concepts and mechanisms developed to address the delays and disruptions of interplanetary communications can be easily applied in certain connectivity regions with long signal propagation times, regular node obstruction, throughput flexibility, or constrained transmission range and information. Every communication network must be capable of effectively transmitting data from point to point. In the context of DTNs, conventional routing strategies should be able to manage network latency without significantly reducing information transfer efficiency. Many of these protocols may be classified based on the prospective contacts’ level of global awareness [14]. Furthermore, in [14], the main idea of DTN networking is to solve the design and routing ideas in order to provide effective interoperability with/among severe and performance-challenged devices. The fundamental DTN core concept allows network routers to send and store data to be forwarded later when the connection is restored.

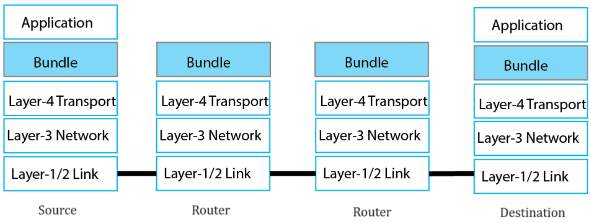


Fig. 2. DTN architecture.

III. Licklider Transmission Protocol (LTP)

LTP is suitable for DTNs’ transport layer operations. While it shares some similarities with TCP (Transmission Control Protocol), LTP differs from TCP in several key ways. Firstly, unlike TCP, which operates exclusively between the two endpoints, LTP runs on every node, including both routers and endpoints (see Fig. 2). The primary rationale behind this approach is to address the considerable distances and associated delays between the end nodes. In cases where a packet is lost during transit, implementing Automatic Repeat reQuest (ARQ) solely between the end nodes would result in prohibitively long recovery times. Therefore, LTP incorporates ARQ at every hop within the network. Secondly, LTP optimizes the number of acknowledgment packets. In contrast to TCP, where each individual segment is acknowledged separately, LTP groups bundles into LTP blocks. This aggregation reduces the acknowledgment overhead within the network.

In LTP, file transfer is accomplished by first fragmenting the file into a number of bundles, which are later divided into several LTP data blocks. Each data block is further divided into segments and transmitted to the receiver node. LTP allows for the option of reliable transmission, for example, by requiring acknowledgments. LTP segment size is set to ensure it fits into Ethernet MTU (1500 bytes) for later processing. Red blocks indicate reliable transmission, while green blocks indicate unreliable transmission. In this study, we assume that all the blocks are set as 100% red. At the end of the block transmission, the sender sends a checkpoint (CP) segment. When the receiver receives the CP segment, it checks for any loss of segments within the block. Upon receiving the CP, the receiver sends a report segment (RS) to the sender to indicate whether any segments were lost or not. When there is no loss, the receiver cumulatively acknowledges the sender with RS. On the other hand, when there is any loss, RS informs the sender about the lost data, and the sender retransmits any lost segments, followed by another CP at the end of its transmission. In LTP, both CP and RS segments are transmitted with a timer to ensure delivery to the other side. In case of a timeout, they are retransmitted.

IV. Performance Analysis

For the satellite network considered in Fig. 1 (also given in Fig. 3 for illustration), due to the different orbiting speeds and dynamic behaviors of these satellite constellations, we consider DTN protocols between them and implement CGR at the layer-3 and LTP at layer-4. We demonstrate end-to-end latency and goodput of a file transfer through this satellite network.

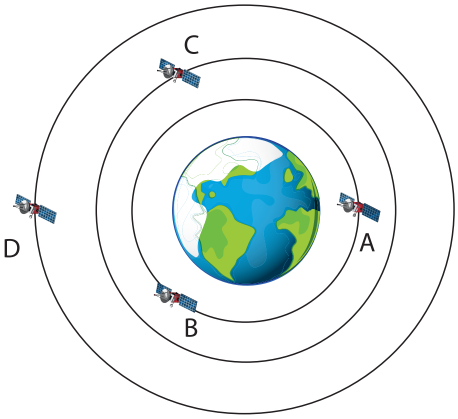


Fig. 3. 2D illustration of nodes A,B,C and D orbiting Earth. Nodes A and B are the LEO satellites, C is the MEO satellite, and D is the GEO satellite (not to the scale).

A. Contact Graph Routing (CGR)

Table I shows the contact plan we obtained for a duration of 2 hours, where nodes A and B are the LEO satellites, C is the MEO satellite, and D is the GEO satellite. In Table I, for example, the first row shows the first contact between nodes A and B; contact starts at 23.33 min and lasts for 20 mins. The average distance between A and B during this contact is 5500 km, which corresponds to the propagation delay of 0.018 sec.

The CGR relies on adapted Dijkstra's shortest path algorithm [15]. This algorithm computes a route from a source to a target node using a contact graph that is based on a contact plan. Each vertex in the contact graph represents connectivity period between two nodes. Two vertices are connected by a directed edge if the end of the earlier contact matches the start of the next contact. The computed route is a sequence of contacts from the source to the destination.

TABLE I. Contact Plan

Cont.	From	To	Start (min)	End (min)	avg. distance (km)
1	A	B	+23.33	+43.33	5500
2	A	B	+76.66	+93.33	3400
3	A	D	+0	+8.33	42000
4	A	D	+53.33	+105	39000
5	A	C	+0	+13	28000
6	A	C	+36.66	+90	26000
7	B	D	+0	+4	46500
8	B	C	+15	+98.33	25500
9	C	D	+0	+120	46000

For the contact plan in Table 1, we run CGR to find the shortest path from source node A to destination node D by implementing the algorithm described in [15]. The shortest path is found as through contacts #1,#8 and #9, A→B, B→C and C→D. In this route, the longest path in terms of propagation delay is contact #9 with 0.15 sec.

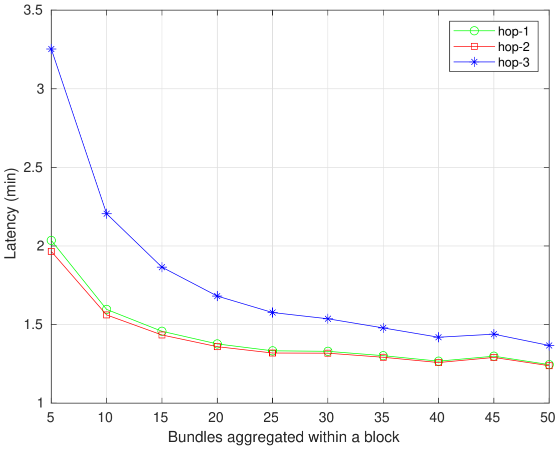


Fig. 4. End-to-end latency for each hop.

B. LTP goodput and delay performance

After finding the shortest path, we employ LTP and demonstrate a file transfer of 1 MB size through this path. In the analysis, we investigate the goodput and latency performance of LTP. We assume that each link has symmetrical data rates of 115200 bps. In the context of LTP, if the bundles are configured as 'red,' a series of checkpoint (CP), report segment (RS), and report acknowledgment (RA) exchanges occur after each block transfer between any two nodes along the communication path. The segment size is considered to be equal to the bundle size, which is set at 1000 bytes, as are the sizes of the CP, RA, and RS segments. To maintain generality, we assume that there occurs no errors in the links.

Fig. 4 shows the end-to-end delay for the file transfer for each hop individually. To simplify the discussion, no forwarding is demonstrated in Fig. 4. For the hop-1 (A→B) when 5 bundles are aggregated within an LTP block, the complete file transfer takes 2.04 min. For the same bundle configuration, hop-2 (B→C) and hop-3 (C→D) require 1.97 and 3.26 min, respectively. When the number of bundles aggregated within a block is increased, the acknowledgment overhead, including the number of CR, RS, and RA exchanges, is reduced, and the file transfer duration also decreases. Fig. 5 shows the goodput achieved by each hop individually. Hop-1 (A→B) achieves the highest goodput at around 13.5 KBps when 50 bundles are aggregated within a block. For the same configuration, hop-2 (B→C) and hop-3 (C→D) achieve 13.5 KBps and 12.2 KBps, respectively.

Next, we investigate the end-to-end file transfer latency, including forwarding. In LTP, forwarding typically occurs at the block level; in other words, the intermediate node can start forwarding after receiving a block and completing RS and RA exchanges. In Fig. 6, this strategy is indicated as incremental forwarding. The contact plan in Table 1 shows that the second and third hops (B→C and C→D) are already available when source node A starts its transmission. Therefore, incremental forwarding is feasible for this route, where nodes B and C forward the blocks as they receive without waiting for

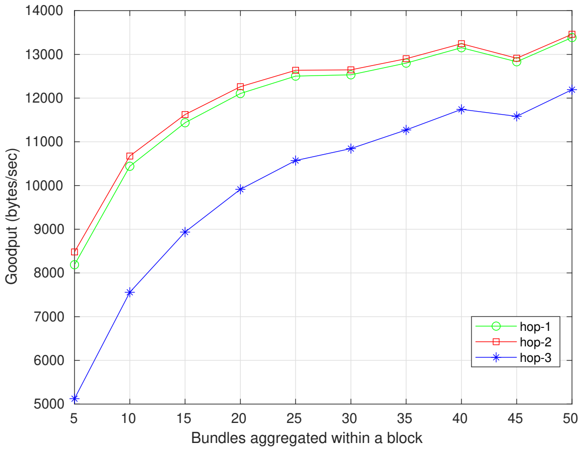


Fig. 5. Goodput (bytes/sec) for each hop.

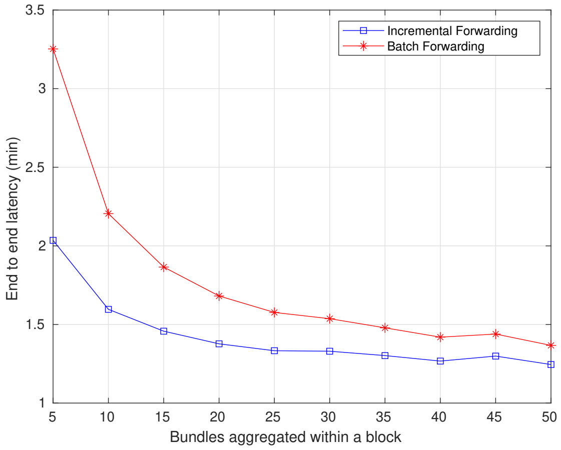


Fig. 6. End-to-end latency.

the entire file. In Fig. 6, we further compare incremental forwarding with batch forwarding, where a complete file is stored before forwarding it to the next hop. As more bundles are aggregated within a block, end-to-end delay converges to 1.2 and 1.4 mins. respectively for incremental and batch forwarding.

V. Conclusion

In this work, we investigated the expanding landscape of space networking by providing a comprehensive analysis of DTN technologies and their multi-layer interconnections. Our research exemplifies the practical significance of DTN protocols, focusing on their potential to provide efficient and reliable communication within the demanding space environment. In our work, we first generated a satellite network using real satellite locations and then built a contact plan based on their movements. We ran the CGR algorithm to find the shortest path for the given scenario and demonstrated a file transfer using LTP over this path. The results of this research have important significance for the development and improvement of space communication systems in the future, establishing possibilities for networks based on satellites and interplanetary space that are more effective. Our future work includes incorporating the impact of segment losses, using instantaneous distances between nodes, and building a visual simulator that shows all these interactions.

Acknowledgement

This work was supported by the Faculty Development Competitive Research Grant Program of Nazarbayev University (grant no. 20122022FD4125).

References

[1] R. Wang, X. Wu, T. Wang, X. Liu, and L. Zhou, “Tcp convergence layer-based operation of dtn for long-delay cislunar communications,” *IEEE systems journal*, vol. 4, no. 3, pp. 385–395, 2010.

[2] J. A. Fraire, P. Madoery, S. Burleigh, M. Feldmann, J. Finochietto, A. Charif, N. Zergainoh, R. Velazco, *et al.*, “Assessing contact graph routing performance and reliability in distributed satellite constellations,” *Journal of Computer Networks and Communications*, vol. 2017, 2017.

[3] S. Burleigh, “Interplanetary overlay network: An implementation of the dtn bundle protocol,” 2007.

[4] A. Sabbagh, R. Wang, K. Zhao, and D. Bian, “Bundle protocol over highly asymmetric deep-space channels,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2478–2489, 2017.

[5] A. Sabbagh, R. Wang, S. C. Burleigh, and K. Zhao, “Analytical framework for effect of link disruption on bundle protocol in deep-space communications,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 5, pp. 1086–1096, 2018.

[6] N. Bezirgiannidis, C. Caini, and V. Tsaoussidis, “Analysis of contact graph routing enhancements for dtn space communications,” *International Journal of Satellite Communications and Networking*, vol. 34, no. 5, pp. 695–709, 2016.

[7] E. Birrane, S. Burleigh, and N. Kasch, “Analysis of the contact graph routing algorithm: Bounding interplanetary paths,” *Acta Astronautica*, vol. 75, pp. 108–119, 2012.

[8] G. Araniti, N. Bezirgiannidis, E. Birrane, I. Bisio, S. Burleigh, C. Caini, M. Feldmann, M. Marchese, J. Segui, and K. Suzuki, “Contact graph routing in dtn space networks: overview, enhancements and performance,” *IEEE Communications Magazine*, vol. 53, no. 3, pp. 38–46, 2015.

[9] N. Alessi, C. Caini, T. de Cola, S. Martin, and J. P. Mayer, “Dtn performance analysis of multi-asset mars-earth communications,” *International Journal of Satellite Communications and Networking*, vol. 40, no. 1, pp. 11–26, 2022.

[10] S. Farrell, V. Cahill, D. Geraghty, I. Humphreys, and P. McDonald, “When tcp breaks: Delay-and disruption-tolerant networking,” *IEEE Internet Computing*, vol. 10, no. 4, pp. 72–78, 2006.

[11] S. Burleigh, A. Hooke, L. Torgerson, K. Fall, V. Cerf, B. Durst, K. Scott, and H. Weiss, “Delay-tolerant networking: an approach to interplanetary internet,” *IEEE Communications Magazine*, vol. 41, no. 6, pp. 128–136, 2003.

[12] K. Scott and S. Burleigh, “Bundle protocol specification,” tech. rep., 2007.

[13] C. J. Krupiarz, E. J. Birrane, B. W. Ballard, L. Benmohamed, A. A. Mick, K. A. Stambaugh, and E. W. Tunstel, “Enabling the interplanetary internet,” *Johns Hopkins APL Technical Digest*, vol. 30, no. 2, pp. 121–134, 2011.

[14] V. Cerf, S. Burleigh, A. Hooke, L. Torgerson, R. Durst, K. Scott, K. Fall, and H. Weiss, “Delay-tolerant networking architecture,” tech. rep., 2007.

[15] J. A. Fraire, O. De Jonckère, and S. C. Burleigh, “Routing in the space internet: A contact graph routing tutorial,” *Journal of Network and Computer Applications*, vol. 174, p. 102884, 2021.

Refik Caglar Kizilirmak (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 2004 and 2006, respectively, and the Ph.D. degree from Keio University, Yokohama, Japan, in 2010. He was with the Communications and Spectrum Management Research Center, Turkey, on several telecommunication and defense industry projects. He is currently with the Department of Electrical and Computer Engineering, Nazarbayev University, Astana, Kazakhstan. He has contributed to the technical requirements document of IEEE 802.15.7r1 standardization, which will enable visible light communication. He has authored several articles in the field of wireless communications and has filed three patent applications with the patent offices of USA and Japan.

Israel Ehile Ehile received the B.Eng. in electrical and electronics engineering from the University of Agriculture, Makurdi, Nigeria in 2018. He was a Research Assistant with the Department of Electrical and Computer Engineering, Nazarbayev University, Astana, Kazakhstan in 2023. He is currently a Master student in Electrical and Computer Engineering at Nazarbayev University, Astana, Kazakhstan. His research interests include Communication in Inter-Satellite, Routing in Deep Space Network, Computer Networks and Cloud Security.

Bekzat Kabdrashev was born in Akmola region, Kazakhstan in 2001. Kabdrashev graduated from National School of Physics and Mathematics in Astana, Kazakhstan in 2019. He is a senior computer science student at Nazarbayev University in Astana, Kazakhstan. Alongside with his studies at Nazarbayev University, he is researching interplanetary networks as a Research Assistant of Professor Refik Kizilirmak. His research interests include computer networks and programming language design.

Sergey Khvan is a graduate student in Computer Engineering at the University of Padova, Italy. The author completed his bachelor’s studies in Electrical and Computer Engineering at Nazarbayev University, Astana in 2023. Previously worked as a Research Engineer at “Robotics and Artificial Intelligence” LLP and Research Assistant at Nazarbayev University. Previous projects include the spheres of the Internet of Things(IoT), Blockchain, and Robotics. Current research interests include Machine/Deep Learning Applications, Computer Vision, and Natural Language Processing.

A Study on the evaluation of the ICT development indexes and some results

Narantuya Erkhembaatar*, Otgonbayar Bataa**

* Department of Communications Engineering, Mongolian University of Science and Technology, Ulaanbaatar, Mongolia

** Department of Communications Engineering, Mongolian University of Science and Technology, Ulaanbaatar, Mongolia

narantuya@must.edu.mn, otgonbayar_b@must.edu.mn

***Abstract*— The ICT Development Index (IDI) functions as an established tool for assessing the digital divide and facilitating comparisons of ICT performance within and between countries. Information entropy, representing the level of uncertainty in a random variable, can be applied across various fields, including information and communication technology (ICT). When designing data analysis using information entropy, it is essential to observe, evaluate, and utilize metrics derived from this method. The proposed methodology aims to allocate weights to the indicators within the ICT Development Index for country ranking. To assess the efficacy of this methodology, we explored its potential applications in evaluating indexes. Our model incorporates an innovative approach that combines the entropy weight coefficient method with the correlation coefficient weighting method. We present the evaluation results of the integrated calculation method in Mongolia.**

***Keywords*— ICT indicators, entropy weight, IDI, correlation coefficient weight, ICT development index**

I. Introduction

Since 2009, the International Telecommunications Union (ITU) has annually released the IDI. This index serves as a valuable resource for policymakers and researchers in evaluating the efficacy of policies and initiatives geared toward promoting ICT development. It enables the benchmarking of progress and the exchange of best practices among countries [1][2]. Through quantitative metrics and benchmarks, the IDI offers an objective assessment of international performance. Decision-makers find it to be an indispensable tool for measuring ICT growth and progress globally [2][3]. Additionally, the IDI acknowledges the convergence of technology and the emergence of new technologies, making it effective in assessing the impact of ICT on society and the economy [4-6]. ITU's Council noted in September 2023 during the annual meetings, that the IDI methodology, the Version 3 document, was finalized, now called Version 3.1 with 10 indicators, and has been submitted for approval by Member States [7].

II. Methodology for Index Evaluation

Within this section, we will introduce the formulation of an integrated evaluation model for the IDI. The subsequent subsection will furnish an outline of several fundamental theories, followed by an in-depth explanation of the proposed integrated algorithm.

A. Entropy weighting method

The concept of information entropy measures the amount of information that can be obtained from a source of random data. It was introduced by Shannon [8-16]. The entropy weighing method is a process used to calculate index values in analyses conducted under objective conditions. It enables the evaluation of intent, degree of order, and efficiency by employing entropy estimation of the information. The entropy weighting approach contributes to enhancing the objectivity of rank lists [13][14][16-27].

B. Correlation coefficient weighting method

To examine the correlation among indicators of ICT development, the correlation coefficient (CC) formula was applied. Correlation analysis, a quantitative analytical tool, is employed to assess the extent of the relationship between independent and dependent variables [13][24][28].

C. Proposed weighting method

This subsection aims to provide a detailed description of the proposed integration algorithm.

The following are the specific steps:

1) Assuming that there are m additional elements that require measurement, and we have an algorithm for measuring n objects, we must create an evaluation matrix for the evaluation model.

$$X_{ij} = \left(x_{ij}\right)_{mxn} \tag{1}$$

The notation x_{ij} represents the elements of a matrix, while the value of the i^{th} indicator of the j^{th} sample is denoted as x_{ij}.

2) The matrix needs to be normalized using the following operations:

$$d_{ij} = \frac{x_{ij}}{\max x_{ij}} \tag{2}$$

The resulting normalization matrix is obtained as follows:

$$D_{ij} = \left(d_{ij}\right)_{mxn} \tag{3}$$

3) The related weight of x_{ij} is:

$$P_{ij} = \left(p_{ij}\right)_{mxn}$$
$$p_{ij} = {d_{ij}} \Big/ {\textstyle\sum_{i=1}^{m} d_{ij}} \qquad (4)$$

4) Equation (5) represents Shannon's information entropy for the i^{th} indicator of the matrix, where m is the number of indicators and n is the number of objects.

$$E_i = -\frac{1}{\ln(n)} \sum_{j=1}^{n} p_{ij} \ln p_{ij} \qquad (5)$$

To standardize the value of E_i and ensure that $0 < E_i < 1$.

5) The equation for representing the entropy weight is given below:

$$W_{ei} = \frac{1 - E_i}{m - \sum_{i=1}^{m} E_i} \qquad (6)$$

6) This results in a symmetrical matrix of size mxm with a generic element denoted by r_{ik} in a matrix of R. To determine r_{ik}, follow the steps below:

$$r_{ik} = \frac{n \sum x_{ij} y_{ik} - \left(\sum x_{ij}\right)\left(\sum y_{ik}\right)}{\sqrt{n \sum x_{ij}^2 - \left(\sum x_{ij}\right)^2} \cdot \sqrt{n \sum y_{ik}^2 - \left(\sum y_{ik}\right)^2}} \qquad (7)$$
$$i, k = 1, 2, \ldots, m$$

7) The symmetric matrix must be used to determine the correlation coefficient.

$$R = \left(r_{ik}\right)_{mxm}$$
$$i, k = 1, 2, \ldots, m \qquad (8)$$

We utilize the sum function to estimate the level of disagreement caused by the index function f_i in comparison to the other indexes. This implies that the alternatives with higher discordant scores on criteria f_i and f_k should receive a lower r_{ik} rating.

8) The sum vector can be normalized to obtain the weight of the CC:

$$W_{cci} = \frac{\sum_{k=1}^{m} (1 - r_{ik})}{\sum_{i=1}^{m} \sum_{k=1}^{m} (1 - r_{ik})} \qquad (9)$$
$$i = 1, 2, \ldots, m$$

CC weight W_{cci} can be obtained with Equation (9).

9) To find the weight W_{eci} by utilizing the output of Equation (6) and (9) and performing the calculation indicated in Equation (10).

$$W_{eci} = \frac{W_{ei} \cdot W_{cci}}{\sum_{i=1}^{m} W_{ei} \cdot W_{cci}} \qquad (10)$$

The proposed weighting method utilizes the CC weight method grounded in the principles of entropy.

III. Assessment of Indicators in the Case of ICT Development Index

A. Selected indicators of IDI

The IDI was established with the purpose of gauging the advancement of the information and communication technology sector. Published by the ITU, this composite indicator was active from 2009 to 2017 but was discontinued in 2018 due to challenges related to the availability and quality of data [7]. The IDI was established with the purpose of gauging the advancement of the ICT sector. Published by the ITU, this composite indicator was active from 2009 to 2017 but was discontinued in 2018 due to challenges related to the availability and quality of data [7]. Table 1 shows indicators of IDI by methodology of the ICT Development Index Version 3.1, ITU.

TABLE 1. Indicators of IDI [7]

Universal connectivity	
X1	Proportion of individuals who used the Internet (from any location) in the last 3 months
X2	Proportion of households with Internet access at home
X3	Active mobile-broadband subscriptions per 100 inhabitants
Meaningful connectivity – Infrastructure	
X4	Percentage of the population covered by at least a 3G mobile network
X5	Percentage of the population covered by at least a 4G/LTE mobile network
X6	Mobile broadband Internet traffic per mobile broadband subscription (GB)
X7	Fixed broadband Internet traffic per fixed broadband subscription (GB)
Meaningful connectivity – Affordability	
X8	Mobile data and voice high-consumption basket (as a % of GNI per capita)
X9	Fixed-broadband Internet basket price (as % of GNI per capita)
Meaningful connectivity – Device	
X10	Percentage of individuals owning a mobile phone

Indicators (X1) through (X3) denote universal connectivity, while (X4) to (X10) signify meaningful connectivity. Specifically, (X4) to (X7) focus on infrastructure, (X8) to (X9) assess affordability, and (X10) gauges device ownership.

B. Correlation analysis

Correlation analysis stands as a crucial statistical tool in the development of composite indicators. By elucidating the statistical relationships among the indicators under consideration for inclusion, it offers an initial insight into the robustness of an index and potential issues related to internal consistency. Correlation analysis can also provide insights into the weighting process and the arrangement of indicators. Table 2 displays the correlation coefficients related to the indicators of capital city and aimags in Mongolia. Mongolia is divided into 21 aimags or provinces. In general, the correlation coefficients exhibit the anticipated signs within the chosen set of indicators. The anticipated weak correlation between the two affordability indicators and the remaining indicators is also in line with expectations, given that these indicators are measured in opposite directions; lower prices correspond to a better situation. The indicators within the universal connectivity category exhibit positive and moderate to strong correlations with each other. The survey-based indicators demonstrate the highest degree of similarity, while the somewhat weaker coefficients between fixed and mobile broadband penetration indicators suggest that these two technologies complement each other. All correlations are stronger than 0.183 in both pillars.

TABLE 2. PEARSON CORRELATION COEFFICIENT MATRIX OF SELECTED INDICATORS OF IDI

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1	1	0.614	0.435	0.391	0.438	0.315	0.376	0.183	0.217	0.914
X2	0.614	1	0.672	0.716	0.688	0.315	0.376	0.391	0.509	0.525
X3	0.435	0.672	1	0.72	0.789	0.506	0.546	0.528	0.538	0.358
X4	0.391	0.716	0.72	1	0.914	0.455	0.536	0.525	0.394	0.354
X5	0.438	0.688	0.789	0.914	1	0.462	0.537	0.523	0.387	0.359
X6	0.315	0.641	0.506	0.455	0.462	1	0.638	0.388	0.54	0.201
X7	0.376	0.661	0.546	0.536	0.537	0.638	1	0.41	0.6	0.207
X8	0.183	0.391	0.528	0.525	0.523	0.388	0.41	1	0.41	0.184
X9	0.217	0.509	0.538	0.394	0.387	0.54	0.6	0.41	1	0.437
X10	0.914	0.525	0.358	0.354	0.359	0.201	0.207	0.184	0.437	1

Likewise, the moderate correlation observed between the two survey-based measures and the penetration measures based on administrative data indicates complementarities between these two approaches. The strong positive correlation observed between the indicators measuring mobile broadband coverage with at least 3G and 4G/LTE technologies implies that these two indicators can be consolidated into a singular measure within the meaningful connectivity infrastructure category.

C. Analysis of selected indicators of IDI

Implementing the suggested weighting method results in indicators X9 and X10 emerging with the highest rank among the IDI indicators, signifying their development in ICT in the country. Conversely, indicators X5 and X7 attain the lowest rank.

TABLE 3. THE WEIGTHS AND RANK OF THE IDI INDICATORS

indicators	*We*	rank	*Wcc*	rank	*Wec*	rank
X1	0.0103	9	0.1104	3	0.0115	9
X2	0.0411	8	0.0905	7	0.0493	8
X3	0.1388	3	0.0843	9	0.1280	2
X4	0.1466	2	0.0862	8	0.1260	3
X5	0.1363	4	0.0842	10	0.1198	4
X6	0.1363	5	0.1047	5	0.1171	5
X7	0.0003	10	0.0969	6	0.0003	10
X8	0.0642	7	0.1178	2	0.0634	7
X9	0.2620	1	0.1072	4	0.3145	1
X10	0.0642	2	0.1178	1	0.0701	6

Table 3 and Figure 1 show the weights and rank of the IDI indicators.

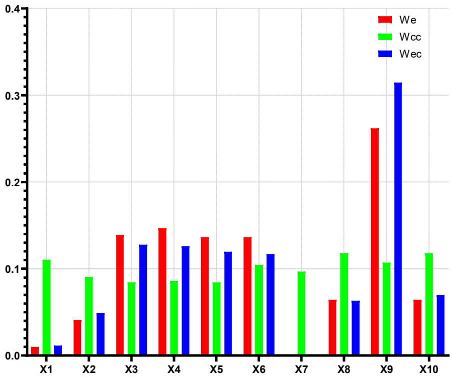


Figure 1. The weigths of the IDI indicators

It appears that weights between indicators calculated by equations (6), (9), and (10) pillar scores in Figure 1, are generally favorable, with values ranging from 0.0003-0.262 for the We, from 0.0842-0.1072 for the Wcc, and from 0.0003-0.3145 for the Wec. This suggests effective summarization of information from indicators into pillars.

D. Analysis of the IDI indicators of aimags

The outcomes of the entropy weight method reveal that Umnugovi aimag (0.0554) in the central region and Ulaanbaatar city (0.0421) attain the highest rank in IDI. Conversely, Bayan-Ulgii aimag (0.0413) in the western region and Bulgan aimag (0.0421) in the central region secure the lowest ranks.

TABLE 4. THE WEIGHTS AND RANK OF CAPITAL CITY AND AIMAGS

No	Aimag	*We*	rank	*Wcc*	rank	*Wec*	rank
1	Bayan-Ulgii	0.0413	22	0.0358	21	0.0326	21
2	Govi-Altai	0.0447	12	0.0392	17	0.0387	16
3	Zavkhan	0.0445	13	0.0415	13	0.0407	13
4	Uvs	0.0435	18	0.0398	16	0.0382	17
5	Khovd	0.0443	15	0.0419	12	0.0409	12
6	Arkhangai	0.0434	20	0.0227	22	0.0218	22
7	Bayankhongor	0.0434	19	0.0372	19	0.0356	19
8	Bulgan	0.0421	21	0.0360	20	0.0335	20
9	Orkhon	0.0487	3	0.0564	3	0.0606	3
10	Uvurkhangai	0.0440	17	0.0388	18	0.0377	18
11	Khuvsgul	0.0443	16	0.0423	11	0.0413	11
12	Govisumber	0.0470	4	0.0563	4	0.0584	4
13	Darkhan-Uul	0.0470	5	0.0560	5	0.0580	5
14	Dornogovi	0.0465	6	0.0477	7	0.0489	7
15	Dundgovi	0.0443	14	0.0406	14	0.0397	15
16	Umnugovi	0.0554	1	0.0596	2	0.0728	1
17	Selenge	0.0454	9	0.0454	10	0.0454	9
18	Tuv	0.0455	8	0.0470	8	0.0471	8
19	Dornod	0.0463	7	0.0530	6	0.0541	6
20	Sukhbaatar	0.0449	11	0.0404	15	0.0400	14
21	Khentii	0.0451	10	0.0455	9	0.0453	10
22	Ulaanbaatar	0.0488	2	0.0638	1	0.0687	2

The normalized scores and rank of the IDI indicators of aimags are present in Table 4 and Figure 2.

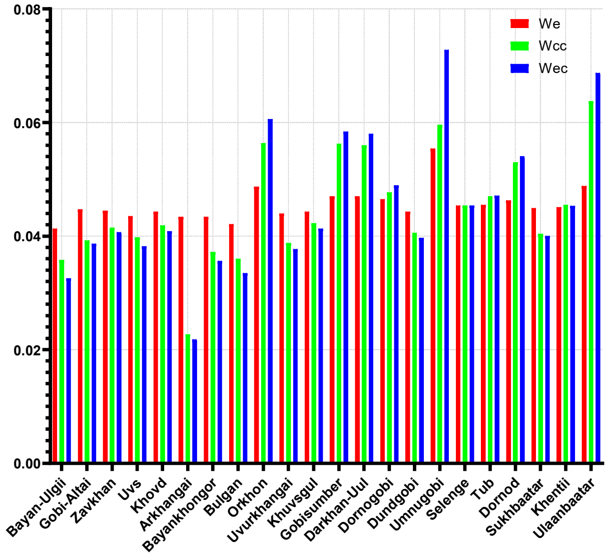


Figure 2. The weights of capital city and aimags

With the CC weight method, the results emphasize Umnugovi aimag (0.0728) as the highest in ICT development in the country, while Arkhangai aimag (0.0227) in the Khangai region ranks the lowest. Applying the proposed weighting method, Umnugovi aimag (0.0728) and Ulaanbaatar city (0.0687) emerge as the most developed in ICT in the country. Conversely, Arkhangai aimag (0.0413) in the Khangai region is identified as the least developed in terms of ICT development.

IV. Conclusions

The IDI indicators of aimags were analyzed using the CC weight method, entropy weight method, and the proposed method equation. The present study has assessed the levels of ICT development across capital city and 21 aimags by employing a composite result derived from ten selected indicators. The weight method has been utilized to scrutinize the IDI indicators of the aimags and ascertain the rank of regions within the study domain. Umnugovi aimag (0.078) in the central region attains the highest rank, while Ulaanbaatar city (0.0687) region secures the second rank for IDI in Mongolia. Nevertheless, Arkhangai (0.0218), Bayan-Ulgii (0.0326), and Bulgan (0.0335) aimags are situated at the lowest rank for IDI in Mongolia.

References

[1] Measuring the Information Society Report 2016, 2009, ITU International Telecommunication Union Place des Nations CH-1211 Geneva Switzerland, ISBN 978-92-61-21431-9, 2016.

[2] Organization for Economic Cooperation and Development, Guide to measuring the information society. 2011, OECD. http://www.oecd.org

[3] Organization for Economic Cooperation and Development, "Guide to measuring the information society", Partnership on Measuring ICT for Development, 2011

[4] Measuring the Information Society Report 2011, 2011 ITU International Telecommunication Union Place des Nations CH-1211 Geneva Switzerland, ISBN 978-92-61-21431-9, 2011.

[5] Core list of ICT indicators, March 2016 version, Partnership on Measuring ICT for Development, 2016

[6] Regional outcomes from IDI 2017, Measuring the Information Society Report 2017, Volume 1, ITU International Telecommunication Union Place des Nations CH-1211 Geneva Switzerland, ISBN 978-92-61-24521-4, 2017.

[7] Methodology of the ICT Development Index: Version 3.1, ITU October 2023.

[8] Zhao, H.,Yao, L., Mei, G., Liu, T., Ning, Y. "A Fuzzy comprehensive evaluation method based on AHP and Entropy for landslide susceptibility map", *Entropy*, 19(8), 396, 2017 https://doi.org/10.3390/e19080396

[9] Shannon C. and Weaver W., "The mathematical theory of communication", University of Illinois Press, 1949

[10] Erkhembaatar Narantuya, "Study on development of information potential", *International forum on strategic technology (IFOST-2013)*, Mongolia, Volume 2, 2013, 380-383, https://doi.org/10.1109/IFOST.2013.6616918

[11] Jose, F., G. Salvatore, and E. Matthias. "Multiple criteria decision analysis: State of the art surveys". *International Series in Operations Research & Management Science*, Volume 78, Springer, 2005 https://doi.org/10.1007/978-1-4939-3094-4

[12] Erkhembaatar Narantuya, "Defining information entropy and potential in the communication network", *International forum on strategic technology (IFOST-2018)*, China, 2018, 89-94

[13] Arindam Sutradhar, Pritirekha Daspattanayak. "A regional model for the variability of Agricultural development: evidence from a drought prone region of Rarh Bengal, Eastern India", *Modeling Earth Systems and Environment*, Springer Nature journal, 2023, https://doi.org/10.1007/s40808-023-01721-6

[14] Erkhembaatar Narantuya, Bataa Otgonbayar, "Entropy weight method for evaluating indicators of ICT development index", *International journal of current advanced research*, Volume 9, Issue 12 (C), 23500-23505, December 2020 https://doi.org/10.24327/ijcar.2020.23504.4654

[15] Erkhembaatar Narantuya, "Evaluation of Indicators for ICT Development Index using an Integrated Entropy Weighting Method". *ICT focus*, Volume 2, 1-13, 2023, https://doi.org/10.58873/sict.v2i1.43

[16] Erkhembaatar Narantuya, Bataa Otgonbayar, "Determining ICT indicators using entropy theory", *International conference on advanced communications technology (ICACT-2019)*, Korea, 2019, 217-222, https://doi.org/10.23919/ICACT.2019.8702010

[17] Liping Wang, Zhongyi Qu, Wei Yang, "Coupled urbanisation and ecological protection along the Yellow river basin in the context of dual carbon", *Sustainability*, MDPI, 15(7):5728, 4-16, 2023, https://doi.org/10.3390/su15075728

[18] Tang, C. M., A. Y. T. Leung, and K. C. Lam. "Entropy application to improve construction finance decisions", *Journal of construction engineering and management*, 132(10):1099–111, 2006.

[19] Sun, L., Liu, Y., Zhang, B., Shang, Y., Yuan, H., Ma, Z, "An integrated decision-making model for transformer condition assessment using game theory and modified evidence combination extended by D numbers", *Energies*, 9, 697, 2016, https://doi.org/10.3390/en9090697

[20] Kang, H.Y., Hung, M.C., Pearn, W.L., Lee, A.H.I., Kang, M.S. "An integrated multi-criteria decision making model for evaluating wind farm performance", *Energies*, 4, 2002–2026, 2011, https://doi.org/10.3390/en4112002

[21] Lee, A.H.I., Lin, C.Y., Kang, H.Y., Wen, H.L. "An integrated performance evaluation model for the photovoltaics industry", *Energies*, 5, 1271–1291, 2012, https://doi.org/10.1016/j.solener.2022.12.039

[22] Zeng, F.; Cheng, X.; Guo, J.; Tao, L.; Chen, Z. "Hybridising human judgment, AHP, grey theory, and fuzzy expert systems for candidate well selection in fractured reservoirs", *Energies*, 10, 447, 2017, https://doi.org/10.3390/en10040447

[23] Guili, Y., Jianhua, Z., Tianhong, W., Juan, D. "Comprehensive energy-saving evaluation of thermal power plants based on TOPSIS gray relational projection and the weight sensitivity analysis". *Journal of Chinese Society of Power Engineering*, 35, 404–411, 2015

[24] Wang, Y.M., Luo, Y. "Integration of correlations with standard deviations for determining attribute weights in multiple attribute decision making", *Mathematical and Computer Modelling*, 51, 1–12, 2010

https://doi.org/10.1016/j.mcm.2009.07.016

[25] Diakoulaki, D., Mavrotas, G., Papayannakis, L. "Determining objective weights in multiple criteria problems: The critic method", *Computer and Operations Research* 22(7):763–770, 1995, https://doi.org/10.1016/0305-0548(94)00059-H

[26] Lo, T.P., Guo, S.J. "Effective weighting model based on the maximum deviation with uncertain information", *Expert Systems with Applications*, 37(12): 8445–8449, 2010, https://doi.org/10.1016/j.eswa.2010.05.034

[27] Gan, Y., Duan, Q., Gong, W., Tong, C., Sun, Y., Chu, W., Ye, A., Miao, C., Di, Z. "A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model", *Environmental Modelling & Software*, 51:269–285, 2014, https://doi.org/10.1016/j.envsoft.2013.09.031

[28] Babamoradi, H., Berg, F.V.D., Rinnan, "A. Bootstrap based confidence limits in principal component analysis-A case study", *Chemometrics & Intelligent Laboratory Systems*, 120:97–105, 2013, https://doi.org/10.1016/j.chemolab.2012.10.007

Narantuya Erkhembaatar received a bachelor's degree in telecommunications engineering from the Polytechnic Institute of Mongolia in 1989, and in 2002 she earned a master's degree in technology from Andhra University in India. Her bachelor's thesis focused on investigating electronic exchange in the telecommunication network of Ulaanbaatar city, while her master's thesis involved the development of CBT for satellite communication. Her doctor degree thesis was centered on researching the determination of ICT development indicators based on entropy.

Otgonbayar Bataa, in 1978 graduated from the Polytechnic Institute of Mongolia majoring in Radio communication engineer. Bachelor degree thesis: Feasibility study of improving the efficiency of discrete information system. Master degree thesis (M.Sc) in 1995: Some issues of speech synthesis. PhD degree thesis in 1996: Study of Mongolian speech synthesis and applying it in telecommunication technics, in 2003, post Ph.D program thesis: Optimal version of OFDM system frequency and time-distortion. Professor. Consulting engineer of Mongolia. Research topic: Broadband, high speed integrated services technologies (WiMAX, WiBro, Mobile IPTV, 4GLTE, 5GNR etc).

Decoding Convolutional Hadamard Codes and Turbo Hadamard Codes using Recurrent Neural Networks

Sheng Jiang and Francis C. M. Lau
Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University,
Hong Kong SAR, China
Emails: s.jiang@polyu.edu.hk, francis-cm.lau@polyu.edu.hk

***Abstract*—In this paper, a Recurrent Neural Network (RNN) based decoder is proposed for the decoding of convolutional Hadamard codes (CHC) and Turbo Hadamard Codes (THC). Moreover, a long short-term memory (LSTM) network is adopted to realize the RNN decoder, forming the LSTM-CHC decoder and LSTM-THC decoder. Also, the proposed LSTM-THC decoder consists of several serial-concatenated LSTM-CHC decoders, which are pre-trained separately. The end-to-end LSTM-THC decoder is then trained based on the pre-trained weights. Simulations are performed on the LSTM-CHC/LSTM-THC decoders and their error performances are compared with those of the conventional decoders.**

***Index Terms*—convolutional Hadamard code, turbo Hadamard code, Recurrent Neural Networks.**

I. Introduction

Deep learning (DL) has been providing new approaches to decoding channel codes. In [1] and [2], Dense Neural Network (DNN) is used to mimic the Belief Propagation (BP) decoders of Bose-Chaudhuri-Hocquenghem (BCH) codes and High-Density Parity-Check (HDPC) codes. Neural BP decoder has also been proposed to decode polar codes in [3] and [4]. On the other hand, Recurrent Neural Network (RNN) is used to mimic the Bahl-Cocke-Jelinek-Raviv (BCJR) decoders of convolutional codes and turbo codes in [5]. However, no deep learning decoder has been investigated to decode channel codes that approaches the ultimate Shannon limit. Turbo Hadamard codes (THC) [6], concatenated zigzag Hadamard codes (CZHC) [7] and Low-Density Parity-Check Hadamard codes (LHC) [8] are a group of ultimate-Shannon-limit approaching codes formed by concatenating Hadamard codes with other codes. In this paper, we propose to decode convolutional Hadamard code (CHC) and THC using RNN. The method could also be applied to CZHC as the code structure and decoding algorithm of CZHC and THC are similar.

We organize the paper as follows. Sect. II reviews the encoder/decoder of CHC and THC. Sect III describes the structures of our proposed RNN decoder for CHC/THC. Sect. IV shows the details of the training of the RNN decoders and the simulated error performance. Sect. V provides some concluding remarks.

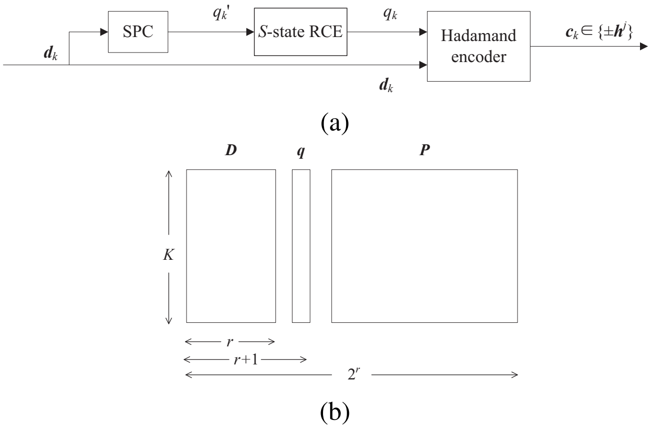


Fig. 1. Convolutional-Hadamard code. (a) Encoder block diagram and (b) code structure. SPC: single-parity check; RCE: recursive convolutional encoder.

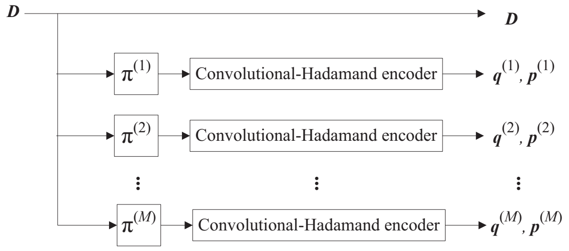


Fig. 2. Structure of turbo-Hadamard code.

II. Review of Turbo Hadamard Codes

An order r Hadamard matrix is recursively constructed from an order $r-1$ Hadamard matrix, i.e.,

$$\boldsymbol{H}_r = \begin{bmatrix} +\boldsymbol{H}_{r-1} & +\boldsymbol{H}_{r-1} \\ +\boldsymbol{H}_{r-1} & -\boldsymbol{H}_{r-1} \end{bmatrix} \quad (1)$$

with $\boldsymbol{H}_0 = [+1]$. The codeword set of an order-r Hadamard code is formed by the columns of $\pm\boldsymbol{H}_r$. In an order-r Hadamard code, each codeword has a length of 2^r. Moreover, the r bits with indices $\{0, 1, 2, 4, 8, \ldots, 2^{r-1}\}$ are denoted as the information bits while the remaining bits are denoted as the parity-check bits.

Fig. 1 shows the encoder block diagram and code structure of a CHC [9]. For an information block $\boldsymbol{D}$ of size $K \times r$,

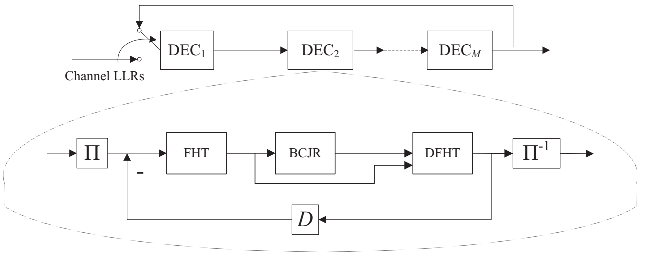


Fig. 3. Block diagram of a turbo Hadamard decoder that consists of M CHC decoders (sub-decoders). Each CHC decoder further consists of an interleaver, a FHT block, a BCJR block, a DFHT block and a de-interleaver.

we apply an 2-state rate-1/2 recursive-convolutional-encoder (RCE) to generate $\boldsymbol{q}$ of size $K \times 1$ and an order-r Hadamard encoder to generate the Hadamard parity-bits $\boldsymbol{P}$ of size $K \times (2^r - r - 1)$. Fig. 2 shows the encoder structure of a THC [6], [9], which is the combination of a number of M CHC derived from the same but interleaved message bits. In particular, $\pi^{(i)}$ represents the i-th interleaver; and $\boldsymbol{q}^{(i)}$ and $\boldsymbol{p}^{(i)}$, respectively, represent the parity bits generated by the RCE and Hadamard encoders in the i-th CHC encoder.

The decoding of THC follows a similar principle as turbo code. Referring to Fig. 3, a THC decoder with M component codes consists of M serially connected CHC decoders (sub-decoders), where the last sub-decoder is connected to the first sub-decoder. Extrinsic information is exchanged between CHC decoders to enhance the decoder performance. Each of the CHC decoders consists of three blocks: the fast-Hadamard-transform (FHT) block, the Bahl-Cocke-Jelinek-Raviv (BCJR) decoder and the dual-fast-Hadamard-transform (DFHT) block. The CHC decoders compute the *a-posteriori-probability* (APP) likelihood ratio (LLR) and the extrinsic information of the information bits, and then send them to the next CHC decoder. One iteration is completed when the last CHC decoder outputs the LLRs and extrinsic information. Here we denote the number of iterations by I.

III. Proposed RNN decoder for CHC/THC

The most complex part of the conventional THC decoder is the BCJR decoder. As the BCJR decoder adopts the forward-backward computational algorithm, it introduces a high latency to the decoding process. On the other hand, the forward-backward computational algorithm fits into the natural principle of RNN. In this paper, we investigate the use of a special type of RNN, namely long short-term memory (LSTM), to realize the decoder. This is because LSTM deals with the vanishing gradient problem [10] much better than the traditional RNN. As its name suggests, LSTM not only carries the "short-term memories" of information bits at its decoding stage, but also the "long-term memories" of information bits stored from the previous decoding stages. Moreover, a bi-directional LSTM can even carry the "long-term memories" of the next decoding stages from the backward layers. In other words, the LSTM networks can perfectly imitate the BCJR decoding process. We first investigate the LSTM decoder for CHC. Based on the LSTM-CHC decoders, we construct an end-to-end LSTM-THC decoder.

A. CHC decoder based on bi-directional LSTM

In the CHC decoder (sub-decoder) shown in Fig. 3, the FHT block prepares the trellis probabilities from the input *a-priori* (AP) LLRs. The prepared information are then sent to the BCJR decoder to decide which trellis (i.e., the related codeword) has the largest probability. Finally all the information collected from both FHT block and the BCJR decoder are sent to DFHT block to calculate the final APP LLR. In our proposed bi-directional LSTM decoder, the FHT/DFHT blocks are no longer needed. The LSTM networks are capable of merging the preparation stage, the forward/backward stage and the APP calculation stage into one network.

As shown in Fig. 1(b), we assume a CHC segmented into K (horizontal) blocks with r information bits and $2^r - r$ parity-check bits per block. Each block of CHC is actually a Hadamard code of length 2^r. A bi-directional LSTM decoder for the CHC is illustrated in Fig. 4, where the following steps are performed.

1) Send the K blocks of CHC AP LLRs to both the forward layer and the backward layer of the LSTM decoder consecutively. The input length is 2^r. (Note that we can also add an FHT block before feeding the LLRs to the LSTM. However, we find that the decoder performance remains the same.)
2) Perform both the forward/backward recursions.
3) Collect data from both the forward/backward layers and do batch normalizations to avoid gradient exploding.
4) Repeat step 1 to step 3 to get better decoding performance.
5) Use a fully-connected dense layer to generate the final APP LLR.

B. End-to-end THC decoder based on bi-directional LSTM

An end-to-end LSTM-THC decoder consists of a number of LSTM-CHC decoders plus interleavers and simple logic blocks. For example, the structure of an end-to-end LSTM-THC decoder with $M = 3$ is shown in Fig. 5. Note that the internal structures of the LSTM-CHC decoders are already illustrated in Fig. 4. Details of the intra-LSTM-CHCs decoding process are described as below.

1) We consider the i-th sub-decoder. The extrinsic LLRs computed during the last iteration of the i-th LSTM-CHC (which are stored in memory) are first subtracted from the output LLRs of the $(i-1)$-th LSTM-CHC to generate the AP LLRs. These AP LLRs and the parity-check LLRs of the i-th CHC are then sent to the i-th LSTM-CHC decoder. Note that the subtraction logics are fixed and no training is needed.
2) The i-th LSTM-CHC decoder produces the APP LLRs. The extrinsic LLRs are computed by subtracting the input AP LLRs from the output APP LLRs, and are stored in memory.
3) The APP LLRs are interleaved by an interleaver and then sent to the $(i+1)$-th LSTM-CHC. Note that the interleaver logics are also fixed and no training is needed.

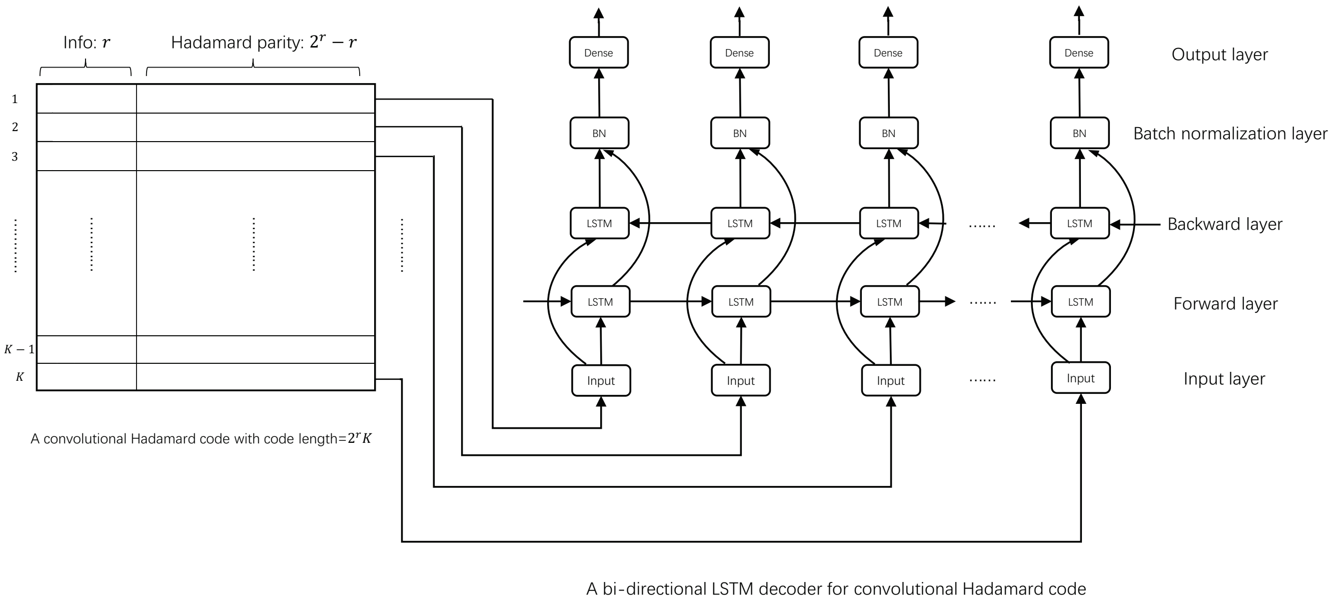

Info: r
Hadamard parity: $2^r - r$
1
2
3
$K-1$
K
A convolutional Hadamard code with code length=2^rK
Dense
BN
LSTM
Input
Output layer
Batch normalization layer
Backward layer
Forward layer
Input layer
A bi-directional LSTM decoder for convolutional Hadamard code

Fig. 4. Illustration of a bi-directional LSTM-CHC decoder.

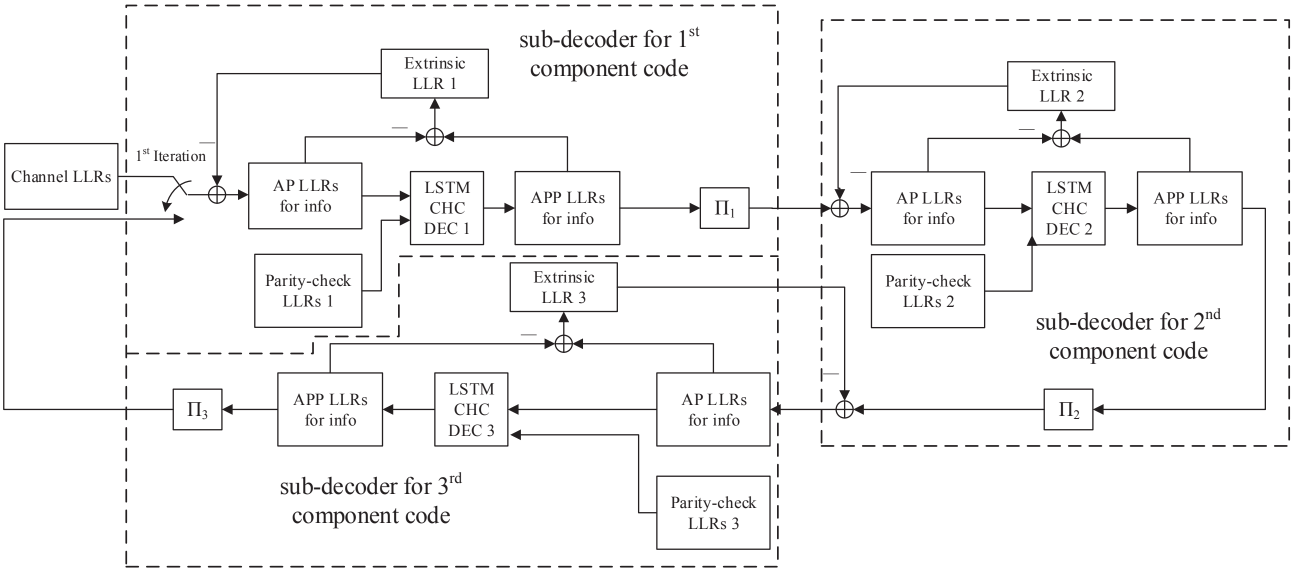


Fig. 5. End-to-end LSTM-THC decoder structure with $M = 3$.

4) Iterations of the sub-decoders continue until the maximum number of iterations is achieved.

IV. Training and Performance Evaluation

A. LSTM-CHC decoder

In this paper, we consider LSTM decoders for CHC with parameters $r = 4$, $K = 100$; and THC with parameters $r = 4$, $K = 100$, $M = 3$. For the LSTM-CHC decoder, 20000 codewords of CHC are generated. The 20000 codes are sent through an additive-white-Gaussian-noise (AWGN) channel with different bit-energy-to-noise-power-spectral-density ratio (E_b/N_0) for training. The model structure of the LSTM-CHC is illustrated in Fig. 6. The LSTM-CHC decoder consists of 6 layers. The shape of the input is $(\mathrm{B}, 100, 16)$ where B denotes the batch size; the number of LSTMs (which equals the number of blocks K in CHC) is 100; and the input length is $2^r = 16$. The number of hidden units in the LSTMs are set to 200 so the shapes of LSTM layers and batch normalization layers are $(\mathrm{B}, 100, 200)$. The output layer outputs a shape of $(\mathrm{B}, 100, r)$ where the APP LLRs of $r = 4$ information bits per block is given. Finally, hard decisions are made to the APP LLRs to obtain the decoded bits.

In this study, we train two separate LSTM-CHC decoders with CHCs at $E_b/N_0 = 2$ dB and $E_b/N_0 = 10$ dB, respectively. An adaptive moment estimation (ADAM) optimizer with learning rate 0.001 and mean squared error (MSE) loss

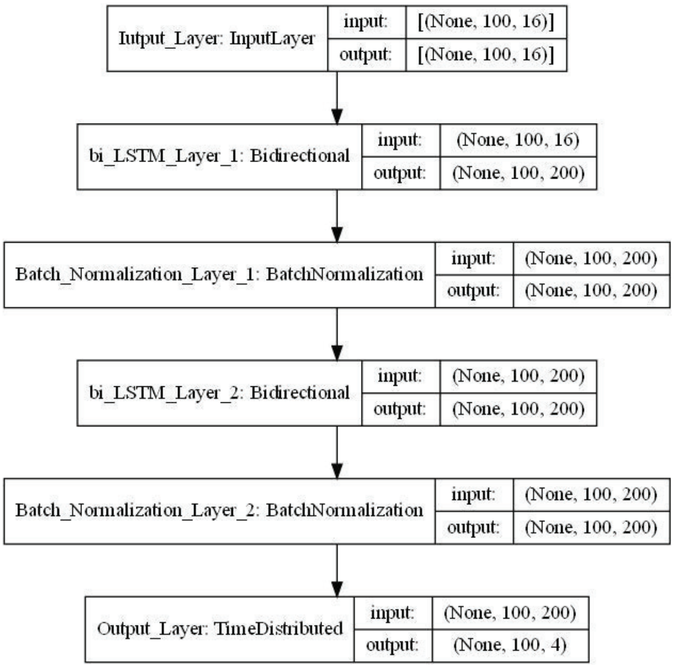


Fig. 6. An example of bi-directional LSTM-CHC decoder architecture.

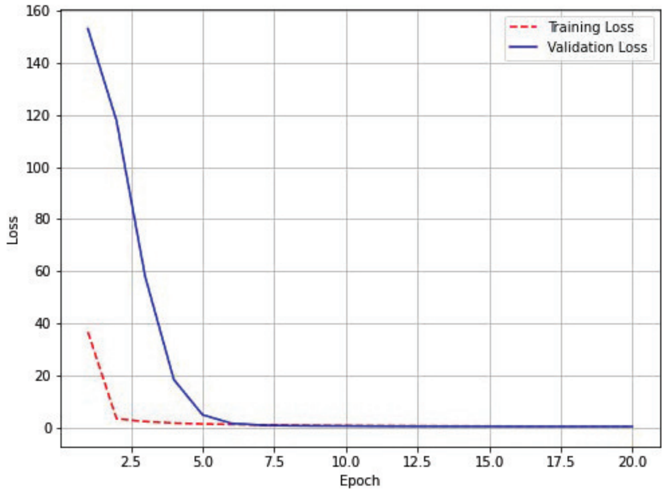


Fig. 7. Training loss and validation loss per epoch in LSTM-CHC.

function are used in the training. CHCs under various E_b/N_0 are used to test the performance of the trained LSTM-CHC decoders. Fig. 7 shows the evolution of training and validation losses with respect to the number of training epochs when the model is trained at $E_b/N_0 = 10$ dB. It takes only 7 epochs to make the validation loss close to 0. The BER results compared with conventional CHC decoder are shown in Fig. 8. The LSTM-CHC decoder trained at $E_b/N_0 = 2$ dB achieves the same performance as the conventional CHC decoder in the low-E_b/N_0-regime, i.e., $E_b/N_0 < 6$ dB; but perform badly in the high-E_b/N_0-regime. The LSTM-CHC decoder trained at $E_b/N_0 = 10$ dB, on the other hand, achieves almost the same performance as the conventional CHC decoder, particularly when $E_b/N_0 > 9$ dB.

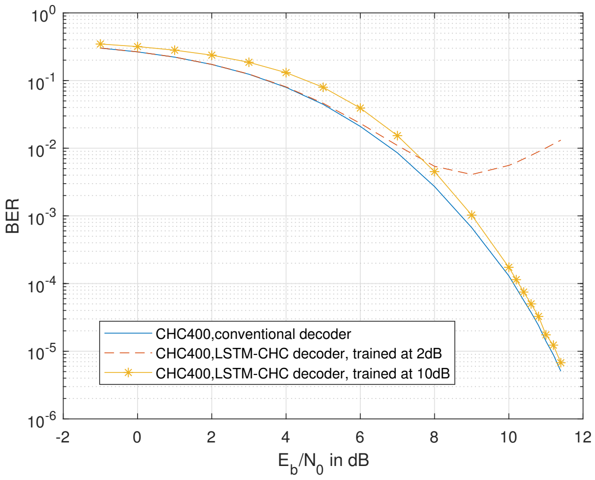


Fig. 8. BER performance of LSTM-CHC decoders.

B. LSTM-THC decoder

The direct training of the proposed end-to-end LSTM-THC decoder with such complex recurrent networks is challenging because the training from random weights costs a huge amount of time. Instead, we first train the MI LSTM-CHC sub-decoders individually and then use the trained weights as the initial weights of the end-to-end LSTM-THC decoders. (Recall that I denotes the number of iterations of a conventional THC decoder.) Details of the training process are listed as follows.

1) Use a THC encoder to randomly generate 20000 THC codewords. These 20000 codes are sent through an AWGN channel under $E_b/N_0 = -1$ dB. The received noisy observations are then decoded by the standard THC decoder described in [6]. During the decoding process, the intermediate data, i.e., the AP LLRs and the APP LLRs of each sub-decoder, are also recorded.
2) Assuming $I = 4$, train the $MI = 12$ LSTM-CHCs using the recorded LLRs with 100 epochs and ADAM optimizer with learning rate 0.001. The mean squared error function is used as the loss function. Record the trained weights of each LSTM-CHC.
3) Randomly generate 20000 THC codewords. Train the end-to-end LSTM-THC using the trained weights in Step 2 as initial weights. ADAM optimizer and MSE loss function are used. The learning rate is set to be 0.001 initially and is decreased after a number of epochs. The training is terminated early if the validating loss does not decrease within 10 epochs to prevent overfitting.

The BER performance of LSTM-THC decoder is shown in Fig. 9. It can be observed that the LSTM-THC decoder has a similar performance as the conventional THC decoder at $E_b/N_0 = -1$ dB, at which the end-to-end LSTM-THC is trained. However, the LSTM-THC decoder performs worse and worse than the conventional decoder as E_b/N_0 increases.

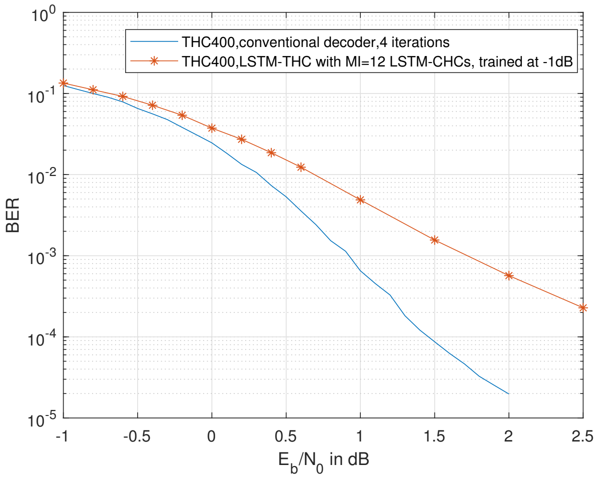


Fig. 9. Comparison of the BER performances of the LSTM-THC decoder and the conventional THC decoder. $MI = 12$ LSTM-CHCs are used in the LSTM-THC decoder.

V. Conclusion

In this paper, we propose to decode both CHC and THC via LSTM. The BER performance of LSTM-CHC is very close to that of the conventional CHC decoder when the training and testing E_b/N_0 are the same. Since the end-to-end training of LSTM-THC is very time-consuming, we are only able to train the LSTM-THC model at $E_b/N_0 = -1$ dB. The BER performance of LSTM-THC shows a loss compared to the conventional THC decoder when E_b/N_0 increases from $E_b/N_0 = -1$ dB. In our future work, we will train the LSTM-THC model at a higher E_b/N_0 and see if the model can produce comparable or better error performance than the conventional THC decoder.

References

[1] E. Nachmani, Y. Be'ery, and D. Burshtein, "Learning to decode linear codes using deep learning," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2016, pp. 341–346.

[2] E. Nachmani, E. Marciano, L. Lugosch, W. J. Gross, D. Burshtein, and Y. Be'ery, "Deep learning methods for improved decoding of linear codes," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 119–131, 2018.

[3] T. Gruber, S. Cammerer, J. Hoydis, and S. t. Brink, "On deep learning-based channel decoding," in *2017 51st Annual Conference on Information Sciences and Systems (CISS)*, 2017, pp. 1–6.

[4] S. Cammerer, T. Gruber, J. Hoydis, and S. ten Brink, "Scaling deep learning-based decoding of polar codes via partitioning," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6.

[5] H. Kim, Y. Jiang, R. Rana, S. Kannan, S. Oh, and P. Viswanath, "Communication algorithms via deep learning," in *Sixth International Conference on Learning Representations (ICLR)*, 2018.

[6] L. Ping, W. K. Leung, and K. Y. Wu, "Low-rate turbo-Hadamard codes," *IEEE Transactions on Information Theory*, vol. 49, no. 12, pp. 3213–3224, Dec 2003.

[7] W. K. R. Leung, G. Yue, L. Ping, and X. Wang, "Concatenated zigzag Hadamard codes," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1711–1723, April 2006.

[8] G. Yue, L. Ping, and X. Wang, "Generalized low-density parity-check codes based on Hadamard constraints," *IEEE Transactions on Information Theory*, vol. 53, no. 3, pp. 1058–1079, March 2007.

[9] S. Jiang, P. W. Zhang, F. C. M. Lau, and C.-W. Sham, "An ultimate-Shannon-limit-approaching Gbps throughput encoder/decoder system," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 10, pp. 2169–2173, 2020.

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

Sheng Jiang received the Bachelor of Engineering Degree in Microelectronics from Shanghai Jiaotong University, China; the Master of Engineering Degree in Electronic Engineering from Hong Kong University of Science and Technology, Hong Kong SAR; and the PhD degree from The Hong Kong Polytechnic University, Hong Kong SAR. He is currently a postdoctoral fellow at The Hong Kong Polytechnic University, Hong Kong SAR. His research interests include channel codes, hardware implementation of channel encoder/decoder, and machine learning.

Francis C. M. Lau received the BEng(Hons) degree in electrical and electronic engineering and the PhD degree from King's College London, University of London, UK. He is a Professor at the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China. He is also a Fellow of IEEE and a Fellow of IET.

He is a co-author of two research monographs and a co-inventor of six US patents. He has published more than 330 papers. His main research interests include channel coding, cooperative networks, wireless sensor networks, chaos-based digital communications, applications of complex-network theories, and wireless communications. Over the past years, he has secured research grants and consultancy projects from various organizations including the Hong Kong Research Grant Council; Hong Kong Jockey Club; Highways Department, Hong Kong SAR; National Natural Science Foundation of China; and Huawei Technologies Co. Ltd. He is a co-recipient of one Natural Science Award from the Guangdong Provincial Government, China; eight best/outstanding conference paper awards; one technology transfer award; two young scientist awards from International Union of Radio Science; and one FPGA design competition award.

He was the General Co-chair of International Symposium on Turbo Codes & Iterative Information Processing (2018) and the Chair of Technical Committee on Nonlinear Circuits and Systems, IEEE Circuits and Systems Society (2012-13). He served as an associate editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II (2004-2005 and 2015-2019), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I (2006-2007), and IEEE CIRCUITS AND SYSTEMS MAGAZINE (2012-2015). He has been a guest associate editor of INTERNATIONAL JOURNAL AND BIFURCATION AND CHAOS since 2010. He also served as a member of the IEEE CAS Society Fellow Evaluation Committee in 2022 and 2023.

QPSO-based Beamforming in Dual RIS-assisted Uplink Anti-jamming Communication System

Di Zhou[1], Zhiquan Bai[1,*], Jinqiu Zhao[1], Zeyu Liu[2], Dejie Ma[1], and KyungSup Kwak[3]
[1]Shandong Provincial Key Lab. of Wireless Communication Technologies,
School of Information Science and Engineering, Shandong University, Qingdao 266237, Shandong, China
[2]Department of Engineering Construction, China Mobile Inner Mongolia Co., Ltd. Baotou Branch, Baotou 014000, China
[3]Department of Information and Communication Engineering, INHA University, Incheon 22212, Korea
Email: emailofzhoudi@163.com, zqbai@sdu.edu.cn*, 202020373@mail.sdu.edu.cn,
13500621551@139.com, madj0212@163.com, kskwak@inha.ac.kr

***Abstract*—Considering a dual reconfigurable intelligent surface (RIS) assisted uplink cellular communication system under a malicious jamming user with known positions, we propose a joint active and passive anti-jamming beamforming scheme to maximize the system signal-to-interference-plus-noise ratio (SINR) and enhance the system achievable rate in this paper. To obtain the optimal solution, the quantum particle swarm optimization (QPSO) algorithm is utilized, which also mitigates the risk of falling into the local optimum and achieves better reliable global optimization. The simulation results illustrate that the proposed beamforming design in the dual RIS-assisted uplink cellular system exhibits superior anti-jamming performance compared with the other typical optimized beamforming schemes.**

***Index Terms*—anti-jamming, RIS, QPSO, beamforming, SINR**

I. Introduction

RECENTLY, reconfigurable intelligence surface (RIS) becomes a critical technology for 6G networks, offering flexible and cost-effective options to alter the wireless environment. The phases of RIS units can be designed to enable directional reflection of incident signals, leading to an improved channel propagation environment [1]. The change of the RIS reflection phase by controlling the reflection pattern is called the passive beamforming (PB), as compared to the active beamforming (AB) at the base station (BS) [2].

Specifically, [3] studied the RIS-assisted multi-cell wireless network, where the AB and the PB were designed to maximize the minimum achievable rate of the edge users. Furthermore, considering the non-ideal channel state information (CSI), robust beamforming optimization algorithms were proposed in [4] for RIS-assisted multi-user multiple input single output (MU-MISO) system. In response to potential eavesdropping attacks, [5] employed RIS to enhance the secrecy rate of the wireless communication system, simultaneously achieving lower transmission power. Furthermore, in [6], the transmission power was minimized by RIS deployment under the secrecy rate constraints. AB and PB of RIS were designed to enhance the secrecy rate in the anti-jamming scenarios for the multi-user multiple input multiple output (MU-MIMO) system [7].

Previous anti-jamming communication methods mainly utilize single RIS to improve the system performance. However, recent research has investigated the utilization of dual RIS-assisted wireless communication systems, which can enhance the system performance compared to single RIS cases, due to increased diversity and better beamforming vector design. Joint optimization of the reflection phase-shift matrices and the AB was studied in [8]. Moreover, the dual RIS in wireless secure transmission and an alternating optimization (AO) algorithm were proposed based on the product Riemannian manifold to jointly optimize the AB vector at the transmitter and the PB vector at the RIS for maximizing the system secure rate [9]. However, as the number of RISs increases, more inter-RIS links are formed, resulting in deep coupling of the reflected phase-shifts and increasing complexity in the joint optimization of the phase-shift matrix of each RIS. Therefore, researchers have focused on the beamforming design in multiple RIS-assisted wireless systems and intelligent algorithms to analyze the balance between performance benefits and complexity.

Inspired by the above research, we propose the quantum particle swarm optimization (QPSO) based joint AB and PB design for a dual RIS-assisted anti-jamming MISO wireless system in this paper. The QPSO algorithm is taken to iteratively optimize the beamforming in the cellular communication system, leveraging the strong global convergence properties of QPSO to improve algorithm efficiency and accuracy. We also present a beamforming design that optimizes the PB of both RISs, leading to more comprehensive optimization and better anti-jamming performance. Through simulation results, we demonstrate the effectiveness of our approach in reducing the jamming levels and improving the system signal-to-interference-plus-noise ratio (SINR) compared to conventional joint AB and PB optimization algorithms.

The paper is organized as follows. Section II briefly presents

The work is supported in part by the Innovation and Development Joint Foundation of Shandong Provincial Natural Science Foundation under Grant ZR2021LZH003, in part by the Joint Funds of the National Natural Science Foundation of China under Grant No. U23A20277, in part by the National Defense Preliminary Research Project under Grant 90903030202, in part by the National Nature Science Foundation of China under Grant 61771291, and in part by the National Research Foundation of Korea under Grant MSIT-NRF-2020R1A2B5B02002478. (Corresponding author: Zhiquan Bai.)

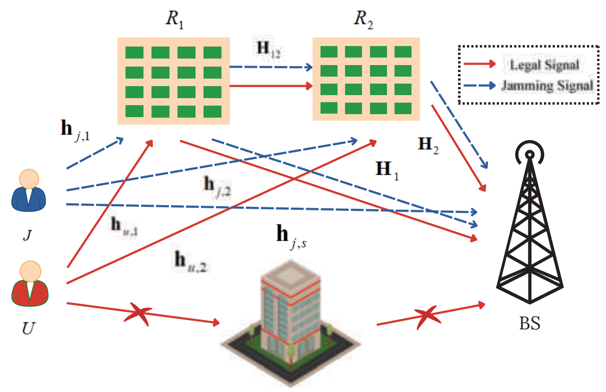


Fig. 1. Dual RIS-assisted uplink anti-jamming cellular system model.

the system model and the objective problem. Section III proposes the optimization of joint AB and PB design and the solution to maximize the system SINR and the achievable rate. Section IV contains the simulation results and analyses. Finally, Section V concludes the contents of the paper.

II. System Model and Problem Formulation

A. System Model

The dual RIS-assisted anti-jamming cellular system consists of one BS with N_t antenna, two RISs (R_1 and R_2), a legal user U, and a malicious jamming user J, as shown in Fig. 1. R_1 is near U with M_1 reflection elements, whereas R_2 is close to the BS with M_2 reflection elements. U and J have single transmit antenna. Due to the obstacles, such as building, there is no direct link between U and the BS. The signal from U should be transmitted to the BS by R_1 and R_2. We assume that $\mathbf{H}_{12} \in \mathbb{C}^{M_2 \times M_1}$, $\mathbf{H}_1 \in \mathbb{C}^{N_t \times M_1}$, and $\mathbf{H}_2 \in \mathbb{C}^{N_t \times M_2}$ denote the channel matrix from R_1 to R_2, from R_1 to the BS, and from R_2 to the BS, respectively. In a real urban environment, long-distance transmission tends to cause severe path loss. Therefore, we only assume the direct and the first-reflection links from U to the BS. The other signal transmissions, including scattered, diffracted, and multi-reflected signals through the RIS, are ignored.

The two RISs reflect the received signals from the BS to users through the RIS plane array, resulting in a multiplicative channel model. We define the diagonal PB matrices for R_1 and R_2 as $\mathbf{\Theta}_1 = diag\left(\boldsymbol{\theta}_1\right) \in \mathbb{C}^{M_1 \times M_1}$ and $\mathbf{\Theta}_2 = diag\left(\boldsymbol{\theta}_2\right) \in \mathbb{C}^{M_2 \times M_2}$, respectively. Here, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are reflection coefficient, where $|\theta_{1,m_1}| = 1, \forall m_1 \in \{1, 2, ..., M_1\}$ and $|\theta_{2,m_2}| = 1, \forall m_2 \in \{1, 2, ..., M_2\}$ are the m_1-th element of $\boldsymbol{\theta}_1$ and the m_2-th element of $\boldsymbol{\theta}_2$, respectively.

B. Problem Formulation

The transmission of the legal signal and the jamming signal is shown in Fig. 1. The composite channel from U to the BS can be expressed as

$$\mathbf{H}_u = \mathbf{H}_2\mathbf{\Theta}_2\mathbf{H}_{12}\mathbf{\Theta}_1\mathbf{h}_{u,1} + \mathbf{H}_1\mathbf{\Theta}_1\mathbf{h}_{u,1} + \mathbf{H}_2\mathbf{\Theta}_2\mathbf{h}_{u,2}, \quad (1)$$

where $\mathbf{h}_{u,1} \in \mathbb{C}^{M_1 \times 1}$ and $\mathbf{h}_{u,2} \in \mathbb{C}^{M_2 \times 1}$ are the channel matrices of the U-R_1 link and the U-R_2 link, respectively. The channel between J and the BS is $\mathbf{h}_{j,s} \in \mathbb{C}^{N_t \times 1}$. We can also write the composite channel from J to BS as

$$\mathbf{H}_j = \mathbf{H}_2\mathbf{\Theta}_2\mathbf{H}_{12}\mathbf{\Theta}_1\mathbf{h}_{j,1} + \mathbf{H}_1\mathbf{\Theta}_1\mathbf{h}_{j,1} + \mathbf{H}_2\mathbf{\Theta}_2\mathbf{h}_{j,2} + \mathbf{h}_{j,s}, \quad (2)$$

where the channel matrices of the J-R_1 link and the J-R_2 link are $\mathbf{h}_{j,1} \in \mathbb{C}^{M_1 \times 1}$ and $\mathbf{h}_{j,2} \in \mathbb{C}^{M_2 \times 1}$, respectively. We assume that, in the uplink transmission, the transmit signals of the legal user U and the jamming user J are x_u and x_j, respectively. Let $\mathbf{w}$ denote as the AB vector at the BS. The transmit power of U and J are respectively set to be P_u and P_j, and the received signal at the BS is given by

$$y = \mathbf{w}^{\mathrm{H}} \left(\sqrt{P_u}\mathbf{H}_u x_u + \sqrt{P_j}\mathbf{H}_j x_j + \mathbf{n} \right) \quad (3)$$

where $\mathbf{w}^{\mathrm{H}} \in \mathbb{C}^{1 \times N_t}$, $\mathbf{w}\mathbf{w}^{\mathrm{H}} = 1$. $(\cdot)^{\mathrm{H}}$ indicates the conjugate transpose operation of the vector and $\mathbf{n} \in \mathbb{C}^{N_t \times 1}$ represents the complex Gaussian noise with mean 0 and variance σ^2. By combining the equations (1)-(3), the system SINR can be calculated as

$$\mathrm{SINR} = \gamma\left(\mathbf{w}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right) = \frac{P_u \left| \mathbf{w}^{\mathrm{H}}\left(\mathbf{H}_u\right) \right|^2}{P_j \left| \mathbf{w}^{\mathrm{H}}\left(\mathbf{H}_j\right) \right|^2 + \sigma^2}. \quad (4)$$

According to the Shannon channel capacity theorem, the system achievable rate is given by

$$R_a = \log_2 \left[1 + \frac{P_u \left| \mathbf{w}^{\mathrm{H}}\left(\mathbf{H}_u\right) \right|^2}{P_j \left| \mathbf{w}^{\mathrm{H}}\left(\mathbf{H}_j\right) \right|^2 + \sigma^2} \right]. \quad (5)$$

To improve the system anti-jamming capability and achieve higher system achievable rates, the joint optimization of AB and PB is formulated by maximizing the system SINR as

$$\begin{aligned} &\max_{\mathbf{w}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2} \ \gamma\left(\mathbf{w}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right) \\ &\text{s.t.} \mathrm{C}_1 : \mathbf{w}\mathbf{w}^{\mathrm{H}} = 1 \\ &\qquad \mathrm{C}_2 : |\theta_{1,m_1}| = 1, \forall m_1 \in \{1, 2, ..., M_1\} \\ &\qquad \mathrm{C}_3 : |\theta_{2,m_2}| = 1, \forall m_2 \in \{1, 2, ..., M_2\}, \end{aligned} \quad (6)$$

where C_1 is the energy normalization constraint for the BS, C_2, and C_3 are the unit modulus constraints for the PB at these two RISs. Obviously, obtaining the optimum solution to the above optimization problem is challenging due to the coupled optimization variables and the non-convex objective function. Simultaneously optimizing the variables of the objective function with $\mathbf{w}, \boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$ is difficult. To overcome these problems, we use the QPSO algorithm to obtain the global optimum solution.

III. Algorithm Design

A. Algorithm Basics

The QPSO algorithm was first proposed as an intelligent optimization algorithm. In contrast to the standard particle swarm optimization (PSO) algorithm, the QPSO algorithm has faster search speeds and fewer parameters for convergence [10]. Let $\mathbf{X}_i$ and $\mathbf{pbest}_i$ be the vector position and the individual optimum position of the particle i, respectively.

The optimization formula to update the individual optimum position of the particle i is given by

$$\mathbf{pbest}_i = \left\{ \begin{array}{c} \mathbf{pbest}_i, if\left(f\left[\mathbf{X}_i \leq f\left[\mathbf{pbest}_i\right]\right]\right) \\ \mathbf{X}_i, if\left(f\left[\mathbf{X}_i > f\left[\mathbf{pbest}_i\right]\right]\right) \end{array} \right. . \quad (7)$$

Meanwhile, the updated formula of the global optimum position is

$$\mathbf{gbest} = \left\{ \begin{array}{c} s = \arg \max\limits_{1 \leq i \leq N} \left\{f\left[\mathbf{pbest}_i\right]\right\} \\ \mathbf{gbest}\left(t\right) = \mathbf{pbest}_s \end{array} \right. \quad (8)$$

where $\mathbf{gbest}$ is the global optimum position and N is the size of the particle population [10].

The individual attractor update equation is written as

$$\mathbf{art}_i = r_1 \cdot \mathbf{pbest}_i + (1 - r_1)\mathbf{gbest}, \quad (9)$$

where $\mathbf{art}_i$ denotes the individual attractor of the particle i, and r_1 is a random factor, with uniform distribution $r_1 \sim \mathrm{U}\left(0, 1\right)$.

In the process of the QPSO algorithm, the attractors integrate the individual optimum position and the global optimum position to represent the directive effect on the particle motion and ensure algorithm convergence. The particle position is

$$\mathbf{X}_i\left(t\right) = \mathbf{atr}_i \pm \frac{\mathbf{L}_i}{2} \ln\left(\frac{1}{r_2}\right), \quad (10)$$

where r_2 is a random factor, with uniform distribution $r_2 \sim \mathrm{U}\left(0, 1\right)$.

We insert the average of the individual best position $\mathbf{mbest}$ to calculate the next iteration of the particle i of the feature-length $\mathbf{L}_i$ and to control the feature-length. We have

$$\mathbf{mbest} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{pbest}_i. \quad (11)$$

According to [10], the feature-length is calculated as

$$\mathbf{L}_i = 2\alpha \cdot \left|\mathbf{mbest} - \mathbf{X}_i\left(t\right)\right|. \quad (12)$$

Moreover, the evolution function of the particle i is

$$\mathbf{X}_i(t+1) = \mathbf{art}_i \pm \alpha \cdot \left|\mathbf{mbest}(t) - \mathbf{X}_i\left(t\right)\right| \cdot \ln\left(\frac{1}{r_3}\right), \quad (13)$$

where r_3 is a random factor, with uniform distribution $r_3 \sim \mathrm{U}\left(0, 1\right)$ and α is the contraction-expansion coefficient and controls the activity range from particles to attractors.

By modifying the value of the contraction-expansion coefficient, which affects the search range of the particles, we can control the algorithm convergence rate. A larger contraction-expansion coefficient may increase the search range and enhance the global search ability. We set the value of α by the linear decreasing strategy as [1.0, 0.5] for improving algorithm performance.

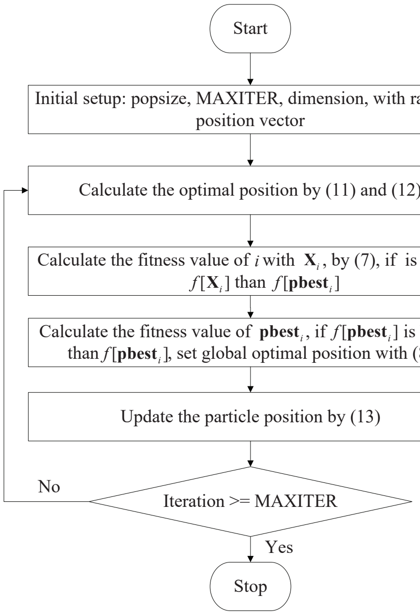

Start
Initial setup: popsize, MAXITER, dimension, with random position vector
Calculate the optimal position by (11) and (12)
Calculate the fitness value of i with $\mathbf{X}_i$, by (7), if is better $f[\mathbf{X}_i]$ than $f[\mathbf{pbest}_i]$
Calculate the fitness value of $\mathbf{pbest}_i$, if $f[\mathbf{pbest}_i]$ is better than $f[\mathbf{pbest}_i]$, set global optimal position with (8)
Update the particle position by (13)
No
Iteration >= MAXITER
Yes
Stop

Fig. 2. QPSO-based beamforming process.

B. Algorithm Process

We denote the position vectors of particles as $\mathbf{w}, \boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$. The QPSO-based joint AB and PB optimization flowchart is shown in Fig. 2 and its procedure is provided in the following.

step 1: Initialize the algorithm parameters, including the population size (popsize), dimension, and the maximum number of iterations (MAXITER).

step 2: Calculate the fitness values by mapping particles from the complex domain to the real domain.

step 3: Iterate to update the particle positions, including individual and global optimum positions, and attractor positions.

step 4: Set the maximum iteration limit. If the iteration meets the maximum iteration limit, output the optimum particle position. Otherwise, repeat steps 2 to 4.

step 5: According to the optimum particle position, calculate the optimization result.

IV. Simulation Results and Analysis

In this section, simulation results are provided to validate the effectiveness of the proposed algorithm. We assume that the receive antenna number of BS is 5. The system setup scenario is designed as follows, U and J are positioned at (1, 50, 0)m and (5, 40, 0)m, respectively. The BS is situated at (1, 0, 2)m. R_1 is placed close to U and R_2 is near the BS, with coordinates (0, 49, 1)m and (0, 1, 1)m, respectively. All channels within the system are modeled as Rayleigh fading distribution, and the path loss is $\varepsilon = \frac{\varepsilon_0}{d^{\alpha}}$, where d is the transmission distance, ε_0 is set as -30dB at the reference distance $d_0 = 1$, and α is the path loss coefficient. Also, the path loss coefficient of the U-R_1 link and J-R_1 link are both set to be 2.2, while to be 3 for the other links [11]. We compare the performance of the proposed QPSO algorithm with the

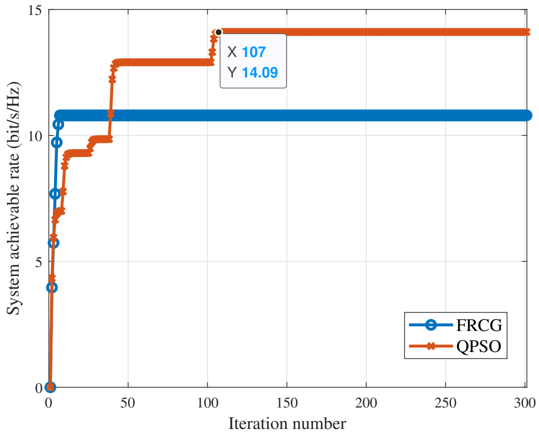


Fig. 3. Convergence performance of FRCG and QPSO.

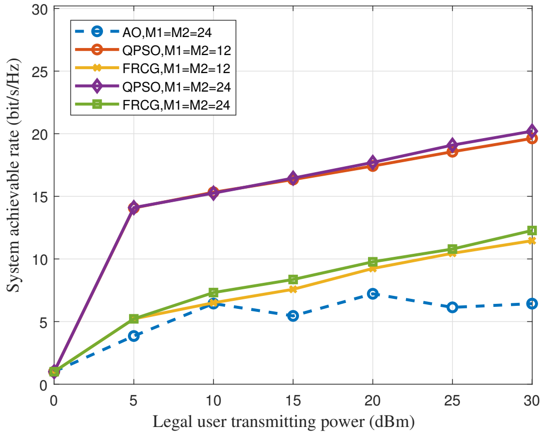


Fig. 5. Relationship between R_a and P_u for typical beamforming schemes.

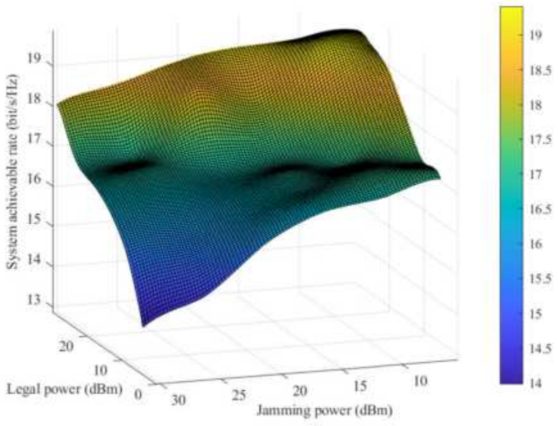


Fig. 4. Anti-jamming performance of the QPSO scheme vs. legal power and jamming power.

following benchmark algorithms, such as the Fletcher-Reeves conjugate gradient (FRCG) algorithm and the AO algorithm.

A. Performance comparison

The convergence performance of the FRCG and QPSO algorithms is evaluated in Fig. 3. We assume the same number of reflection elements for R_1 and R_2 as $M_1 = M_2 = 24$. The FRCG algorithm gets the system achievable rate converged to the fixed value with 11bit/s/Hz. As the iteration number increases above 107, the QPSO algorithm obtains the optimum achievable rate, which is 14bit/s/Hz. Although the FRCG algorithm has fewer iterations to converge than the QPSO algorithm, the system achievable rate of the QPSO algorithm is about 30% higher than that obtained by the FRCG algorithm.

In Fig. 4, the relationship between QPSO-based anti-jamming beamforming optimization performance and the legal power, as well as jamming power is plotted. The system achievable rate increases almost linearly with the power of the legal user. In contrast, when the jamming power rises, the increase in the system achievable rate plateaus. This is because the RIS consistently enhances the radio environment with phase-shift adjustment.

Fig. 5 illustrates the relationship between the transmit power P_u of the legal user and the system achievable rate. The proposed joint AB and PB design in the uplink anti-jamming cellular system enhances signal transmission and anti-jamming ability through diverse links and reduces channel correlation. Furthermore, the QPSO-based beamforming design achieves higher achievable rates compared to traditional AO and FRCG algorithms, highlighting its superior anti-jamming capability. We also simulate the case of $M_1 = M_2 = 12$, which shows that reducing the number of RIS elements is detrimental to the system anti-jamming performance.

In Fig. 6, the system achievable rate is plotted against the transmit power of the jamming user P_j. As P_j increases, the system achievable rate decreases, demonstrating the effectiveness of the optimization of PB at the two RISs in the suppression of jamming. Moreover, our proposed algorithm outperforms the other schemes and shows its advantages in improving the jamming immunity performance.

The relationship between the number of RIS elements and the system achievable rate is shown in Fig. 7 with legal user power $P_u = 15\text{dBm}$ and jamming user power $P_j = 15\text{dBm}$. In Fig. 7(a), the system achievable rate rises with the element number of R_1 under the condition of $M_1 + M_2 = 48$. The proposed QPSO algorithm outperforms the FRCG algorithm in achieving higher transmit rates at the same M_1, demonstrating its superior anti-jamming capability. Fig. 7(b) depicts that the system achievable rate increases as the total element number of the two RISs increases, indicating that the dual RIS-assisted approach can effectively enhance system anti-jamming capability. Regardless of the value of M_1+M_2, our proposed QPSO-based scheme outperforms the other methods. Overall, the results have shown that the RIS can improve the system anti-jamming performance significantly when it is located near the BS.

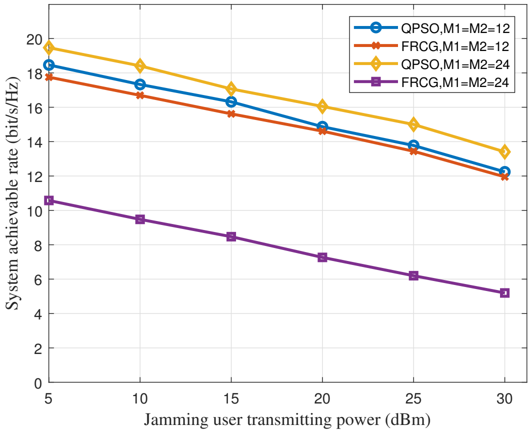
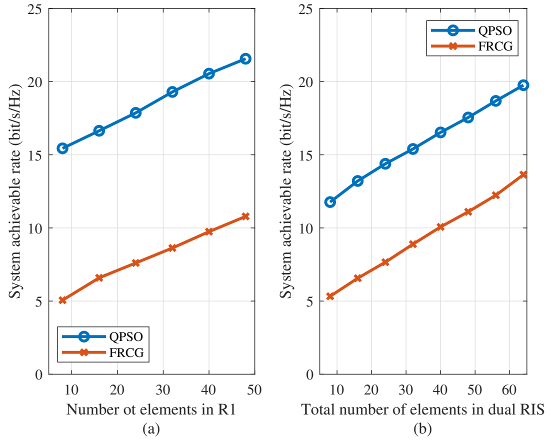


Fig. 6. Relationship between R_a and P_j for typical beamforming schemes.



(a)

(b)

Fig. 7. System achievable rate of QPSO and FRCG schemes. (a) Different element number of R_1. (b) Different element number of dual RIS.

B. Complexity comparison

To perform a comprehensive comparison based on profiling the anti-jamming performance of the proposed algorithm, we also analyze the balance of the performance and the computational complexity. Firstly, the QPSO algorithm convergence is affected by the number of iterations and the number of particles. The QPSO-based scheme requires more iterations and particles to avoid the local optimum and improve the performance of the global search for non-convex problems [10]. Thus, the number of maximum iterations and the number of particles for QPSO are set as $Q_i = 500$ and $N_p = 50$, respectively.

Meanwhile, the FRCG algorithm is faster in convergence and suitable for some low-dimensional problems but may be limited to locally optimum for non-convex problems. Its convergence performance is affected by the iteration number and the iteration step size, which may converge to the local optimum within relatively small number of iterations [12]. Therefore, for the FRCG algorithm, the maximum number of iterations is $F_i = 500$ and the step size of iterations is $S_i = 0.06$.

Particularly, the computational complexities of the QPSO and FRCG algorithms, which achieve the best outputs in terms of the system achievable rate and convergence behavior, are calculated as $\mathcal{O}\left(Q_i N_p\right)$ and $\mathcal{O}\left(F_i S_i\right)$, respectively. The QPSO algorithm can achieve the stable condition after 107 iterations, while the FRCG algorithm requires 16 iterations, as shown in Fig. 3. Although the complexity of the FRCG algorithm is lower than that of the QPSO algorithm, its anti-jamming performance is obviously poor. Overall, the QPSO algorithm is more suitable for solving the high-dimension non-convex optimization problem in this paper.

V. Conclusion

In this paper, we propose a QPSO-based joint AB and PB design for dual RIS-assisted uplink anti-jamming cellular communication to achieve enhanced global convergence and system performance. We formulate the joint AB and PB optimization problem to maximize the system SINR, utilizing the QPSO algorithm to optimize the global solution within beamforming vector constraints. Simulation results reveal that the proposed QPSO-based beamforming outperforms the standard AO and the FRCG algorithms in anti-jamming performance with respect to the system achievable rate.

References

[1] M. A. ElMossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han and G. Y. Li, “Reconfigurable Intelligent Surfaces for Wireless Communications: Principles, Challenges, and Opportunities,” *IEEE Trans. on Cogn. Commun. Netw.*, vol. 6, no. 3, pp. 990-1002, Sept. 2020.

[2] H. Guo, Y. -C. Liang, J. Chen and E. G. Larsson, “Weighted Sum-Rate Maximization for Reconfigurable Intelligent Surface Aided Wireless Networks,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3064-3076, May 2020.

[3] M. Hua, Q. Wu, D. W. K. Ng, J. Zhao, and L. Yang, “Intelligent Reflecting Surface-Aided Joint Processing Coordinated Multipoint Transmission,” *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1650-1665, Mar. 2021.

[4] G. Zhou, C. Pan, H. Ren, K. Wang, M. D. Renzo and A. Nallanathan, “Robust Beamforming Design for Intelligent Reflecting Surface Aided MISO Communication Systems,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1658-1662, Oct. 2020.

[5] M. Cui, G. Zhang, and R. Zhang, “Secure Wireless Communication via Intelligent Reflecting Surface,” *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1410-1414, Oct. 2019.

[6] Z. Chu, W. Hao, P. Xiao, and J. Shi, “Intelligent Reflecting Surface Aided Multi-Antenna Secure Transmission,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 1, pp. 108-112, Jan. 2020.

[7] Z. Chu *et al.*, “Secrecy Rate Optimization for Intelligent Reflecting Surface Assisted MIMO System,” *IEEE Trans. Inf. Forensic Secur.*, vol. 16, pp. 1655-1669, Nov. 2021.

[8] Y. Han, S. Zhang, L. Duan and R. Zhang, “Cooperative Double-IRS Aided Communication: Beamforming Design and Power Scaling,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 8, pp. 1206-1210, Aug. 2020.

[9] L. Dong, H. -M. Wang, J. Bai and H. Xiao, “Double Intelligent Reflecting Surface for Secure Transmission with Inter-Surface Signal Reflection,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2912-2916, Mar. 2021.

[10] J. Sun, W. Xu and B. Feng, “A Global Search Strategy of Quantum-behaved Particle Swarm Optimization,” in *Proc. IEEE CCIS*, vol.1, pp. 111-116, Jul. 2004.

[11] B. Zheng, C. You and R. Zhang, “Double-IRS Assisted Multi-User MIMO: Cooperative Passive Beamforming Design,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4513-4526, Jul. 2021.

[12] S. Masood, M. N. Doja and P. Chandra, “Analysis of Weight Initialization Routines for Conjugate Gradient Training Algorithm with Fletcher-Reeves Updates,” in *Proc. IEEE ICCCA*, pp. 304-308, 2016.

Di Zhou is pursuing her Ph.D. degree in electronic information from the School of Information Science and Engineering, Shandong University, Qingdao, China. She graduated with an M.S. degree in telecommunications engineering from the University of Sydney, Sydney, Australia in 2021. Her research interests include wireless network, reconfigurable intelligent surface, image processing, and deep learning.

Zhiquan Bai received the M.Eng. degree in communication and information system from Shandong University, Jinan, China, in 2003, and the Ph.D. degree (Hons.) from INHA University, Incheon, South Korea, in 2007, under the Grant of Korean Government IT Scholarship. He held a postdoctoral position with INHA University, and was a Visiting Professor with The University of British Columbia, Canada. He is currently a Professor with the School of Information Science and Engineering, Shandong University. His research interests include cooperative technology and spatial modulation, orthogonal time frequency space modulation, MIMO technology, resource allocation and optimization, and deep-learning based 5G wireless communications. He is a member of the editorial board of Journal of Systems Engineering and Electronics and also an associate editor of the International Journal of Communication Systems.

Jinqiu Zhao received B.E. degree from Shandong Normal University, Jinan, China, in 2020. She is currently pursuing her Ph.D. degree in the School of Information Science and Engineering, Shandong University, Qingdao, China. Her main research interests include reconfigurable intelligent surface and machine learning.

Zeyu Liu received B.E. degree from Inner Mongolia University , Huhehaote, China, in 2000. He is currently an Engineer in China Mobile , Baotou, China. His main research interests include Mobile Communication and Transmission Network Technology.

Dejie Ma is currently pursuing the M.S. degree in Electronic Information at the School of Information Science and Engineering, Shandong University, Qingdao, China. His research interests include reconfigurable intelligent surface, integrated sensing and communication and signal processing.

KyungSup Kwak received his BS degree from the Inha University, Inchon, Korea, in 1977 and his MS degree from the University of Southern California in 1981 and his PhD degree from the University of California at San Diego in 1988, under the Inha University Fellowship and the Korea Electric Association Abroad Scholarship Grants, respectively.From 1988 to 1989, he was with Hughes Network Systems, San Diego, California. From 1989 to 1990, he was with the IBM Network Analysis Center, North Carolina. Since then, he has been with the School of Information and Communication Engineering, Inha University, Korea, as a professor. He is the director of UWB Wireless Communications Research Center (UWB-ITRC).Since 1994, he served as a member of the board of directors and the vice president and the president of Korean Institute of Communication Sciences (KICS) in 2006 and the president of Korea Institute of Intelligent Transport Systems (KITS) in 2009. He received many research awards, such as the award of research achievements in UWB radio from the Ministry of Information and Communication and Prime Ministry of Korea in 2005 and 2006, respectively. In 2008, he is elected as Inha Fellow Professor (IFP). In 2010, he received the Korean President official commendation for his contribution to ICT innovation and industrial promotion.He published more than 100 SCI journal papers, 300 conference/domestic papers, obtained 20 registered patents and 35 pending patents, and proposed 21 technical proposals on IEEE 802.15 (WPAN) PHY/MAC. He is one of the members of the IEEE, IEICE, KICS, and KIEE. His research interests include multiple access communication systems, cognitive radio, UWB radio systems and WBAN, WPAN, and sensor networks.

A compact dual-band metamaterial absorber using square split rings for C-band and X-band sensors applications

1stRamesh Amugothu
Dept of ECE
NIT Warangal
Warangal, India
ar720057@student.nitw.ac.in

2ndVakula Damera
Dept of ECE
NIT Warangal
Warangal, India
vakula@nitw.ac.in

***Abstract*—A novel dual-band metamaterial absorber is proposed to achieve narrow-band absorption in the microwave range, making it highly suitable for sensor applications. Comprising two split-ring resonators one operating in the S-band and the other in the C-band the absorber is designed to operate simultaneously in both bands, providing narrowband absorption for different angles of incidence. The proposed absorber can achieve absorption peaks of greater than 99.5% and 96.9% in the respective bands. The physical mechanism of the proposed absorbers is demonstrated, along with the representation of its permeability and penetrability values, as well as its electric and magnetic field distributions. The metamaterial features a planar structure, showcasing polarisation-insensitivity and angle-insensitive absorptive properties. Furthermore, the absorber has a compact size, making it suitable for sensing, EMI, and EMC applications.**

***Index Terms*—metamaterial absorber, polarisation insensitive, dual band, compact.**

I. Introduction

In recent decades, metamaterial perfect absorbers (MPAs) have garnered significant attention due to their crucial role in various promising applications, such as thermal emitters, sensors, and solar cells. The fundamental principle behind MPAs lies in the suppression of reflection from incident electromagnetic waves. This is achieved by employing an optically thin metallic layer backed on the substrate of a metamaterial absorber to block transmission. Consequently, the widely adopted metal-dielectric-metal (MDM) configuration in research [1], [2] utilizes perfect absorption based on geometric and material characteristics. In the MDM configuration, perfect absorption is attainable by manipulating the geometric and material features. The structure of MDM absorbers is designed to control the resonance absorption band. Sensors exhibiting multiband absorption responses are potentially valuable in chemical and biomedical applications. Recent studies have showcased various plasmonic structures with multiple absorption bands [3], [4], [5]. For sensor applications requiring precise absorption frequency detection, MPAs with a narrow absorption band are essential. In microwave metamaterial absorbers, near-unity absorption bands are scarce. However, a multiband absorber can detect multiple frequencies, enhancing the metamaterial absorber's utility as a sensor. The literature presents several examples of multi-band metamaterial absorbers [6]–[12].

The paper proposes a dual-band metamaterial absorber on a metal-backed dielectric substrate with a square split-ring geometry. The design focuses on minimizing the reflection coefficient at the resonance frequency. The detailed design of the unit cell geometry is presented in Section II, while Section III describes the absorption characteristics, including the surface current field distribution of the metamaterial. To conclude, Section IV provides a brief summary of the design and its characteristics. Utilizing low-cost materials with simple designs offers additional advantages compared to pre-existing designs [13]–[15].

II. Unit cell Design

The proposed structure, illustrated in Fig. 1, is designed to minimize the reflection coefficient of the unit cell [13]. The unit cell patch comprises a circular split-ring resonator enclosed by a square-ring resonator. The two rings are connected using conduction in the stings of dimension 'd,' and the corners of the square consist of square patches of dimensions 'b1' and 'b2' to enhance mutual coupling between the two resonators at higher frequencies. A metal plate is placed on the back of the substrate. The structural parameters are tailored for an FR4-epoxy dielectric substrate with a thickness of 1.6 mm. The optimized parameter values are 'a' = 23 mm, 'b' = 22 mm, 'd' = 20.5 mm, 'l' = 7 mm, 'w' = 0.4 mm, 'b1' = 3.2 mm ('b2' is also set to 3.2 mm), 'g' = 0.2 mm, 'g1' = 1 mm, and 'h' = 1.6 mm. These dimensions are optimized to enhance the performance of the metamaterial absorber for use in sensor and electromagnetic detector applications. Both the top and bottom are constructed using copper metal with a thickness of 0.035 mm on an FR4 substrate with a thickness of 1.6 mm, a relative permittivity of 4.4, and a loss tangent of 0.02. The proposed unit cells are simulated using the commercially available software HFSS under a periodic boundary condition with a normal incidence of electromagnetic waves.Â

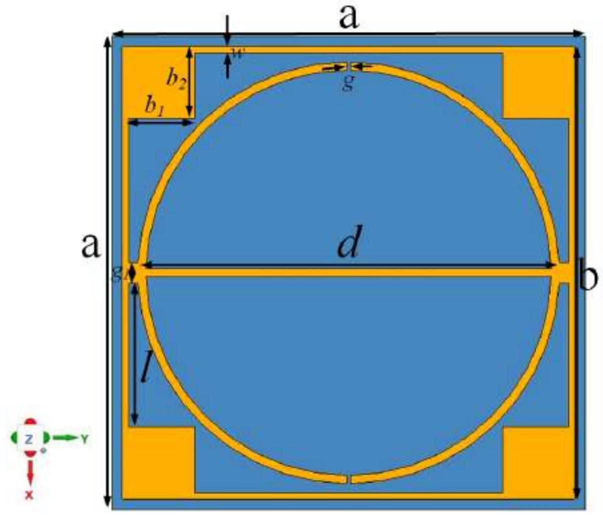


Fig. 1: unit cell geometry top view.

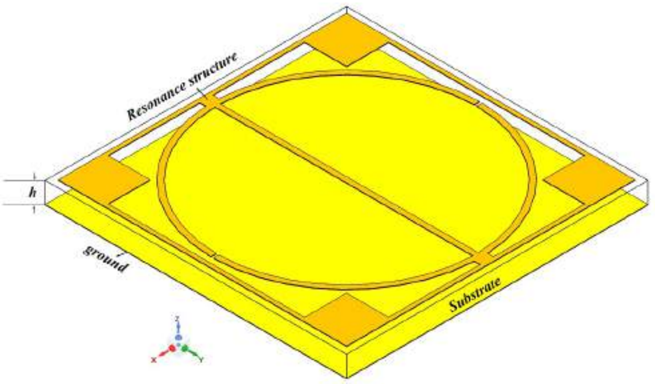
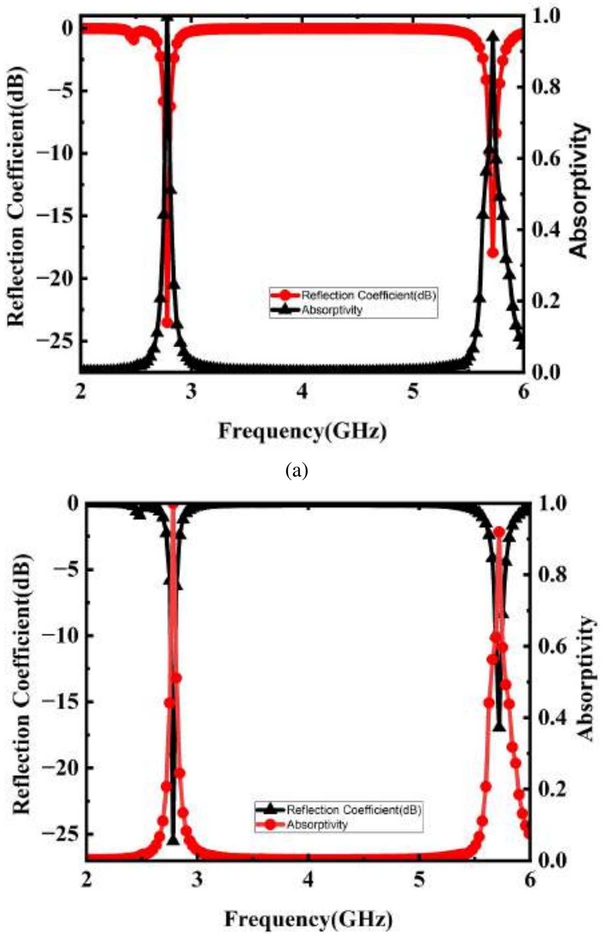


Fig. 2: unit cell geometry Isometric view.

The simulated results of the proposed structure is shown in Figure 3. From Fig.3(a,b) it is observed that the reflection coefficient is -24.2 dB and -18.6 dB at the frequencies 2.78 GHz and 5.78 GHz respectively and also the absorptivity is 99.5% and 96.9%.

III. Results and analysis

To calculate the absorption by reflection and transmission coefficients, the following formula is used: $A(\omega) = 1 - R(\omega) - T(\omega)$ [16]. where $R(\omega)$ is the reflection coefficient as a function of frequency and $T(\omega)$ is the transmission coefficient as a function of frequency. Maximum absorption occurs when the minimum reflection of the incident wave happens at the surface. The simulated reflection coefficient and absorptivity for TE and TM modes are shown in Fig.3.



(a)

(b)

Fig. 3: Simulated reflection coefficient and absorptivity for (a) TE (b) TM..

The Nicolson-Ross-Weir technique is used to obtain the effective medium parameters such as relative permittivity, relative permeability, refractive index, and relative impedance, as shown in Fig. 4. From Figure 4(a, b), at frequency 2.78 GHz, the relative permittivity and relative permeability values are negative, and at frequency 5.78 GHz, the relative permeability is negative and the relative permittivity is near zero. Due to this, there is a maximum absorption 99.8% at 2.78 GHz and 99.6% absorption at 5.78 GHz, and their impedance is near unity, as shown in Figure 4(d).

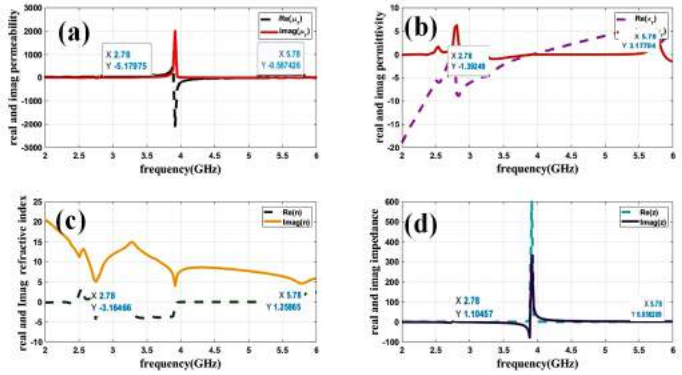
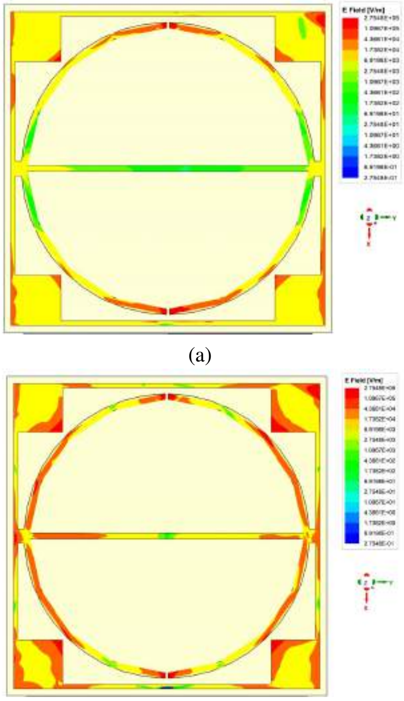


Fig. 4: Real and Imaginary parts of effective medium parameters.

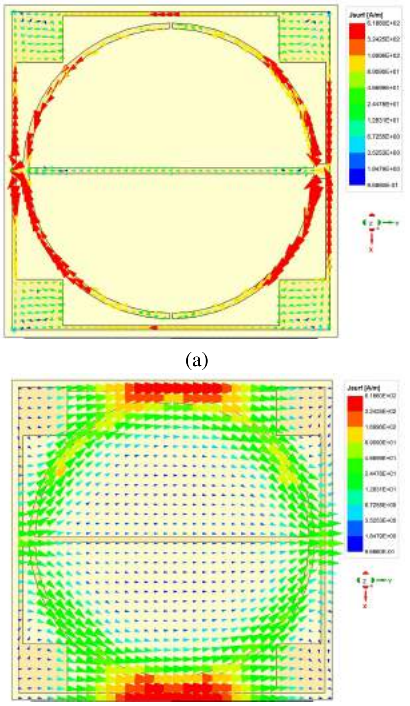
A description of the physical mechanism of the proposed metamaterial absorber is given in Figures 5, 6, and 7 as a function of the electric field distribution and the surface current distribution. The electric field distributions and surface current distributions are analyzed to gain insight into the generated resonance bands. At a frequency of 2.78 GHz, it can be claimed that the inner circle has an electric and magnetic resonance that causes absorption. The absorption at 5.78 GHz is due to the presence of the outer ring structure. The strong electric field and current distribution are presented for both bands, as illustrated in Figs. 5 and 6. In Figs. 6 and 7, there is an opposite current at the bottom, because of which there is a strong magnetic field present. Because of this, there is a simultaneous magnetic and electric field, causing strong resonance.



(a)

(b)

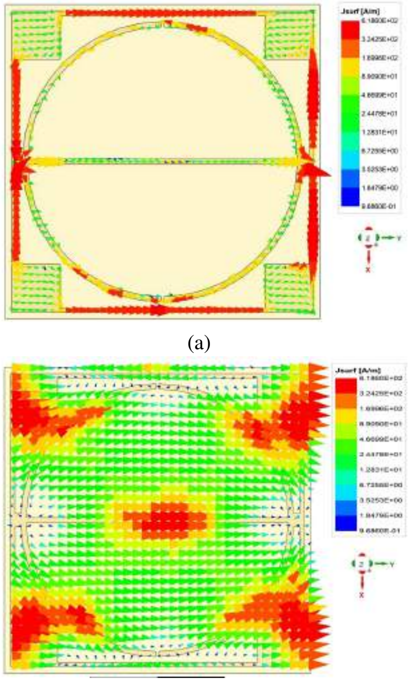
Fig. 5: The electric field distribution (a) 2.78 GHz (b) 5.78GHz .



(a)

(b)

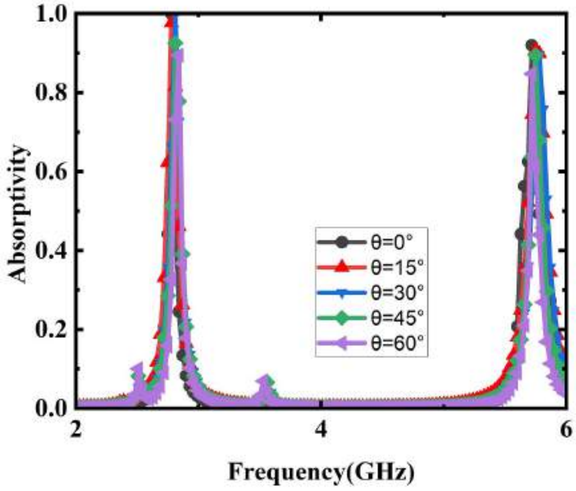
Fig. 6: The surface current distribution (a) 2.78 GHz Top view (b) 2.78 GHz Bottom view .



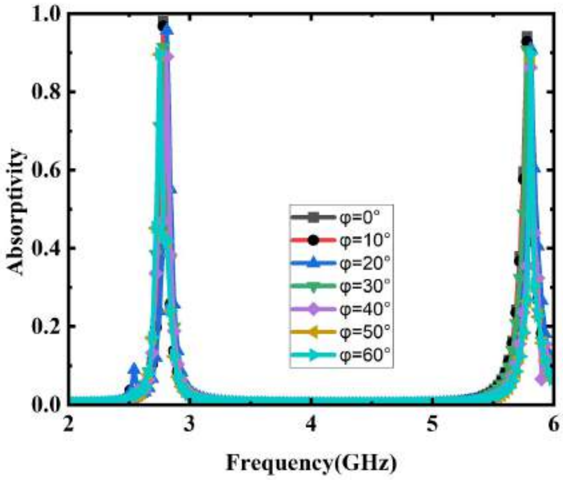
(a)

(b)

Fig. 7: The surface current distribution (a) 5.78 GHz Top view (b) 5.78 GHz Bottom viewr.



(a)



(b)

Fig. 8: The oblique incidence angle for proposed absorber TE and TM.

The simulated reflection coefficients of dual-band absorbers with TM polarised waves and TE-polarized waves are presented in Fig. 8. From the figures, it is evident that the proposed absorber has a wide range of angular stability 60^0. The absorption band at two resonance frequencies is 0.1 GHz and 0.14 GHz. The bandwidth is calculated as the range of the frequency width at half maximum (FWHM). The consequent quality factor (Q = f/FWHM, where f indicates resonance frequency) is between 32.6 and 34.04. Furthermore, due to its symmetrical geometry, it exhibits polarization insensitivity under normal incidence. As a result, the designed effective absorber can be used as a sensor and EM detector for GHz frequencies.

IV. Conclusion

This paper proposes a dual-band metamaterial absorber on a metal-backed dielectric substrate for applications in the S and C bands. Using a square and circular split-ring geometry, the structure produces two absorption bands at 2.78 GHz and 5.78 GHz. Additionally, the device is polarization-insensitive and stable under the oblique incidence of EM waves, which makes it a good candidate for practical wireless applications. To understand the structure's characteristics, we examined its current distribution patterns and surface power loss density. There is a design parameter that can be used to tune the resonance frequencies of the proposed structure. The proposed design has a high quality factor, which is an important factor for sensing applications. The metamaterial absorber is proposed for use in sensors and EM detectors.

References

[1] A. K. Osgouei, H. Hajian, A. E. Serebryannikov, and E. Ozbay, "Hybrid indium tin oxide-au metamaterial as a multiband bi-functional light absorber in the visible and near-infrared ranges," *Journal of Physics D: Applied Physics*, vol. 54, no. 27, p. 275102, 2021.

[2] A. K. Osgouei, A. Ghobadi, B. Khalichi, and E. Ozbay, "A spectrally selective gap surface-plasmon-based nanoantenna emitter compatible with multiple thermal infrared applications," *Journal of Optics*, vol. 23, no. 8, p. 085001, 2021.

[3] E. Buhara, A. Ghobadi, B. Khalichi, H. Kocer, and E. Ozbay, "Mid-infrared adaptive thermal camouflage using a phase-change material coupled dielectric nanoantenna," *Journal of Physics D: Applied Physics*, vol. 54, no. 26, p. 265105, 2021.

[4] A. Kalantari Osgouei, H. Hajian, B. Khalichi, A. E. Serebryannikov, A. Ghobadi, and E. Ozbay, "Active tuning from narrowband to broadband absorbers using a sub-wavelength vo 2 embedded layer," *Plasmonics*, vol. 16, pp. 1013–1021, 2021.

[5] P. B. Johnson and R.-W. Christy, "Optical constants of the noble metals," *Physical review B*, vol. 6, no. 12, p. 4370, 1972.

[6] R. K. Singh and A. Gupta, "An ultra-thin polarization independent quad-band metamaterial-inspired absorber for x-and ku-band applications," in *2020 6th International Conference on Signal Processing and Communication (ICSC)*. IEEE, 2020, pp. 34–38.

[7] R. Amugothu and D. Vakula, "Design and modelling of wide incidence angle double-band metamaterial absorbers for applications in the x frequency bands," in *2022 IEEE International Symposium on Smart Electronic Systems (iSES)*, 2022, pp. 62–65.

[8] S. Ampavathina and V. Damera, "Compact complementary symmetric ring resonator band stop filter for ku band applications," in *2022 IEEE Wireless Antenna and Microwave Symposium (WAMS)*, 2022, pp. 1–5.

[9] S. Genikala and A. Ghosh, "Design of polarization-insensitive dual band microwave absorber for emi/emc applications," in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*. IEEE, 2020, pp. 433–436.

[10] G. Sen and A. Ghosh, "Dual band metamaterial absorber using concentric split-ring structures for wireless applications," in *2019 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*. IEEE, 2019, pp. 1–4.

[11] A. K. Osgouei, B. Khalichi, E. Buhara, A. Ghobadi, and E. Ozbay, "Dual-band polarization insensitive metamaterial-based absorber suitable for sensing applications," in *2021 IEEE Photonics Conference (IPC)*. IEEE, 2021, pp. 1–2.

[12] M. Chaluvadi and P. Rao, "Wide-angle and polarization insensitive dual-band metamaterial absorber," in *2019 IEEE 5th Global Electromagnetic Compatibility Conference (GEMCCON)*. IEEE, 2019, pp. 1–3.

[13] J. Chen, Z. Hu, G. Wang, X. Huang, S. Wang, X. Hu, and M. Liu, "High-impedance surface-based broadband absorbers with interference theory," *IEEE Transactions on Antennas and Propagation*, vol. 63, no. 10, pp. 4367–4374, 2015.

[14] G. Sen and S. Das, "Frequency tunable low cost microwave absorber for emi/emc application," *Progress In Electromagnetics Research Letters*, vol. 74, pp. 47–52, 2018.

[15] G. Sen, A. Banerjee, M. Kumar, S. N. Islam, and S. Das, "A dual band metamaterial inspired absorber for wlan/wi-max applications using a novel i-shaped unit cell structure," in *2016 Asia-Pacific Microwave Conference (APMC)*. IEEE, 2016, pp. 1–3.

[16] N. I. Landy, S. Sajuyigbe, J. J. Mock, D. R. Smith, and W. J. Padilla, "Perfect metamaterial absorber," *Physical review letters*, vol. 100, no. 20, p. 207402, 2008.

Session 2B: Artificial Intelligence 2

Chair: Prof. Bing-Yuh Lu, Guangdong University of Petrochemical Technology, China

1 Paper ID: 20240037, 104~106

Optimizing Implementation of SNN for Embedded System

Dr. Hyeonguk Jang, Dr. Jae-Jin Lee, Dr. Kyuseung Han,

Electronics and Telecommunications Research Instit. Korea(South)

2 Paper ID: 20240049, 107~115

Multivariate PCA-based Composite Criteria Evaluation Method for Anomaly Detection in Manufacturing Data

Mr. HyeokSoo Lee, Mr. Youngki Jo, Prof. Jongpil Jeong,

Department of Smart Factory Convergence, Sungkyunkwan University, Korea(South)

3 Paper ID: 20240248, 116~121

Pitching-Motion: Pose-Based Pitch Trajectory Overlay System

Mr. Bor-Yao Tseng, Mr. Hung-Tse Chiang, Prof. Jiann-Liang Chen, Mr. Han-Chuan Hsieh,

National Taiwan University of Science & Technology. Taiwan

4 Paper ID: 20240469, 122~130

A Review of Detection-related Multiple Object Tracking in Recent Times

Ms. Suya Li, Ms. Ying Cao, Ms. Xin Xie,

Henan University. China

5 Paper ID: 20240475, 131~136

An Enhanced Topic Modeling Method in Educational Domain by Integrating LDA with Semantic

Mr. Ruofei Ding, Mr. Pucheng Huang , Ms. Shumin Chen, Mr. Jiale Zhang, Mr. Jingxiu Huang, Mr. Yunxiang Zheng,

South China Normal University. China

Optimizing Implementation of SNN for Embedded System

Hyeonguk Jang, Jae-Jin Lee, Kyuseung Han

Electronics and Telecommunications Research Institute, Daejeon, 34129, Korea

lemon@etri.re.kr, ceicarus@etri.re.kr, han@etri.re.kr

Abstract— **Spiking neural networks (SNNs) are a highly promising AI technology for embedded systems, owing to their energy-efficient properties. However, the manual implementation of SNNs encounters practical challenges because of the all-to-all connections in large networks. Thus, this paper presents a novel methodology to reduce wire congestion in the SNN implementations while mitigating adverse effects on inference accuracy.**

Keywords— **back propagation through time, deep learning, embedded system, hardware, spiking neural networks**

I. Introduction

Embedded systems are special computing systems used in a wide range of applications, from home appliances such as smartphones and IoT devices to industrial control systems, automotive systems, and medical devices [1]. Unlike existing computers, it has strong constraints on power consumption, area, and real-time processing, while performing multiple tasks or functions within a specific application or domain. Therefore, it is important to optimize both hardware and software to meet the stated requirements rather than including general-purpose features.

The recent advancements in AI have spurred a growing demand for AI capabilities in embedded systems. Spiking neural networks (SNNs) offer a distinct advantage over other AI technologies, primarily due to their low-power operation. Learning is done on an external powerful server, and the typical configuration for embedded systems is to focus on the inference function of the target application [2].

One of the major challenges in using SNNs in the embedded systems is that SNNs require complex circuits that must be implemented manually. SNNs consume low power by operating only when an event occurs. This feature can be realized when the circuits are implemented using analog or asynchronous digital techniques [3], which cannot be automated. Thus, in this paper, we will propose a methodology to alleviate this difficulty.

II. Implementation of SNN

SNN is a neural network that mimics spike-based operations of a biological brain, which is show in Figure 1. Its architecture mainly consists of two types of layers, convolution layers and fully-connected layers (FCL). Among these, we will focus on the FCL in this paper. FCL is literally a layer in which the neurons in the current layer and the neurons in the next layer are all-to-all connected as shown in Figure 2.

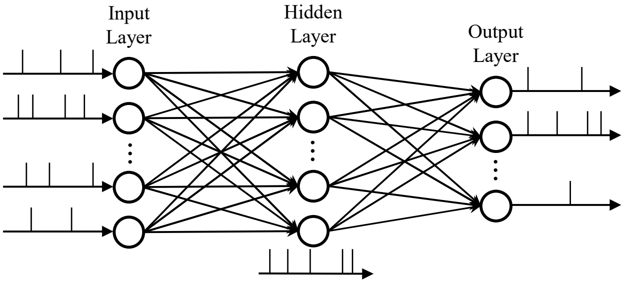


Figure 1. Spiking Neural Network

It incurs three problems when implementing FCL. First, FCL necessitates the substantial number of wires to connect between nodes. When implemented on a chip, the complexity of wire connections increases, the area expands, and the number of metal layers increases. And wire delay becomes longer and scalability decreases.

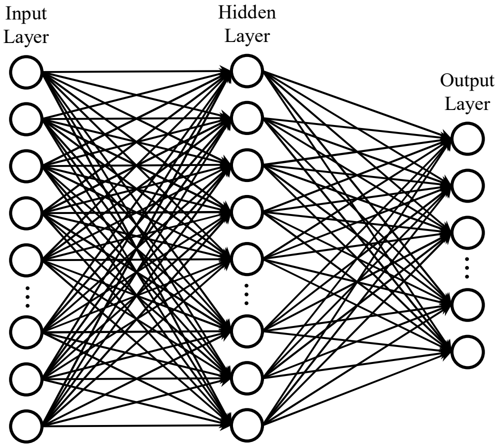


Figure 2. Fully-connected layers

Second, the first problem becomes more serious in that SNNs require manual implementation. For low power, SNN must operate only when an event occurs, and the current way to implement SNN is asynchronous digital circuit design or analog circuit design. Both of these ways must be designed manually and cannot be implemented through HDL-level automation. Therefore, implementing FCL takes a lot of time and effort.

Lastly, if FCL is designed to have connections with different delays among them for simplifying implementation, the learning will become much more challenging, or additional hardware will be required. A representative example is mesh structure. Ideally, spikes fired from neurons in the current layer should reach neurons in the next layer at the same time. However, in mesh structure, the arrival times are not the same. There are two ways to solve this problem. The first approach is to learn SNN as a learning method that takes into account the timing characteristics of the hardware. This approach is not only difficult to learn, but also has the problem of the software framework becoming diverse. The second approach is to synchronize the spike arrival times by introducing a hardware component, such as a global synchronizer. This approach has a high level of difficulty in coordinating signals between asynchronous and synchronous logics. And because this approach is centralized, it creates additional scalability issues.

III. Proposed Adjacently-connected Layer

To reduce the number of connections between neurons, we introduce a structure where only adjacent neurons are connected, as illustrated in Figure 3. We call this structure an adjacently-connected layer (ACL). Because significantly reduced connections could potentially affect inference accuracy, we will investigate this effect in the experiments section. Additionally, the connection method is explained as follows.

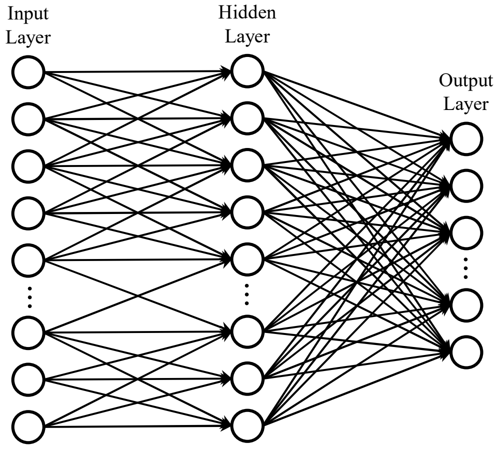


Figure 3. Proposed adjacently-connected layers

The proposed adjacent-connection restricts the number of neurons in the next layer that can be connected to a neuron in the current layer to a fixed count N. A method for determining which neurons in the next layer should be connected to neurons in the current layer is as follows. First, we calculate the relative position within the layer for a neuron in the current layer. Then we look for the most adjacent neuron whose the relative position within the next layer is most similar to the relative position of the neuron within the current layer. Lastly, we select a total of N neurons with the most adjacent neuron and N/2 neurons above the adjacent neuron and N/2-1 neurons below the adjacent neuron. Here, if the number of neurons above is less than N/2 or the number of neurons below is less than N/2-1, the number of selected neurons is less than N. Ultimately, we can construct ACL by applying the method to all neurons in the current layer.

This approach is different from conventional pruning techniques [4]-[6]. Pruning is a method of lightening the model by training it not to use some of the connections, although they are implemented. Therefore, pruning is not a technique that affects implementation. Since the required connections change each time we train, the implementation of all-to-all connection is ultimately required.

IV. Experiments

A. Experimental Environments

For the experiments, we performed the simulation of SNNs using the snnTorch library based on the PyTorch deep learning library. MNIST [7] and Fashion-MNIST [8], which are used for image classification, were used as datasets for the experiment. The network structure used for classification of the two datasets consists of input, hidden, and output layers, and the number of neurons in each layer is 784, 300, and 10, respectively. The integration and fire (I&F) neuron model without bias was used as the neuron model of SNN.

B. Learning of SNNs with Adjacently-connected Layers

For SNN learning, we used the Back Propagation Through Time (BPTT) [9] algorithm. To control synaptic connections, we utilized the custom mask in the pruning function provided by PyTorch. The custom mask enables the establishment of connections between adjacent neurons, while the weights of the unconnected synapses are set to 0, ensuring that only the weights of the connected synapses are trained.

C. Experimental Results

In this experiment, the accuracy of each MNIST dataset and Fashion-MNIST dataset was measured for the cases where the number of neurons in the hidden layer to which a neuron in the input layer is connected is all-to-all, 300, 250, 200, 150, 100, 50, and 10.

Figure 4 presents the results of the experiment on the MNIST dataset. It shows a slight improvement in accuracy when using ACL instead of FCL in the structure of the network. And the improved accuracy is maintained from 300 to 50 connections. And we confirmed that when the number of connections is less than 100, the accuracy decreases as the number of connections decreases. Considering the results, it is optimal to use ACL with 100 connections for the MNIST dataset.

The experimental result on the MNIST dataset showed that the accuracy when using adjacent-connection is slightly higher than that when using all-to-all connection. This phenomenon may appear somewhat unusual, but it may come from the insufficient performance of the learning framework. The framework employs a heuristic that does not search the entire solution space, but rather focuses on finding local optimum. Therefore, reducing the solution space by using ACL can yield better results, while further reduction ultimately led to worse results. If there was an ideal algorithm, all-to-all connection would always provide higher accuracy.

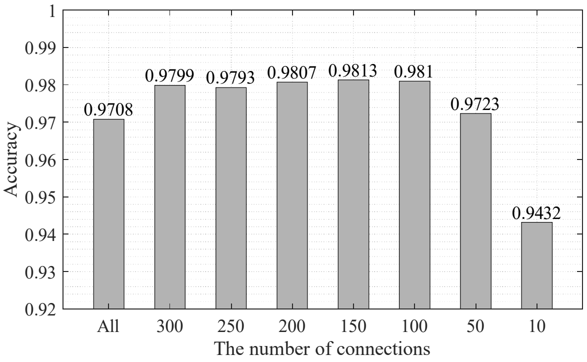


Figure 4. Inference accuracy on MNIST dataset

Figure 5 presents the results of the experiment on the Fashion-MNIST dataset. When compared to all-to-all connection, we can see that the accuracy is barely reduced until the number of connections is reduced by 200. From the results, in the case of Fashion-MNIST, the optimal number of adjacent-connection is 200 while maintaining accuracy.

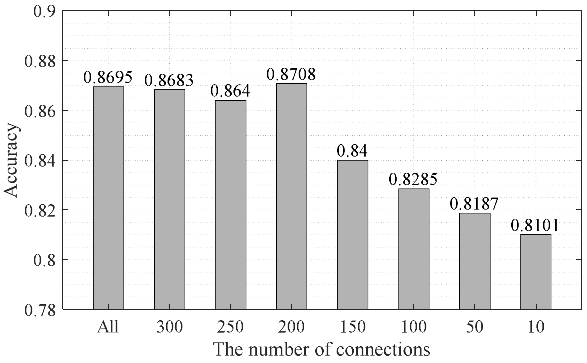


Figure 5. Inference accuracy on Fashion-MNIST dataset

As a result of comprehensive analysis of the experimental results for the two datasets, some of the SNNs using ACL have no loss of accuracy compared to all-to-all connection. And in some cases, we can find an optimal solution that reduces the number of connections while achieving better accuracy. Ultimately, we confirmed that applying ACL proposed in this paper can simplify and optimize the HW design without the loss of accuracy.

V. Conclusions

The manual layout of highly-congested wires complicates the implementation of SNNs. In order to mitigate the difficulties, we propose an adjacently-connected layer (ACL) optimized for embedded systems. By considering application characteristics, ACL effectively reduces the number of wire connections while minimizing the loss of accuracy. In the future, we plan to utilize the 2D features of images for the connections in order to use fewer wires.

Acknowledgment

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-01294, Development of IoT based edge computing ultra-low power artificial intelligent processor).

References

[1] M. Sarrafzadeh, F. Dabiri, R. Jafari, T. Massey, and A. Nahapetan, “Low power light-weight embedded systems,” in Proc. IEEE Int. Symp. Low Power Electron. Design (ISLPED), pp. 207–212, 2006.

[2] Q. Liu and Z. Zhang, “Ultra-low power always-on intelligent and connected SNN-based system for multimedia IoT-enabled applications,” IEEE Internet Things Journal, vol. 9, no. 17, pp. 15570–15577, Sep. 2022.

[3] A. Javanshir, T. T. Nguyen, M. A. P. Mahmud and A. Z. Kouzani, “Advancements in Algorithms and Neuromorphic Hardware for Spiking Neural Networks,” in Neural Computation, vol. 34, no. 6, pp. 1289-1328, May 19, 2022.

[4] T. Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste, "Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks," Journal of Machine Learning Research, vol. 22, no. 241, pp. 1-124, Sep. 2021.

[5] N. Rathi, P. Panda and K. Roy, "STDP-Based Pruning of Connections and Weight Quantization in Spiking Neural Networks for Energy-Efficient Recognition," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 38, no. 4, pp. 668-677, Apr. 2019.

[6] C. J. Schaefer, P. Taheri, M. Horeni and S. Joshi, "The Hardware Impact of Quantization and Pruning for Weights in Spiking Neural Networks," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 70, no. 5, pp. 1789-1793, May 2023.

[7] L. Deng, “The MNIST database of handwritten digit images for machine learning research [best of the Web],” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 141–142, Nov. 2012.

[8] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," arXiv:1708.07747, 2017.

[9] J. K. Eshraghian et al., “Training Spiking Neural Networks Using Lessons from Deep Learning,” in Proceedings of the IEEE, vol. 111, no. 9, pp. 1016-1054, Sept. 2023.

Hyeonguk Jang received his B.S. and M.S. degrees in electrical engineering from Gyeongsang National University, Jinju, South Korea, in 2013 and 2015, respectively, and his Ph.D. degree in ICT from the University of Science and Technology, Daejeon, South Korea in 2021. He is currently working at Electronics and Telecommunications Research Institute, Daejeon, South Korea, as a post doctor. His interests include spiking neural network, deep neural network and design of RISC-V based low-power embedded systems.

Jae-Jin Lee received the B.S., M.S., and Ph.D. degrees in computer engineering from Chungbuk National University, Cheongju, South Korea, in 2000, 2003, and 2007, respectively. He is a leader of the AI Edge SoC Research Section, Electronics and Telecommunications Research Institute, Daejeon, South Korea. His research interests include spiking neural network and RISC-V based embedded systems.

Kyuseung Han (Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering and computer science from Seoul National University (SNU), Seoul, South Korea, in 2008 and 2013, respectively. Since 2014, he has been with the Electronics and Telecommunications Research Institute, Daejeon, South Korea. His current research interests include reconfigurable architecture, network-on-chip and computer-aided design of low-power embedded systems.

Multivariate PCA-based Composite Criteria Evaluation Method for Anomaly Detection in Manufacturing Data

HyeokSoo Lee*,**, Youngki Jo**, Jongpil Jeong*

* Department of Smart Factory Convergence, Sungkyunkwan University, 2066 Seobu-ro, Jangan-gu, Suwon, 16419, Korea

** Research & Development Team, THiRA-UTECH Co., Ltd, CK Bldg 7, Hakdong-ro 5-gil Gangnam-gu, Seoul, 06044, Korea

huxlee@g.skku.edu, ykjo0712@thirautech.com, jpjeong@skku.edu

Abstract— **In recent years, manufacturing sites have become more intelligent and efficient by adopting various IT technologies. Among them, equipment and process abnormality detection is a topic of high interest for efficient factory operation. In this paper, we propose a method that can detect comprehensive abnormalities by utilizing the PCA algorithm, which is an unsupervised learning-based data analysis method that can easily analyze multivariate data and detect abnormalities in the data, the Hotelling T^2 method, which is suitable for multivariate data analysis, and the Box-Pierce statistical method to increase the detection criteria of abnormality detection data. To verify the effectiveness of the proposed method, experiments were conducted and validated using a chemical product production dataset. We expect that this method can be utilized for equipment and process anomaly detection in real-time at manufacturing sites.**

Keywords— **Multivariate Analysis, Anomaly Detection, Unsupervised Learning, Principal Component Analysis (PCA), Hotelling's T^2 Control Chart, Box-Pierce Test**

I. Introduction

Along with the Fourth Industrial Revolution, major developed countries are increasingly emphasizing the importance of manufacturing, and the importance of smart factories that combine information and communication technologies in production and manufacturing is also increasing. The use of information and communication technology in the manufacturing industry began in earnest in the 80s. In particular, industries with high investment costs, such as the semiconductor and display industries, have automated their manufacturing sites by actively introducing information and communication technologies because production efficiency is significant. Currently, factories in these industries are mostly operated as unmanned automated factories, and all facilities and manufacturing operation systems are organically connected. Factory automation is a form of automating repetitive tasks according to certain rules by utilizing IT technology, and the smart factory, which has been emphasized recently, aims to transform into intelligence beyond such automation [1].

Recent advances in IT technologies such as artificial intelligence, sensors and the Internet of Things (IoT), big data, and robots have made it possible to build a flexible and intelligent manufacturing environment beyond the existing level of automation. By utilizing real-time data generated by various facilities, sensors, and the Internet of Things during the production process, it is possible to monitor the status of processes and facilities in real-time, detect abnormal conditions, and prepare for upcoming risks or problems before they occur. There are limitations to diagnosing and taking action on manufacturing equipment and process defects by relying on the experience of those in charge of the manufacturing site or related experts. Therefore, to increase manufacturing efficiency and utilization, it is necessary to proactively predict equipment failures and process defects using manufacturing data and AI technologies [2].

In this paper, we proposed a method for detecting and predicting abnormalities in facilities or processes by analyzing time series data generated by facilities at manufacturing sites in real time using unsupervised learning. There may be many variables in the data generated by facilities, and multivariate analysis methods are used to comprehensively analyze these many variables to predict abnormal situations. Therefore, we proposed a "Multivariate PCA-based Composite Criteria Evaluation Method" that can effectively predict abnormal situations by using the Principal Component Analysis (PCA) algorithm, multivariate analysis technique Hotelling T^2 control chart, and Box-Pierce statistical method, and verified it through experiments.

The paper is organized as follows:

In Section 2, we reviewed the techniques for multivariate analysis and their popularity in anomaly detection. We also reviewed the basic concepts of PCA algorithm, Hotelling T^2 control chart, and Box-Pierce statistics. In Section 3, the proposed method "Multivariate PCA-based Composite Criteria Evaluation Method" is defined and described. In Section 4, we conducted experiments on two experimental datasets and described the experimental results along with visualization results. The conclusions and future research directions are described in Section 5.

II. Related Work

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

A. Multivariate Analysis

Multivariate analysis is a statistical method for simultaneously analyzing multiple variables on multiple measures and is an extension of univariate and bivariate analysis. It means performing multiple univariate analyses simultaneously, considering the dependent variables' statistical relationship [3].

There are the following types of multivariate analysis methods:

1) *Multiple Regression*: It is an analytical method that is appropriate when there is a relationship between one quantitative dependent variable and one or more quantitative independent variables. It allows you to predict changes in the dependent variable as a result of changes in multiple independent variables and to determine the extent to which the independent variables explain the dependent variable and the relative contribution of the independent variables to the dependent variable [4].

2) *Multiple Discriminant Analysis (MDA)*: A method for classifying samples when each sample falls into two or more groups and the number of groups is known in advance. This method is appropriate when there are multiple categories and multiple groups based on the dependent variable [5].

3) *Cluster Analysis*: Cluster analysis is a method for finding a taxonomy for samples with no information about the population. Unlike discriminant analysis, cluster analysis does not have a predefined group. It checks the similarity between each sample and other samples and determines whether to classify them into a group [6].

4) *Multivariate Analysis of Variance (MANOVA)*: It is a statistical method for determining the relationship between two or more categorical independent variables and multiple quantitative dependent variables and is an extension of univariate analysis of variance. Multivariate analysis of variance effectively tests hypotheses about each group's variance on two or more quantitative dependent variables [7].

5) *Factor Analysis*: Factor analysis is a statistical approach that analyzes the interrelationships among many variables to describe them in terms of a common underlying dimension. It is a way to compress the information contained in a large number of individual variables into a smaller set of variables with minimal loss of information. A typical factor analysis method is PCA [8, 17].

6) *Canonical Correlation Analysis*: Canonical correlation analysis is a logical extension of multiple regression analysis that investigates the relationship between one quantitative dependent variable and multiple quantitative independent variables. The basic principle is to find the linear combination of each variable group that maximizes the correlation between the dependent variable and the independent variable group and to find the set of weights of each variable group that maximizes the correlation between the dependent variable and the independent variable group [9].

B. Anomaly Detection

Anomaly detection is the search for objects or materials that exhibit unexpected patterns in data, and it is a method of creating models based on training data to find data with different characteristics from existing data. It can be used in a variety of fields, including security, medicine, finance, and manufacturing, to predict malicious behavior such as credit card fraud, cyber intrusions, and terrorist acts, as well as abnormal situations such as system failures, equipment breakdowns, and defective products. Anomaly detection is often confused with the problem of classification. Classification is about finding the boundary that separates two categories, while anomaly detection is about finding outlier data given multiple categories [10, 11].

For anomaly detection, there are the following methods:

1) *Statistical Methods*: Statistical methods are based on the assumption that outliers appear in the lower regions of the probability distribution. They fit a normal data model to the given data and use statistical inference to determine if new objects follow that model. The distribution you use depends on the complexity of your data. If your data is simple, you can use simple distributions like Normal, Poisson, or Multinomial. However, more complex models, such as mixed models or Hidden Markov Models, require iterative computations to estimate, which can take a long time depending on the convergence rate and criteria [10].

2) *Classification-based Methods*: Classification-based methods train data into one class or multiple classes depending on the number of labels and treat objects that do not correspond to a class as outliers. Representative classification-based method algorithms are Autoencoder, Bayesian network, Support Vector Machine (SVM) [12], and Decision Trees [13]. The Autoencoder algorithm has an input layer and an output layer with the same number of nodes. It can also have one or more hidden layers and uses an encoder to compress the input data and a decoder to restore the input data. If the error is large, it is judged as an abnormal situation [10].

3) *Clustering-based Methods*: Clustering-based methods can detect outliers with the following principles. The first is that outliers are those with a long distance from the center of the cluster. There are methods such as the Self-Organizing Map (SOM) [14] and k-Means [15] algorithm that cluster training data and compare test data with clusters to determine outliers. The second method is to determine whether an object is an anomaly by the size or density of the cluster to which it belongs. It is based on the assumption that normal belongs to large or dense clusters, while anomalies belong to small or sparse clusters [10].

4) ***Nearest Neighbor-based Methods*****:** Nearest neighbor analysis leverages the basic idea that normal data occurs in dense groups of data, while anomalies occur at a distance. Nearest neighbor anomaly detection methods need to measure the distance or similarity between two pieces of data. The distance or similarity between two pieces of data is calculated differently for continuous and categorical data. For continuous data, we usually use Euclidean distance, and for categorical data, we use a simple matching coefficient. Also, for multivariate data, we need to calculate the value of each attribute and then combine them. A typical method for nearest neighbor analysis is the k-nearest Neighbor algorithm [10, 16].

5) ***Spectral-based Methods*****:** Spectral-based methods are based on the assumption that data can be sent to a lower-dimensional subspace, where the normal and abnormal are sharply separated. A typical method is Principal Component Analysis [17]. When the data is divided into a low-dimensional space, normal objects that satisfy the correlation structure of the data have low values, and anomalous objects that deviate from the correlation structure have high values. It is suitable for processing high-dimensional data and can be used as a preprocessing for applying other techniques. However, it is only useful if the normal and outliers are properly separated in the low-dimensional space to which you send the data [10].

C. Principal Component Analysis (PCA)

Principal Component Analysis was first invented by Carl Pearson in 1901, but in the 1930s, Harold Hotelling, who was unaware of its existence, developed it separately and named it Principal Component Analysis [17]. PCA is a technique for transforming high-dimensional data into low-dimensional data, and it uses orthogonal transformations to transform data in high-dimensional space into low-dimensional space. For example, in the case of data with more than two dimensions, when the data is scaled on a single axis, the axis with the largest variance is called the first principal component, and the axis with the second largest variance is called the second principal component. Furthermore, the principal components after the first principal component are defined as having the largest variance under the constraint that they are orthogonal to the previous principal components. The important components are orthogonal because they are the eigenvectors of the covariance matrix. Principal component analysis is the simplest of the eigenvector-based multivariate analyses, and if a multivariate dataset appears to be simply a set of coordinates, it can be reduced to a lower dimension so that it can be analyzed effectively [17, 18].

If the data contains many variables, the dimensionality of the data may be large, resulting in slow learning and analysis speeds and inefficient analysis results. In this case, manifold learning reduces the dimensionality of the data for efficient learning and analysis. A manifold is a space where data exists, and data can exist in various dimensions. By reducing the dimensionality given by the data, you can view the data in a smaller dimension, requiring fewer parameters and making learning and analysis more efficient. However, on the learning side, while reducing dimensionality speeds up learning, it does not necessarily mean that reducing dimensionality improves the performance of the model. Therefore, it is necessary to determine the appropriate dimensionality by checking the results of the analysis. The most representative dimensionality reduction algorithm for manifold learning is PCA [17, 18].

The PCA algorithm proceeds in the following steps: [17, 18]

- Normalization is performed to center the axis without changing the data. Normalization scales all variables so that their mean is zero.
- Find the covariance matrix for the normalized data.
- Find the eigenvalues and eigenvectors for the calculated covariance matrix.
- The obtained eigenvalues are sorted in descending order to find the largest values.
- The eigenvector values corresponding to these large eigenvalues become the coefficient values that make up the principal components.
- Check the ratio of the Eigenvalues, determine which principal components are significant, and reduce the dimensionality by excluding the remaining principal components.

D. Hotelling's T^2 Control Chart

Hotelling's T^2 test, proposed by Harold Hotelling in 1947, is an extension of the t-test and a typical multivariate test method and assumes that the data is normally distributed and follows the F-distribution [19, 20, 21]. If data has p quality attributes and the number of observations is n, then the data has a p×n dimensional matrix structure. In this case, the statistics for individual observations are described by equation (1). [19, 20, 21]

$$T^2 = \left(X_i - \bar{X}\right)^T S^{-1}\left(X_i - \bar{X}\right) \tag{1}$$

The T^2 statistic takes into account the correlation between the independent variables by using the inverse of the covariance. Under the assumption that the data is normally distributed, the statistic of the Hotelling T^2 control chart follows the F distribution. This is equivalent to equation (2). [19]

$$T^2 \sim \frac{(n-1)p}{n-p} F_{(p,n-p)} \tag{2}$$

For a significance level α, the Upper Control Limit (UCL) value can be defined in equation (3).

$$\mathrm{U}CL = \frac{(n-1)p}{n-p} F_\alpha(p, n-p) \tag{3}$$

Hotelling's T^2 control chart assumes that normal data follows a normal distribution and uses a covariance matrix, and if Hotelling's T^2 statistic is higher than the control limit, it can also detect abnormal conditions. In particular, multivariate statistical methods are effective when multiple variables are highly correlated with each other. Therefore, multivariate statistical methods are a useful way to use multiple variables together to monitor processes more accurately and detect process abnormalities [19, 20, 21].

E. Box-Pierce (BP) Test

The Box-Pierce test was published in 1970 by G. E. P. Box and David A. Pierce [22]. It can be used in time series analysis to determine the existence of autocorrelation, and is a method for testing the null hypothesis "the data are independently distributed" and the alternative hypothesis "the data are not independently distributed and are time series correlated"[22, 23]. The Box-Pearce test is defined as the square of the covariance coefficient ρ for a given data set multiplied by time h and multiplied by the number of data n. The relevant equation is shown in equation (4) and (5). The Box-Pierce test follows a chi-square distribution [22, 23].

$$\hat{\rho} = \mathrm{X_t} - \mathrm{X_{t-1}} \tag{4}$$

$$Q = n\sum_{k=1}^{h} \hat{\rho}_k^2 \tag{5}$$

The Box-Pierce Q statistic is approximated from a chi-square distribution to an F distribution using the actual and predicted values. The value of the Box-Pierce Q statistic and the UCL value are defined by equations (6) and (7) below.

$$BP \sim \frac{(n^2-1)p}{n(n-p)} F(p, n-p) \tag{6}$$

$$UCL = \frac{(n^2-1)p}{n(n-p)} F_\alpha(p, n-p) \tag{7}$$

III. Proposed Method

This paper proposes a method to analyze unlabeled data to detect and predict anomalies. To do this, we need a method based on unsupervised learning and a method of multivariate data analysis. In the case of univariate data analysis, there is a problem that as the number of variables increases, the number of control charts also increases, and when detecting anomalies for each variable, the frequency is often too frequent to be significant. Therefore, a multivariate data analysis method is needed to improve efficiency by managing multiple variables with one control chart regardless of the number of variables and to detect abnormal situations that cannot be detected by univariate data analysis [3, 10, 11].

We reviewed methods for anomaly detection and multivariate data analysis in the previous section. In this paper, we first utilize PCA, which is commonly used for anomaly detection and multivariate data analysis, as the basic algorithm [18, 26, 27]. Based on the eigenvalues and eigenvector values of the covariance matrix obtained from the PCA algorithm calculation, multivariate analysis is performed. For this purpose, the Hotelling T^2 control method, which detects anomalies by considering the correlation between quality attribute values [24, 25], and the Box-Pierce Q statistical method, which checks for autocorrelation using the residuals between predicted and actual values in time series data, are used together. This method is defined as a "Multivariate PCA based Composite Criteria Evaluation Method". Figure 1 shows the detailed steps of the "Multivariate PCA-based Composite Criteria Evaluation Method".

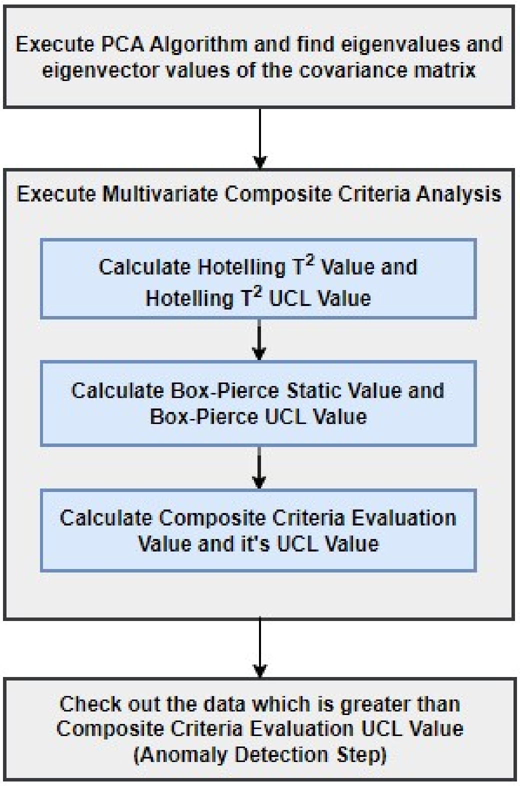

Execute PCA Algorithm and find eigenvalues and eigenvector values of the covariance matrix
Execute Multivariate Composite Criteria Analysis
Calculate Hotelling T² Value and Hotelling T² UCL Value
Calculate Box-Pierce Static Value and Box-Pierce UCL Value
Calculate Composite Criteria Evaluation Value and it's UCL Value
Check out the data which is greater than Composite Criteria Evaluation UCL Value (Anomaly Detection Step)

Figure 1. Multivariate PCA-based Composite Criteria Evaluation Method

We want to calculate the Composite Criteria Evaluation Value UCL values using a chi-square distribution, and since Hotelling T^2 UCL and Box-Pierce UCL are F-distributed, we approximate them with a chi-square distribution. The relevant equations can be found in (8) and (9).

$$T^2 \quad UCL = \frac{(n-1)p}{n-\mathrm{p}} F_\alpha(p, \mathrm{n}-p) \sim \chi_\alpha^2(p) \tag{8}$$

$$BP \quad UCL = \frac{(\mathrm{n}^2-1)p}{\mathrm{n}(\mathrm{n}-p)} F_\alpha(p, \mathrm{n}-p) \sim \chi_\alpha^2(p) \tag{9}$$

Equation (10) explains how the Composite Criteria Evaluation Value and corresponding UCL values are obtained. To simplify the formulas, we will denote the Hotelling T^2 statistic as α1, the Box-Pierce statistic as α2, the Hotelling T^2

UCL value as β1, and the Box-Pierce UCL value as β2. For the Composite Criteria Evaluation UCL, the Hotelling T2 UCL value and Box-Pierce UCL value obtained earlier are combined.

$$\text{Composite Criteria Evaluation Value} = \frac{\alpha 1}{\beta 1} + \frac{\alpha 2}{\beta 2} \qquad (10)$$

IV. Experiment and Results

The experiment was conducted using data from a chemical manufacturing process. A total of ten datasets were used for the experiment, and the labeled dataset was used to check the results of the analysis. In the case of unsupervised learning, the evaluation and adequacy of experimental results are limited and difficult, so we decided to use a labeled dataset to verify the analysis results. When learning of experiment, we removed the labels from the dataset conducted the experiment, and compared the experimental results with the label values in the dataset to check the accuracy.

The experiment was verified by first visualizing the raw data with the labels removed. The deviation of data values between data variables was too large to analyze the data immediately. Therefore, the data sets were scaled and normalized, and the normalized data was visualized to check the distribution. Then, we conducted experiments with the proposed method and checked the Hotelling T^2 statistic, Box-Pierce statistic, and related UCL values. The final Composite Criteria Evaluation Value and UCL values were then checked, and the accuracy of the analysis results was confirmed by comparing the obtained values with the label values. Figure 2 describes the analysis steps to run the experiment.

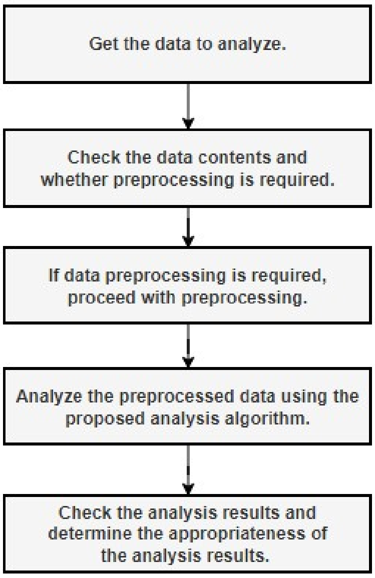

Get the data to analyze.
Check the data contents and whether preprocessing is required.
If data preprocessing is required, proceed with preprocessing.
Analyze the preprocessed data using the proposed analysis algorithm.
Check the analysis results and determine the appropriateness of the analysis results.

Figure 2. Processes for analyzing data

The experiment was performed with a total of 10 datasets. Each dataset consisted of 165 observations and included two abnormalities. Even though the datasets have the same number of data and the same number of abnormalities, they are all independent and different. In the experiment, the measurement results (labeled values) that can identify abnormalities were removed from the dataset and the experiment was conducted with unsupervised learning after the experiment, the experimental results were compared with the measurement results (labeled values) to check the accuracy of the analysis prediction. The following Table 1 summarizes the experimental results.

TABLE 1. Summary of Experimental Results

Data Set #	Number of data	Number of anomalies in the measurement	Number of anomaly predictions
1	165	2	2
2	165	2	1
3	165	2	2
4	165	2	1
5	165	2	2
6	165	2	2
7	165	2	2
8	165	2	2
9	165	2	2
10	165	2	3
Total	1,650	20	19

In the dataset perspective, 7 out of 10 predictions were correct, which is about 70% accuracy, but in the case of the data, 3 out of 1,650 predictions were incorrect, which is 99.82% accuracy. The following describes the detailed analysis procedure and results for two datasets that were correctly predicted

1) *Experiment #1*: The first data is a time series dataset with 7 variables and 165 measurements. The label values include 2 bad results. The raw data can be visualized in Figure 3 and the distribution of data is shown in Figure 4.

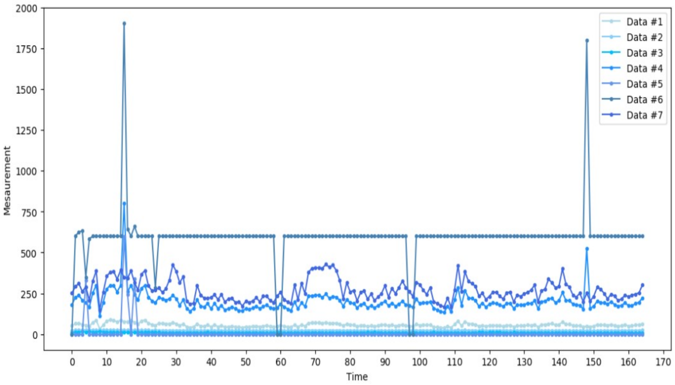


Figure 3. Raw data visualization results

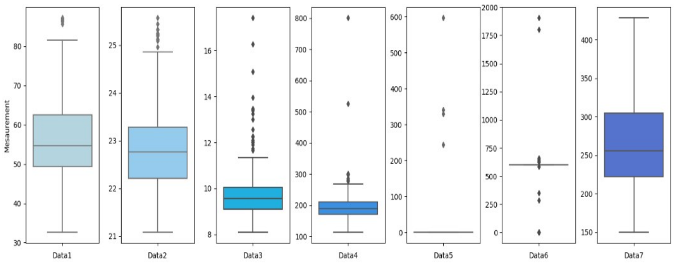


Figure 4. Distribution of raw data

When we visualized the raw data and its distribution, we found that the deviation between the data was considerable, so we normalized the data for efficient data analysis. Figures 5 and 6 show the distribution of the data after data normalization.

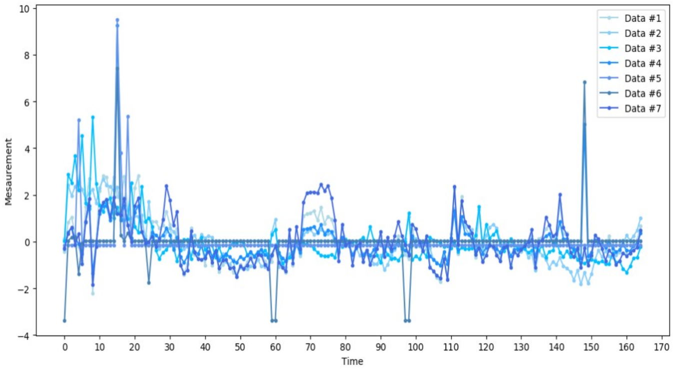


Figure 5. Normalized data visualization results

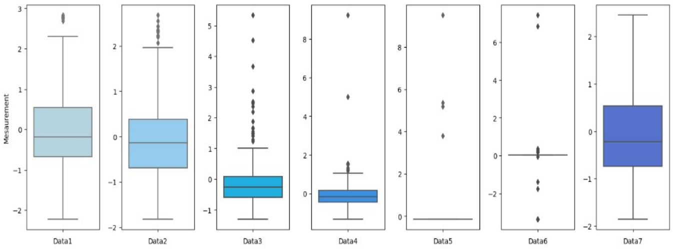


Figure 6. Distribution of normalized data

The proposed method, "Multivariate PCA-based Composite Criteria Evaluation Method", is executed to get Hotelling T^2 statistical value, Hotelling T^2 UCL value, Box-Pierce statistical value, and Box-Pierce UCL value. Figure 4 shows the results for these values.

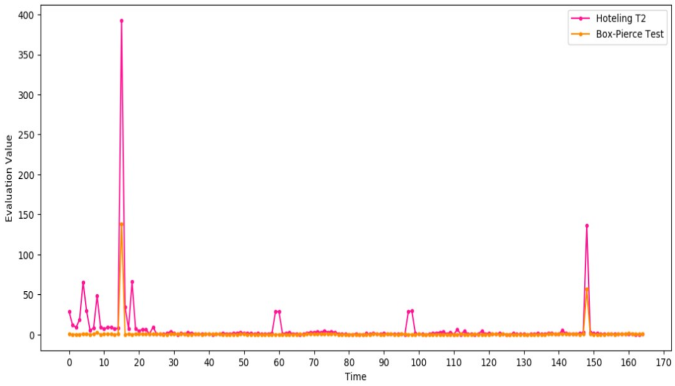


Figure 7. Hotelling T^2 and Box-Pierce statistics visualization results

Utilizing the Hotelling T^2 statistical value, Hotelling T^2 UCL value, Box-Pierce statistical value, and Box-Pierce UCL value values obtained earlier, we finally get the Composite Criteria Evaluation result and UCL value. Figure 5 shows the Composite Criteria Evaluation result and UCL value. The UCL value 15.2496315 is indicated by the red dotted line, and a total of two anomalies are predicted that exceed the UCL value.

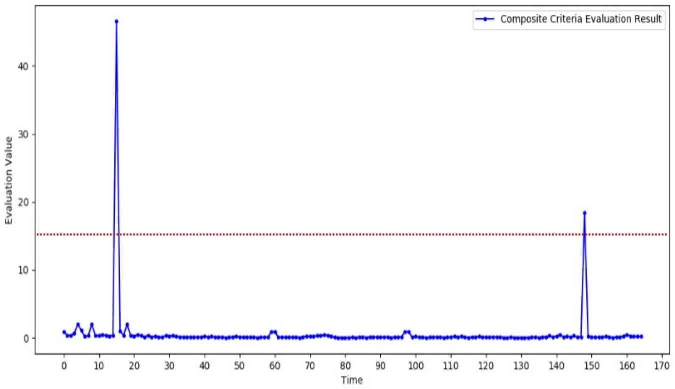


Figure 8. Composite Criteria Evaluation visualization results

The labeled values in the dataset are the results of post-production inspections, with most results between 10.00 and 11.00, and the average of the normal results in the dataset is 10.3933. The 16th and 149th inspection results in the dataset exceed this average criterion with values of 15.8655 and 15.8805, respectively, and can be categorized as abnormal. By checking the Composite Criteria Evaluation results analyzed by unsupervised learning with the proposed method, we can see that the two cases of exceeding the threshold UCL coincide with the points of abnormal conditions mentioned above.

*2) **Experiment #2**:* The second data is a time series dataset with 12 variables and 165 measurements. This dataset also has a total of 2 bad results in its labels. The raw data can be visualized in Figure 9 and the distribution of data is shown in Figure 10.

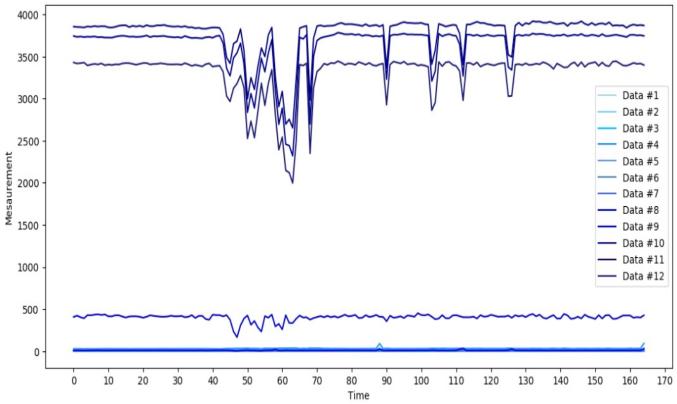


Figure 9. Raw data visualization fesults

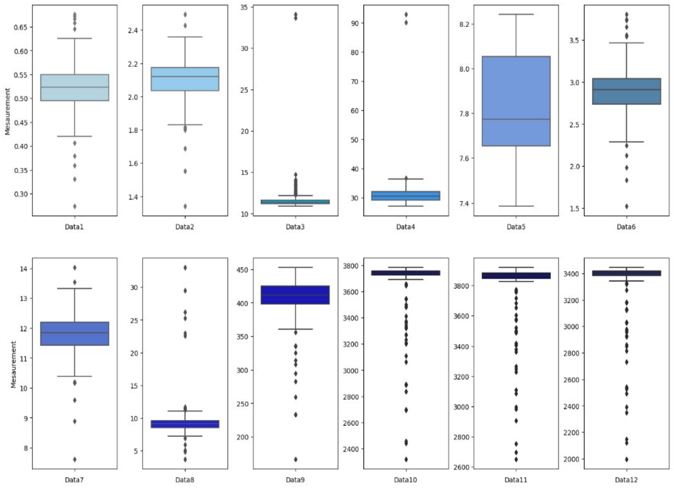


Figure 10. Distribution of raw data

When we checked the raw data, we found that this dataset also had a very large variance between the data, so we needed to normalize the data for efficient data analysis. Figures 11 and 12 show the distribution of the data after data normalization.

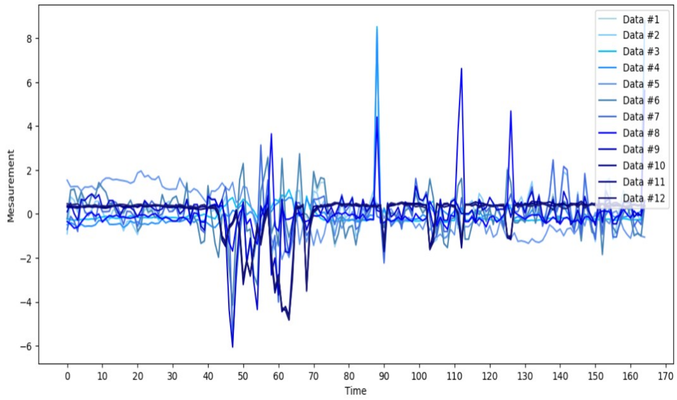


Figure 11. Normalized data visualization results

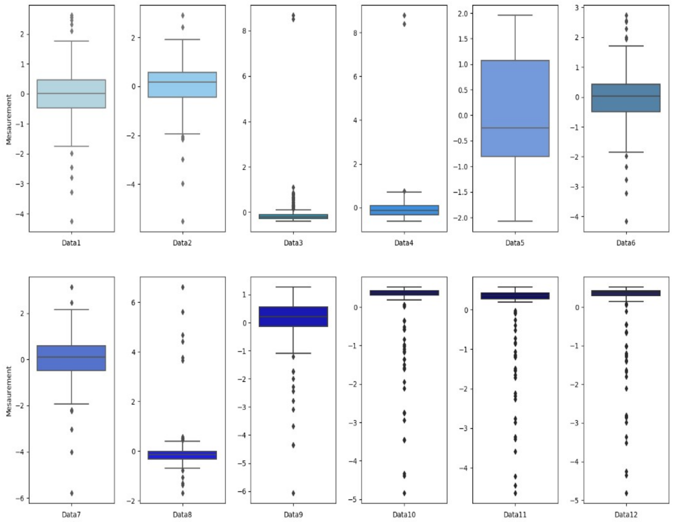


Figure 12. Distribution of normalized data

The proposed method is run and got Hotelling T^2 statistical value, Hotelling T^2 UCL value, Box-Pierce statistical value, and Box-Pierce UCL value. Figure 13 shows what they look like.

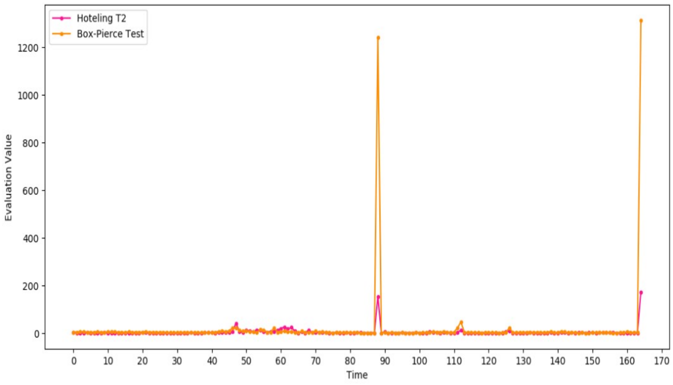


Figure 13. Hotelling T^2 and Box-Pierce statistics visualization results

Based on the Hotelling T^2 statistical value, Hotelling T^2 UCL value, Box-Pierce statistical value, and Box-Pierce UCL value values obtained earlier, we finally calculate the Composite Criteria Evaluation result and UCL value. Figure 14 shows their values. The UCL value 30.8467355 is marked with the red dotted line, and there are two anomalies are predicted.

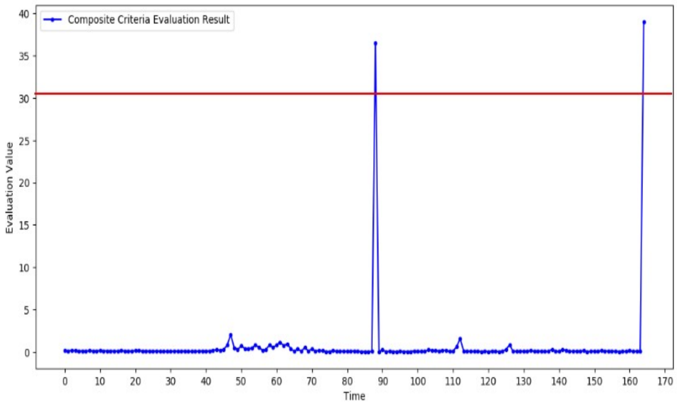


Figure 14. Composite Criteria Evaluation visualization results

The labeled values in the dataset are the results of post-production inspections, with most results between 10.00 and 11.00, and the average of the normal results in the dataset is 10.3932. The 89th and 165h inspection results in the dataset exceed this average criterion with values of 15.834 and 15.819, respectively, and can be categorized as abnormal. By checking the Composite Criteria Evaluation results analyzed by unsupervised learning with the proposed method, we can see that the two cases of exceeding the threshold UCL coincide with the points of abnormal conditions mentioned above.

V. Conclusions

The topic of this paper is research on how to detect abnormal situations by analyzing data generated in the manufacturing production process based on unsupervised learning. Most of these data are in the form of multivariate data with many variables. During machine learning, PCA, which is easy to analyze multivariate data and can detect abnormal situations in the data, is used for initial data analysis, and based on the analyzed results, the Hotelling T2 method, which is suitable for multivariate data analysis, and Box-Pierce statistical method, which is used to prevent abnormal situations from being detected too frequently in various variables, are used together to derive comprehensive abnormality detection results, and "Multivariate PCA based Composite Criteria Evaluation Method" is proposed. An experiment was conducted to verify the effectiveness of the proposed method. Experiments were conducted and verified using 10 datasets from a manufacturing plant that produces chemical products. In the case of the proposed algorithm, it is difficult to verify the results against the experimental results because it is in the form of unsupervised learning. So, we used datasets with measurement results that classified abnormal and normal states, removed the results, tested with datasets with only variables, and compared the experimental results with the measurement results of the dataset. The experimental results confirmed that 7 out of 10 datasets match exactly. In the case of mismatched datasets, one dataset detected only one of the two anomalies, and the other dataset detected anomalies correctly, but there was misrecognized a normal situation as an anomaly once. In terms of datasets, the accuracy was 70%, and in terms of data, the accuracy was about 99.82%. This method can be utilized for real-time abnormality detection of facilities and processes in the field. The detected anomalies can be shared with the manufacturing operation system, and the person in charge can check the equipment status or process variables and take early action if there is a problem. In the future, we would like to study how to increase the efficiency of manufacturing operations by more closely linking to the functions of the manufacturing operation system, how to separate the exact type of anomaly detection into equipment anomaly detection or process anomaly detection, and how to expand parameter management by linking anomaly detection with process or equipment parameters.

References

[1] R. Agnieszka, B. Arne, B. Marcel and S.M. Erik, “The Smart Factory: Exploring Adaptive and Flexible Manufacturing Solutions”, *Procedia Engineering*, vol. 69, 2014.

[2] Y. Shen, R.J. Juan and J. Yuchen, “Real-Time Monitoring and Control of Industrial Cyberphysical Systems: With Integrated Plant-Wide Monitoring and Control Framework”, *IEEE Industrial Electronics Magazine*, vol. 13, Dec, 2019.

[3] P. Luke and G. Hessameddin, *Multivariate Data Analysis*, 8th ed, Cengage Learning EMEA, 2018.

[4] M. Tucker and M. Brian, “Multiple Regression in L2 Research: A Methodological Synthesis and Guide to Interpreting R2 Values”, *The Modern Language Journal*, vol. 102, p 713-731, Dec, 2018.

[5] C. E. Brown, *Applied Multivariate Statistics in Geohydrology and Related Sciences*, 1st ed, Heidelberg: Springer-Verlag, Dec, 2011.

[6] J. R. Kettenring, “The Practice of Cluster Analysis”, *Journal of Classification*, vol. 23, Dec, 2006.

[7] F. L. Huang, “MANOVA: A Procedure Whose Time Has Passed?”, *SAGE Journals*, vol. 64, Dec, 2019.

[8] T. Mohsen and W. Angela, “Factor Analysis: a means for theory and instrument development in support of construct validity”, *International Journal of Medical Education*, vol. 11, Jun, 2020.

[9] Guangdong Liu, Shanshan Yang, Wei Liu, Shengshu Wang, Penggang Tai, Fuyin Kou, Wangping Jia, Ke Han, Miao Liu and Yao He, “Canonical Correlation Analysis on the Association Between Sleep Quality and Nutritional Status Among Centenarians in Hainan”, *Frontiers in Public Health*, vol. 8, Nov, 2020.

[10] Chandola. V, Banerjee and A. Kumar, “Anomaly Detection: A Survey”, *ACM Computing Surveys*, vol. 41, pp. 1–58, Nov. 2009.

[11] L. M. Ghinea, M. Miron and M. Barbu, “Semi-Supervised Anomaly Detection of Dissolved Oxygen Sensor in Wastewater Treatment Plants”, *MDPI Sensors*, vol. 23, Sep. 2023.

[12] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt and B. Scholkopf, “Support vector machines”, *IEEE Intelligent Systems and their Applications*, vol. 13, Jul-Aug, 1998.

[13] M. Oded and R. Lior, *Data Mining and Knowledge Discovery Handbook*, 1st ed, New York, US: Springer-Verlag, Dec, 2005.

[14] T. Kohonen, “The self-organizing map”, *Proceedings of the IEEE*, vol. 78, Sep, 1990.

[15] Claude Sammut, Geoffrey I. Webb, *Encyclopedia of Machine Learning*, 1st ed, New York, US: Springer, Nov, 2010.

[16] Oliver Kramer, *Dimensionality Reduction with Unsupervised Nearest Neighbors*, 1st ed, Heidelberg: Springer-Verlag, Nov, 2013.

[17] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed, New York, US: Springer-Verlag, 2002.

[18] L.E. Mujica, J. Rodellar, A. Fernа́ndez and A. Gu¨emes, “Q-statistic and T2-statistic PCA-based measures for damage assessment in structures”, *SAGE Journals*, vol. 10, Nov. 2010.

[19] National Institute of Standards and Technology (NIST) Engineer Statistics Handbook. [Online]. Available: https://www.itl.nist.gov/div898/handbook/pmc/section5/pmc543.htm/

[20] R. E. Schumacker, “Using R with Multivariate Statistics”, SAGE Publishing, 2015.

[21] M. R. Piña-Monarrez, “Practical decomposition method for T^2 hotelling chart”, *The International Journal of Industrial Engineering.*, vol. 20, pp. 401–411, Nov. 2013.

[22] M. Tucker and M. Brian, “The multiple testing problem for Box-Pierce statistics”, *Electronic Journal of Statistics*, vol. 8, pp. 497–522, Nov. 2013.

[23] WIKIPEDIA Ljung–Box test. [Online]. Available: https://en.wikipedia.org/wiki/Ljung%E2%80%93Box_test/

[24] T.Y Heo, H.B. Jeon, S.M. Park, Y.J. Lee, “Development of Real-Time Water Quality Abnormality Warning System for Using Multivariate Statistical Method”, *Journal of Korean Soc. Environ. Eng*, vol. 37, pp. 137–144, Feb. 2015.

[25] J.H. Kim, S.B. Kim, “Local T^2 Control Charts for Process Control in Local Structure and Abnormal Distribution Data”, *Journal of Korean Society for Quality Management*, vol. 40, pp. 337–346, Aug. 2012.

[26] A. Liang, Y. Hu, G. Li, “The impact of improved PCA method based on anomaly detection on chiller sensor fault detection”, *International Journal of Refrigeration*, Sep. 2023.

[27] T.H.K. Mohammed, K. Abdelmalek, F.H. Mohamed, B. Abderazak, M. Majdi, “Improving Kernel PCA-based algorithm for fault detection in nonlinear industrial process through fractal dimension”, *Process Safety and Environmental Protection*, vol. 179, pp. 525–536, Sep. 2023.

HyeokSoo Lee received a M.S. degree in Computer Science from Fairleigh Dickinson University, New Jersey US, in 1995. He is currently working toward a Ph.D. degree in smart factory convergence at Sungkyunkwan University, Suwon Korea. He is currently an employee as a head of the Research & Development team of THiRA-UTECH Co, Ltd. His research interests are mainly in Smart Factory, Reinforcement Learning, Deep Learning, Robotics, and Data Analysis.

Youngki Jo received the B.S. and M.S. degree in Industrial Engineering from Kumoh National Institute of Technology, Korea in 2018, and 2020. He is currently an employee as a research engineer of the Research & Development team of THiRA-UTECH Co, Ltd. His research interests include generative adversarial networks, deep learning, and time-series modeling and forecasting.

Jongpil Jeong received the bachelor’s degree in engineering from Sungkyunkwan University, Suwon, South Korea, and the master’s and Ph.D. degrees in computer engineering from Sungkyunkwan University, in 2003 and 2008, respectively. He has been with Sungkyunkwan University, since 2008, and has been an Associate Professor with the Department of Smart Factory Convergence, Sungkyunkwan University, since 2016. He is the Principal Investigator of MAKE UNIC, a key support area for smart manufacturing with Sungkyunkwan University. His research interests include smart factory, industrial AI, anomaly detection, manufacturing data analysis, AI-based fault diagnosis and prediction, 5G-based smart manufacturing, industrial IoT applications, AI platforms, cloud platforms, and industrial security.

Pitching-Motion: Pose-Based Pitch Trajectory Overlay System

Bor-Yao Tseng [a], Hung-Tse Chiang [a], Jiann-Liang Chen [a], Han-Chuan Hsieh [b],

[a] Department of Electrical Engineering, National Taiwan University of Science & Technology, Taipei, Taiwan
(Lchen@mail.ntust.edu.tw)

[b] Southern Region Campus, Industrial Technology Research Institute, Hsinchu, Taiwan
(hchsieh@itri.org.tw)

Abstract— **This study reviews the current practical use of the electronic strike zone in baseball, which typically features a fixed high-low strike zone. This system deviates from the baseball rules, which dictate that the strike zone should be adjusted according to the batter's shoulder, waist, and knee positions. Consequently, This study proposes a system based on commonly used baseball game camera perspectives. Using YOLO-Pose technology, it automatically detects players' body structures. It combines them with the baseball nine-grid positioning principle to promptly establish a strike zone that complies with baseball rules and player positions. The accuracy of this system, as measured by the Intersection over Union (IoU) metric, is 0.8541, representing a 14.29% improvement over the current electronic strike zone's IoU metric of 0.7473. Considering both viewers' perspective and sports analysis requirements, the system integrates an automated pitch trajectory detection system. It can overlay multiple pitch trajectories, allowing for a visual comparison of their differences. Combined with the strike zone detection system proposed in this study, it provides a more comprehensive view of the overlap between pitch trajectories and the strike zone. This study aims to enhance baseball officiating and analysis by introducing a more accurate and visually informative system for determining the strike zone and analyzing pitch trajectories.**

Keywords— **Pose estimation, Sports video, Ball Trajectory, Kalman Filter, Position measurement**

I. Introduction

The digitization of sports data is an increasingly popular and vital trend in sports. It enhances athletic performance and provides deeper insights, contributing to the improvement of game quality and athletes' training levels. Data research in the realm of sports is highly regarded [1]. Ball trajectory analysis has gained prominence in sports like tennis and badminton through systems such as Hawk-eye [2]. In baseball, the analysis of pitching trajectories is a crucial research area. Existing systems like TrackMan [3], PITCHf/x [4], and StatCast [5] are widely recognized and extensively utilized.

Moreover, innovations in trajectory analysis systems continue to emerge, exemplified by the integration of pitching trajectories with the electronic strike zone, as seen in K Zone [6]. This integration provides a clear and comprehensible representation of the relationship between pitching trajectories and the strike zone. Such systems exhibit a high level of accuracy in tracking pitching trajectories. However, the corresponding strike zone remains fixed, while an accurate strike zone should vary based on the height of the batter's shoulder, waist, and knees, as stipulated by baseball rules. According to these rules, the lower boundary of the strike zone corresponds to the height of the batter's knees, while the upper boundary aligns with the height of the batter's shoulders and the center of the waistband. In order to enhance the alignment of electronic strike zone data with baseball regulations, this study introduces a computational mechanism and framework as a solution. Considering the batter's physique, it automatically detects the strike zone from images captured from common game perspectives. It overlaps different pitch trajectories from the same pitcher and generates images that comprehensively depict the differences between each pitch and its alignment with the strike zone data.

II. Related work

The digitization of sports data through technology has become commonplace, and it can effectively enhance the fairness of sports, which is why Hawk-eye [2] has become standard equipment for significant tennis competitions. In baseball, errors are inevitable because the home plate umpire relies solely on the naked eye to judge balls and strikes [7]. Various factors on the field can also influence the umpire's strike zone [8]. In recent years, baseball has also been experimenting with and testing systems similar to Hawk-eye, such as the electronic strike zone, which is currently being tested in Minor League Baseball in the United States. It is based on the high-precision pitch tracking analysis of TrackMan [3] to determine whether the ball has entered the strike zone. However, the current electronic strike zone has a fixed height and does not conform to the actual rules of baseball.

In current research related to automated strike zone analysis, Chen, HT. et al.[9] proposed using image processing to outline the batter's silhouette to analyze the positions of the batter's shoulders, waist, and knees. This data is used to determine the strike zone. However, this system does not meet the current market requirements regarding stability, execution speed, and automation. Y. Kanno et al. [10] automated the detection of the batter's shoulder, waist, and knee joint positions using deep learning with OpenPose [11]. Their research focused on analyzing the movement of the catcher's glove during the catching process. The strike zone detection requires significant preprocessing on the input images, such as manually selecting the batter's batting zone. However, several aspects of their research are worth considering, such as the automation of deep learning detection and the detection of the baseball to determine the timing of the batter's batting stance. This study builds upon these foundations to improve the automation and data accuracy of strike zone detection while expanding its applicability.

The strike zone aims to determine whether the pitch result is a strike or a ball. Therefore, electronic strike zones need to be complemented by pitch-tracking technologies to demonstrate their effectiveness effectively. In order to achieve a complete 3D reconstruction of the entire batting area, points that are obscured or cannot be precisely captured in some views are estimated using re-projection to determine their 2D pixel coordinates. In baseball pitch tracking research, Labayen et al. [12] used high-speed cameras to detect the difference between the baseball and the background images through image processing to track the baseball's trajectory. H. Lee et al. [13] used two cameras for calibration and placed pillars with grids on the field to precisely calculate the pitch trajectory based on the positional relationship between the baseball in the image and the pillars. Bor-Jiunn Wen et al. [14] used YOLOv3-Tiny to detect the baseball trajectory automatically and could calculate pitch speed and spin rate from a single-angle image. H. -S. Chen et al.[15] detected the baseball trajectory based on brightness, size, and appearance in the image, and automatically extracted pitching and hitting video clips from game footage based on trajectory features, making pitch analysis more accessible. Combining pitch trajectories with electronic strike zones can visualize sports data and has significant potential for both commercial and research applications[16][17][18]. It can also enhance the viewer experience in 3D replays on the field [19] [20]. However, more data does not necessarily mean better viewer experience, as simplicity and clarity are often more appealing to the general audience [21]. Therefore, in this study, deep learning techniques are used to automatically detect pitch trajectories and overlay multiple pitch trajectories with the electronic strike zone to provide a clear and intuitive representation of the differences between pitch trajectories and the relationship with the strike zone.

III. STRIKE ZONE DETECTION

This study employs YOLO-Pose [22] technology for detecting human body poses in images. By identifying critical skeletal points in the human body, it aims to determine the batter's strike zone. As Figure 1 of the game footage shows, recognizing the human body poses skeletal points for the pitcher, batter, catcher, and umpire. In the game footage, since the pitcher's posture faces away from the camera, the skeletal point corresponding to the nose can be used to exclude the pitcher. Considering that the batter typically occupies the outermost region among the batter, catcher, and umpire, this study utilizes the distances between skeletal points to determine those at the outermost periphery.

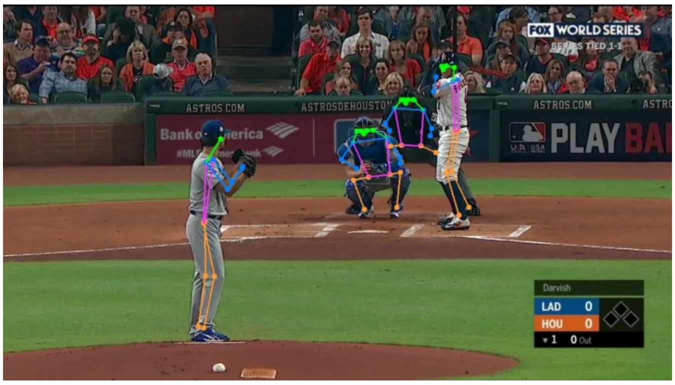
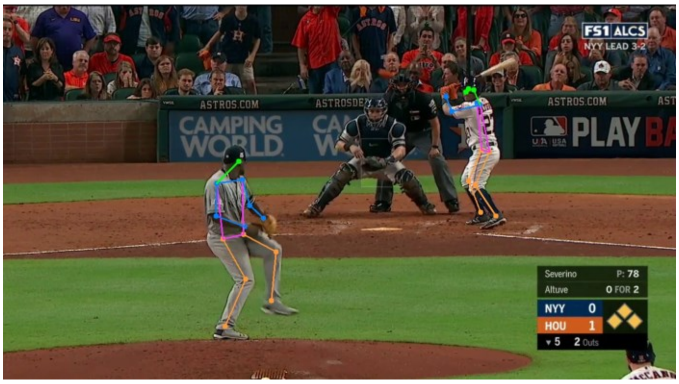


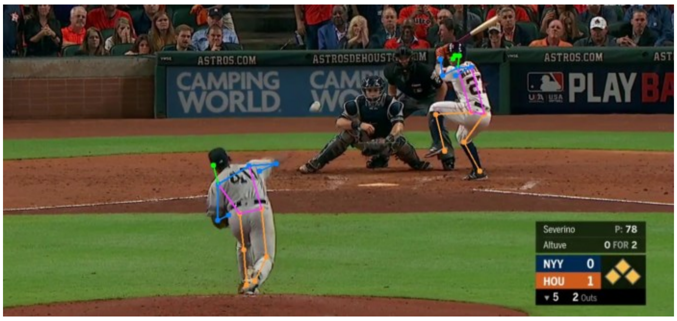
Figure 1. Illustration of Actual Game Footage

By comparing the heights of the detected outermost individuals' nose or ear coordinates, it is possible to identify whether a skeletal point belongs to the batter using the highest coordinate point. Determining the batter's strike zone relies on the positions of the batter's shoulders, the upper edge of their pants, and their knee. Therefore, identifying the timing to assess the strike zone is crucial. In this study, we divide the process after the pitcher releases the baseball into two key moments, namely, the Release Frame and the Prepare Frame, as illustrated in Figure 2. The Release Frame is when the pitcher is about to release or has just released the baseball. The batter has already started preparing for the hitting motion and has assumed the corresponding hitting stance. We use YOLOv7 [23] to detect the initial frame of the pitched baseball to determine the exact timing of the Release Frame. The Prepare Frame is defined as the frame within the 20 frames preceding the Release Frame, where the batter's knee position is the lowest. This moment represents the batter in a standing position and provides more accurate information about the positions of the batter's shoulders, the upper edge of their pants, and their knees.

Once the Prepare Frame is detected, the upper limit of the strike zone is defined as the midpoint between the shoulder and hip skeletal points detected in the Prepare Frame. In contrast, the lower limit is the lower edge of the knee skeletal point. The width of the strike zone is determined manually by marking the width of the home plate, as shown in Figure 3. This strike zone, customized based on the batter's height, provides a more accurate basis for determining a good pitch.



(a) Prepare Frame



(b) Release Frame

Figure 2. Illustration Frames

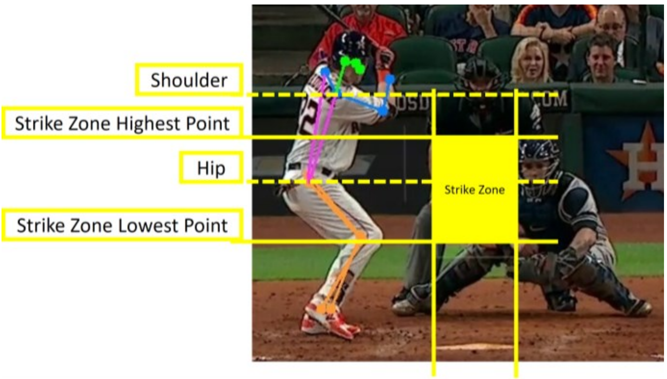


Figure 3. Strike Zone Diagram Based on Height

IV. BASEBALL PITCH OVERLAY

This study combines YOLOv7 and the Kalman filter (KF) to enhance the accuracy of flight trajectory prediction. The Kalman filter is a recursive algorithm applicable to linear systems, effectively addressing Gaussian noise issues induced by image processing. The system can also integrate multiple flight trajectories with the results of human body pose recognition and display them in the image. This provides coaches and pitchers with a better visual experience of flight trajectories.

The operation of the Kalman filter involves estimating the system's state by combining the system model and observed data. It consists of two main stages: prediction and update. In the prediction stage, the current state is predicted based on the previous moment's state as a reference. The predicted state is corrected in the update stage based on the current observed values. Through this recursive process, the accuracy of the estimation continuously improves. The critical assumptions of the Kalman filter are that both the system state and observations follow Gaussian distributions and have a linear relationship. Its advantages include efficient computation and precise estimation.

The state variable $\mathbf{x_k}$ and observation variable $\boldsymbol{z_k}$ refer to the variable settings in [24], which are used to represent the characteristics of the system state and observations. The forms of the state variable $\mathbf{x_k}$ and observation variable $\boldsymbol{z_k}$ are as follows:

$$\boldsymbol{x_k} = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}]^T \quad (1)$$

$$\boldsymbol{z_k} = [u, v, s, r]^T \quad (2)$$

The variable k in the state variables represents different stages of the system. At the same time, u and v denote the ball's pixel positions in the horizontal and vertical directions. s signifies the area of the ball, and r represents the aspect ratio of the ball. Furthermore, $\dot{u}$, $\dot{v}$ and $\dot{s}$ correspondingly denote their velocities.

In the Kalman filter prediction stage, we assume that the state at time step k can be derived from the state at time step k-1. For the constant velocity motion of the baseball considered in this study, we can express the state equation as follows:

$$\boldsymbol{x_k} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \boldsymbol{x_{k-1}} \quad (3)$$

The purpose of the observation equation is to describe the relationship between the state variable $\boldsymbol{x_k}$ and the observation variable $\boldsymbol{z_k}$. Its mathematical form is as follows:

$$\boldsymbol{z_k} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \boldsymbol{x_k} \quad (4)$$

The state variable $\boldsymbol{x_k}$ and observation variable $\boldsymbol{z_k}$ can be represented by the following two sets of equations:

$$\boldsymbol{x_k} = F_k \boldsymbol{x_{k-1}} + B_k \boldsymbol{u_k} + \boldsymbol{w_k} \quad (5)$$

$$\boldsymbol{z_k} = H_k \boldsymbol{x_k} + \boldsymbol{v_k} \quad (6)$$

F_k represents the state-transition matrix used to calculate the predicted state from past estimations. B_k is the control input model, and vector $\boldsymbol{u_k}$ denotes the controller input. $\boldsymbol{w_k}$ represents process noise. H_k is the observation matrix used to compute the predicted measurement based on the predicted state. $\boldsymbol{v_k}$ represents measurement noise.

The prediction stage and update stage can be expressed through the following seven equations:

$$\boldsymbol{x_{k|k-1}} = F_k \boldsymbol{x_{k-1}} + B_k \boldsymbol{u_k} \quad (7)$$

$$P_{k|k-1} = F_k P_{k-1|k-1} {F_k}^T + Q_k \quad (8)$$

$$\boldsymbol{y_k} = \boldsymbol{z_k} - H_k \boldsymbol{x_{k|k-1}} \quad (9)$$

$$S_k = H_k P_{k|k-1} {H_k}^T + R_k \quad (10)$$

$$K_k = P_{k|k-1} {H_k}^T {S_k}^{-1} \quad (11)$$

$$\boldsymbol{x_{k|k}} = \boldsymbol{x_{k|k-1}} + K_k \boldsymbol{y_k} \quad (12)$$

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} \quad (13)$$

Equations (7) and (8) represent the prediction stage, where P_k stands for the posterior estimation covariance matrix, and Q_k denotes the process noise covariance. The equations for the update stage are (9), (10), (11), (12), and (13), with R_k representing the observation noise covariance, $\boldsymbol{y_k}$ representing the innovation or measurement pre-fit residual, S_k representing the innovation covariance, and K_k representing the optimal Kalman gain.

V. Experiment

The dataset utilized in this study is MLB-YouTube[25], which comprises 20 baseball game videos from the 2017 MLB postseason. These videos have a frame rate of 30 frames per second and a resolution of 1280 pixels by 720 pixels. The dataset is divided into two primary sections: activity recognition with segmented videos and activity classification with continuous videos. For this study, we selected the segmented videos used for activity recognition, from which segments containing successful pitches and batters performing hitting actions were extracted, as detailed in Table 1.

In our experiments, we employed YOLOv7[23] technology to identify the positions of Prepare frames, Release frames, and the baseball in the videos. Additionally, we utilized YOLO-Pose[22] as a human pose recognition technique to draw the strike zone based on the batter's height and to depict multiple trajectories of pitched baseballs. These visual elements were embedded into the videos, providing a visualization tool for coaches and pitchers to examine the relationships between pitch types and strike zones.

TABLE 1. Experimental Dataset

	Videos	Frame Numbers
MLB-YouTube[25]	101	14371

This study employed Intersection over Union (IoU) as a metric to measure the accuracy of human pose-based strike zone representation. IoU is a commonly used indicator for evaluating object detection accuracy and is widely employed in deep learning tasks, particularly in object detection. Its calculation involves two parameters: the first is the ground truth target range manually annotated in the dataset, and the second is the target range derived from the algorithm used in the experiment. IoU is computed by measuring the overlap between these target ranges and their union, then dividing it by the union to obtain the IoU value. The established ground truth by manually drawing the strike zone in the Prepare frame. Subsequently, this manually drawn strike zone was compared with the strike zone computed using the human pose recognition technique to calculate the IoU. This metric reflects the correlation between the ground truth and predictions.

Traditional electronic strike zones in televised baseball broadcasts are based on images captured by two high-speed cameras, which are used to calculate 3D coordinate space through triangulation techniques. This 3D strike zone is then displayed in the broadcast. The electronic strike zone technology currently uses a fixed-size strike zone without considering the batter's height factor. This study measured the accuracy of the strike zone drawn based on human pose recognition and tested the accuracy of the electronic strike zone. The testing methodology is the same as that used for drawing the strike zone based on human pose recognition, and the accuracy of the strike zone is shown in Table 2. The human pose recognition strike zone method proposed in this study demonstrated a 14.29% improvement in accuracy compared to the electronic strike zone. The comparison between the strike zones drawn based on human pose recognition, electronic strike zones, and the ground truth is depicted in Figure 4.

TABLE 2. Accuracy of the Strike Zone

	IoU
YOLO-Pose Strike Zone	0.8541
Electronic Strike Zone	0.7473

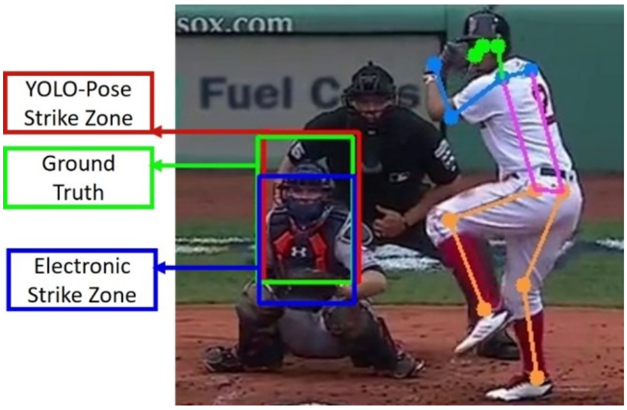


Figure 4. Illustrating the Assessment Range of the Strike Zone

This study goes beyond merely drawing a strike zone based on the batter's height; it also processes the trajectories of pitched baseballs, enhancing their accuracy using the Kalman filter. It allows us to precisely overlay the trajectories of baseballs pitched by the same pitcher against different batters in the same game and combine them with the respective batter's strike zone. It provides pitchers and coaches with a more accessible means of analyzing the different pitch types that pitchers deliver to different batters in the opposing team and how variations in batter height affect the effectiveness of the strike zone. It enables pitchers and coaches to adjust pitch types and strategies based on the characteristics of different batters. Such fine-grained analysis is expected to assist teams in making more strategic decisions, enhancing their competitiveness. The outcomes of this study are depicted in Figure 5.

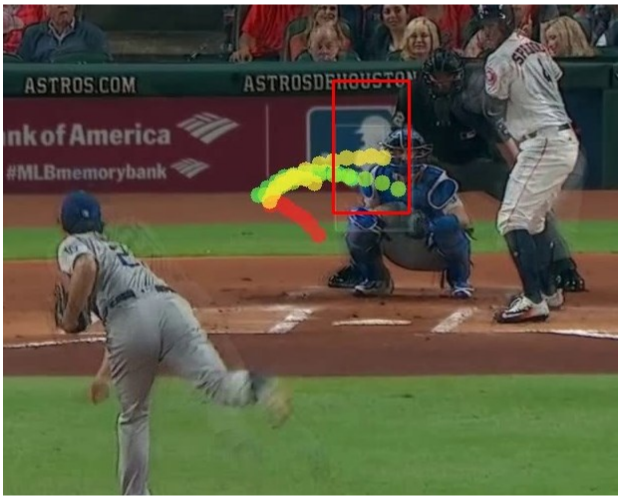


Figure 5. Illustrating Overlapping Baseball Trajectories and Strike Zone

VI. Conclusions

In this study, a review of current electronic strike zone systems and strike zone detection research was conducted. A novel strike zone detection system, YOLO-Pose Strike Zone, was proposed to address the issue of electronic strike zones not conforming to baseball rules in practical applications. The YOLO-Pose Strike Zone achieved an IoU (Intersection over Union) score of 0.8541, a 14.29% improvement in accuracy compared to the current electronic strike zone system with an IoU score of 0.7473. Furthermore, it enhances detection efficiency through automation. In terms of visualizing pitch trajectories, this study considered both the viewer experience and practicality. Overlapping pitch trajectories provides an intuitive way to visualize differences between pitch trajectories and their relationship with the strike zone. This system offers a practical and feasible approach for electronic strike zone-related systems, ensuring that strike zone values align more closely with baseball rules. Additionally, it has the potential for future use in data visualization methods such as 3D replay systems on the field, ultimately enhancing the viewer experience and increasing the precision of baseball data analysis. In the future, we can integrate the data from this system with 3D modelling technology to accurately recreate the pitching trajectory. Coupled with a 3D strike zone model, this will allow us to visualize the fine details of the pitching trajectory with greater precision.

References

[1] G. Thomas, R. Gade, T.B. Moeslund, P. Carr and A. Hilton, "Computer vision for sports: current applications and research topics", Comput. Vis. Image Underst., vol. 159, pp. 3-18, June 2017.

[2] N. Owens, C. Harris and C. Stennett, "Hawk-eye tennis system," Proceedings of the International Conference on Visual Information Engineering VIE 2003, pp. 182-185, Guildford, UK, 2003.

[3] J.J. Martin, "Evaluation of doppler radar ball tracking and its experimental uses, " 2012.

[4] M. Fast, "What the heck is PITCH f/x?" The Hardball Times Baseball Annual, pp. 153-158, 2010.

[5] M. Lage, J. P. Ono, D. Cervone, J. Chiang, C. Dietrich and C. T. Silva, "StatCast Dashboard: Exploration of Spatiotemporal Baseball Data," Proceedings of the IEEE Computer Graphics and Applications, Vol. 36, No. 5, pp. 28-37, Sept.-Oct. 2016.

[6] A. Gueziec, "Tracking pitches for broadcast television," in Computer, vol. 35, no. 3, pp. 38-43, March 2002.

[7] Dale L. Zimmerman. Jun Tang. Rui Huang. "Outline analyses of the called strike zone in Major League Baseball," Ann. Appl. Stat. pp. 2416 - 2451, December 2019.

[8] Jyhhow Huang & Hwai-Jung Hsu "Approximating strike zone size and shape for baseball umpires under different conditions," International Journal of Performance Analysis in Sport, pp. 133-149, 2020.

[9] Chen, HT., Tsai, WJ. & Lee, SY. "Contour-based strike zone shaping and visualization in broadcast baseball video: providing reference for pitch location positioning and strike/ball judgment," Multimed Tools Appl, pp. 239–255, 2010

[10] Y. Kanno, H. Shishido, M. Shinya, Y. Kameda and I. Kitahara, "Detection of Mitt Movement Trajectory in Catcher Framing Using Baseball Video," 2022 IEEE 11th Global Conference on Consumer Electronics (GCCE), pp. 139-143, Osaka, Japan, 2022.

[11] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields", IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172-186, 2021.

[12] Labayen, M., Olaizola, I.G., Aginako, N. et al. "Accurate ball trajectory tracking and 3D visualization for computer-assisted sports broadcast," Multimed Tools Appl, pp. 1819–1842, 2014.

[13] H. Lee, J. Kim, J. Kim and W. -Y. Kim, "A Method of Measuring Baseball Position at the Strike Zone," 2020 International Conference on Electronics, Information, and Communication (ICEIC), Barcelona, Spain, pp. 1-3, 2020.

[14] Bor-Jiunn Wen, Che-Rui Chang, Chun-Wei Lan, and Yi-Chen Zheng, "Magnus-Forces Analysis of Pitched-Baseball Trajectories Using YOLOv3-Tiny Deep Learning Algorithm," Applied Sciences 12, no. 11: 5540, 2022.

[15] H. -S. Chen, H. -T. Chen, W. -J. Tsai, S. -Y. Lee and J. -Y. Yu, "Pitch-by-Pitch Extraction from Single View Baseball Video Sequences," 2007 IEEE International Conference on Multimedia and Expo, pp. 1423-1426, Beijing, China, 2007.

[16] Du, M., Yuan, X. "A survey of competitive sports data visualization and visual analysis," J Vis, pp. 47–67, 2021.

[17] V. Bhatt, U. Aggarwal and C. N. S. V. Kumar, "Sports Data Visualization and Betting," 2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, pp. 1-6, 2022.

[18] Huang, J.-H., & Hsu, Y.-C. "A Multidisciplinary Perspective on Publicly Available Sports Data in the Era of Big Data: A Scoping Review of the Literature on Major League Baseball," SAGE Open, 2021.

[19] C. Dietrich, D. Koop, H. T. Vo and C. T. Silva, "Baseball4D: A tool for baseball game reconstruction & visualization," 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), Paris, France, pp. 23-32, 2014.

[20] M. Lage, J. P. Ono, D. Cervone, J. Chiang, C. Dietrich and C. T. Silva, "StatCast Dashboard: Exploration of Spatiotemporal Baseball Data," in IEEE Computer Graphics and Applications, vol. 36, no. 5, pp. 28-37, Sept.-Oct. 2016.

[21] Zheng, Meng-Cong, and Chih-Yung Chen, "Types of Major League Baseball Broadcast Information and Their Impacts on Audience Experience," Informatics 9, no. 4: 82, 2022.

[22] D. Maji, S. Nagori, M. Mathew and D. Poddar, "YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss," Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2636-2645, New Orleans, LA, USA, 2022.

[23] C. Y. Wang, A. Bochkovskiy and H. Y. Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7464-7475, 2023.

[24] A. Bewley, Z. Ge, L. Ott, F. Ramos and B. Upcroft, "Simple online and realtime tracking," Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), pp. 3464-3468, Phoenix, AZ, USA, 2016.

[25] A. Piergiovanni and M. S. Ryoo, "Fine-Grained Activity Recognition in Baseball Videos," Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1821-18218, Salt Lake City, UT, USA, 2018.

Bor-Yao Tseng was born in Taiwan, in 1999. He received the B.S. degree from Fu Jen Catholic University (FJCU), in 2022. He is currently pursuing the M.S. degree in electrical engineering with the National Taiwan University of Science and Technology (NTUST), Taipei. His main research interests include Artificial Intelligence, Internet of Things (IoT), Sports data analysis, Image Stitching, and Cybersecurity.

Hung-Tse Chiang was born in Taiwan, in 2000. He received the B.S. degree from National Taiwan University of Science and Technology (NTUST), in 2022. He is currently pursuing the M.S. degree in electrical engineering with the National Taiwan University of Science and Technology (NTUST), Taipei. His main research interests include Artificial Intelligence, Internet of Things (IoT), Sports data analysis, Image Stitching, and Cybersecurity.

Jiann-Liang Chen Prof. Chen was born in Taiwan on December 15, 1963. He received the Ph.D. degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan in 1989. Since August 1997, he has been with the Department of Computer Science and Information Engineering of National Dong Hwa University, where he is a professor and Vice Dean of Science and Engineering College. Prof. Chen joins the Department of Electrical Engineering, National Taiwan University of Science and Technology, as a Distinguished professor and Dean now. His current research interests are directed at cellular mobility management, cybersecurity, personal communication systems and Internet of Things (IoT).

Han-Chuan Hsieh received the BSEE degree from National Taipei University of Technology (NTUT) in 1998, and the MS degree in Communication Engineering from Tatung University (TTU), Taipei, Taiwan, in 2008. He received his PhD in EE from National Taiwan University of Science and Technology (NTUST) in 2018. He joined ICL, ITRI in 2012, working on experimental research that mainly covers 4G Networks. He has recently devoted himself to the integration of 5G networks, Computer Vision and IoT applications.

A Review of Detection-related Multiple Object Tracking in Recent Times

Suya Li[†], Ying Cao[*], Xin Xie [†]
[†]Henan University, School of Computer and Information Engineering, Jinming Campus, Kaifeng, China
[*]Corresponding Author
Email: henu_work_cy@163.com
Henan University,
Henan Key Laboratory of Big Data Analysis and Processing,
School of Computer and Information Engineering, Jinming Campus, Kaifeng, China

Abstract—**Multi-object tracking (MOT) is garnering more attention due to its widespread application in the area of autonomous driving, human-computer interaction, and intelligent video surveillance. Especially in recent years, MOT has rapidly developed thanks to related technologies such as object detection, which has helped in handling interfering factors such as crowded scene occlusion, small objects, and similar appearances. Among these, Detection-based MOT is the mainstream for accurately forming objects' trajectories. Therefore, according to the analysis of the last three years' research, this paper particularly focuses on discussing the continuous optimization strategies of MOT around the development of object detection at each stage. In addition, this article also introduces the commonly used benchmark datasets and related applications of MOT.**

Index Terms—**Multi-object Tracking (MOT), Object detection, Deep learning, Research progress**

I. Introduction

With the development of Deep-learning(DL)-based Multi-object tracking (MOT), it has been widely utilized in many related real-life problems like Intelligent transportation[1], Video surveillance[2], autonomous driving[3], human-computer interaction[4]. These applications also highlight the important academic significance of MOT. To accomplish the task of MOT, a common way is locating objects' position frame per frame first and associating them inter frames then. There, the tracking objects could be anything such as pedestrians, sports players, birds, dogs, vehicles, or all of the above. Due to the urgent need for continuous improvement of the applications performance and the challenges such as occlusion, small objects, and similar appearances, the schemes in MOT are always being optimized. Among them, detection-related multi-object tracking methods are the mainstream solution for forming the trajectories of objects[5, 6, 7, 8].

Therefore, different from related review papers on MOT in previous years[9, 10, 11], this paper mainly focuses on object detection-related multi-object research, and through the analysis of the methods to divides them into detection-based MOT, joint detection-based MOT. The core ideas and algorithmic features of each technique are elaborated on separately.

Consequently, to sum up, the work of this paper is organized as follows:

1) The development summary of standard object detection algorithms used in MOT;

2) The list of tracking-by-detection and joint detection and tracking MOT approaches;

3) The overview of MOT datasets;

4) The exploration of various applications.

II. Object Detection For MOT

As the fundamental technology in Computer vision, object detection has been used in much other vision tasks[12] and the related vision application[2, 3, 4]. Also, as the basis of MOT, the performance of object detection is the crucial key. A comprehensive understanding of MOT relies heavily on the development of object detection. Therefore, it is essential to first examine the common object detection. As shown in Fig. 1, the development of object detection could be broadly delineated into two overarching periods: the era of traditional object detection (1998-2014) and the epoch of DL-based (2014-present). The latter could be further sub-categorized into anchor-based (one-stage, two-stage separately) and anchor-free methods.

The conventional object detection algorithms[13, 14, 15] primarily rely on feature extraction methods that are manually crafted[16]. The proposal of RCNN[17] marked the dawn of deep learning in object detection. After filtering related papers, the research could be divided into anchor-based and anchor-free, to be more precise, the anchor-based would be described as two-stage and one-stage methods.

Firstly, two-stage object detection methods: the two-stage methods refer to an object detection approach that involves generating candidates first and then performing classification as well as regression. In this mode, Kaiming et al. aimed for handling the overlapping parts between candidate boxes, and adopted a spatial pyramid pooling layer to process regions of interest in different scales, thereby avoiding redundant

Key Scientific and Technological Project of Henan Province, China232102240023.

[*] Ying Cao is corresponding author.

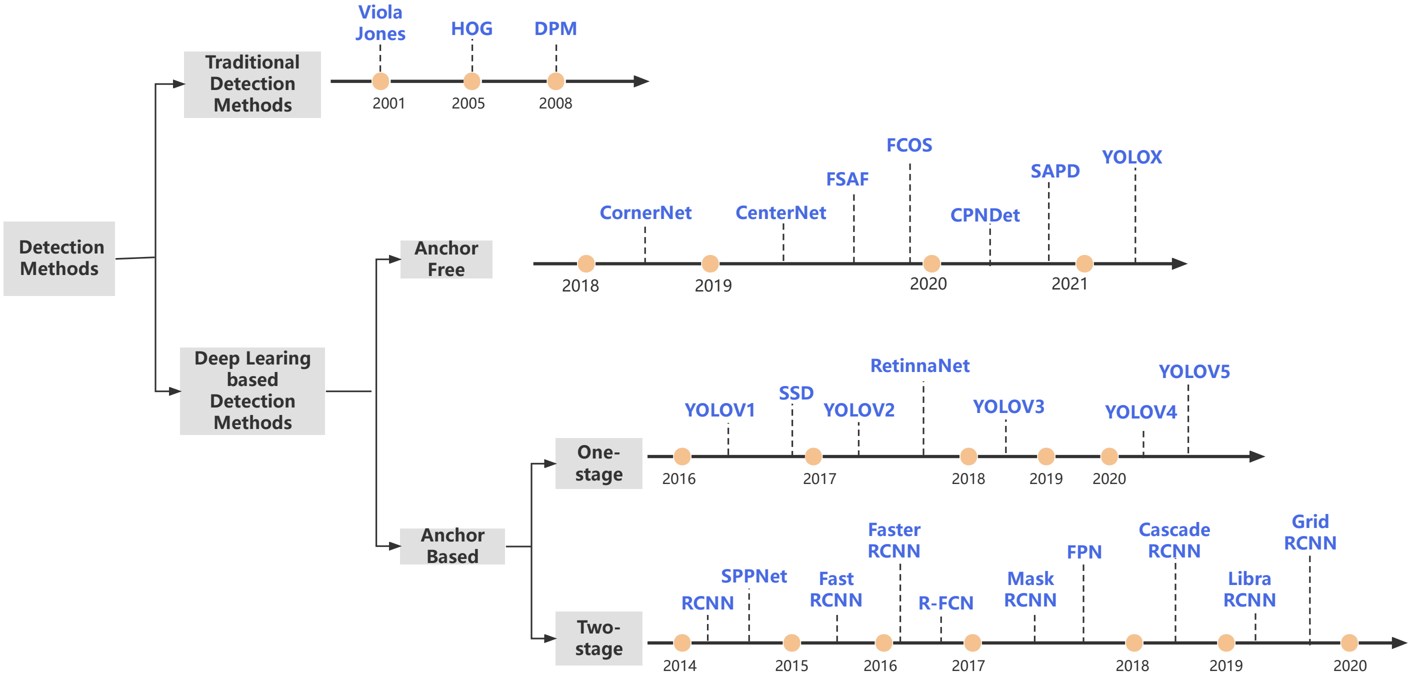


Fig. 1. The development of object detection.

computation in overlapping areas and improving computational efficiency, Meanwhile, they also replaced the fixed size maintain in RCNN with any size of the input image to better preserve the information of input image[18]. To solve the complex training process of SPPNET, Girshick et al. proposed Fast R-CNN, which significantly decreased the computational time by leveraging the Region of Interest (ROI) pooling layer[19]. Subsequently, Ren et al. proposed the Faster R-CNN algorithm, which replaced selective search with the Region Proposal Network (RPN) to further reduce training and testing time while enabling end-to-end training[20]. Later, Lin et al. designed Feature Pyramid Network (FPN) to include both high-level semantic features and low-level visual feature in image features representation for achieving multi-scale object detection[21]. Also based on Faster R-CNN, Cascade R-CNN incorporated the cascaded training to filter more accurate object boxes, thereby achieving higher detection performance[22]. Though the superior performance of these two-stage methods, the complexity, speed, and inaccuracy of candidates may be the limitations of two-stage model.

One-stage object detection methods: compared with two-stage methods, one-stage can achieve high detection speeds while maintaining a high level of accuracy by eliminating the region proposal step. for example: Redmon et al. designed YOLOv1 to predict the class probability and position information for each grid which is generated by input image partition[23]. The same year, Liu et al. utilized the newly introduced techniques of Multi-reference and Multi-resolution detection, along with the employment of multi-scale feature maps, to significantly enhance the precision of multi-scale object detection[24]; Another, YOLOv2[23] modified the network structure of YOLOv1 with a utilization of a more efficient feature extraction network DarkNet-19[25]. YOLOv3 enhanced the DarkNet-19 as DarkNet-53 and took the advantage of FPN (Feature Pyramid Network) structure to achieve multi-scale prediction[26]. YOLOv4 integrated various techniques to improve detection accuracy[27]. Except for the YOLO series, RetinaNet employed Focal Loss to replace the standard cross-entropy loss function, and through weight adjustment automatically making the model more focused on training hard samples[28].

While anchor-based methods can achieve good performance, the bounding boxes (bbox)-based design would be not appropriated. For instance, the different settings of number, size and aspect ratio of bboxes in various datasets, the imbalanced positive and negative samples caused by large number of bboxes, and the more non-object information covered by unfit bboxes shape, these would all lead to detection performance decreasing. In response to these, researches have been conducted on Anchor-free object detection methods, such as CornerNet[29] transformed object detection into corner points detection problem, and used the accurate corner points prediction of objects to eliminate the need for anchor design and regression processes in traditional object detection methods, thus achieving superior object localization accuracy and multi-scale, occlusion issues handled effectively, as well the reduction of parameter number and computational complexity. CornerNet[29] has made a significant contribution in the anchor-free filed. CenterNet also predicted the center point location of the target, but additionally regressed other target attributes through the center point position to achieve higher accuracy and greater versatility[30]. Followed the high detection efficiency characteristic of YOLO series, YOLOx adapted the YOLO detector to an anchor-free design and achieved significant performance improvements by integrating other advanced technique[31].

III. Multi-object Tracking

A. Tracking-by-Detection

Tracking-by-Detection (TBD) is a two-stage paradigm for MOT that generated candidates by a detector firstly, and then associated them inter-frame to obtain the objects' motion trajectories[32]. In TBD paradigm, SORT is the classic method, which formed the objects' trajectories by measuring the relationship among frames based on the Intersection over Union (IOU) between tracked detection and predictions. Specifically, relying on the Kalman filter [39] and the Hungarian matching[40], SORT could predict the current position based on the position of the target in the previous frame, and match the detection bboxes with the prediction bboxes. Though SORT is fast, it still limited in handling object occlusion, and this would lead to high number of ID switches[5]. For this, based on the utilization recursive Kalman filter and a frame-by-frame Hungarian matching, in data association, DeepSORT added a cascade matching scheme to enhance the tracking performance in the crowded scenes and reduce the frequency of ID switches in real-time tracking task[6].

Different with DeepSORT[6] which took advantage of Hungarian matching, Guillem et al. proposed a neural network-based solver to automatically learn the features and spatio-temporal variation patterns of targets, thereby better adapting to different tracking scenarios[33]. Another, to addressed the issues of unstable MOT performance caused by a reliance on appearance features, Han et al. proposed the Motion-aware Tracker (MAT) method to pay more attention on motion features[34], and achieve enhanced performance by integrating Integrated Motion Localization (ML), Dynamic Reconnection Context (DRC) and 3D Integral Image (3DII) modules. Among them, ML is designed to solve camera motion and non-rigid object motion, DRC is utilized to deal with long-range motion-based reconnection and interface caused by occlusion or blur, 3DII is employed in association stage to cut useless track-detection association connections with temporal-spatial constraints.

With the introduction and advancement of anchor-free detectors, MOT can also benefit from the superiors of this detection paradigm, including better adaptability to diverse sizes, shapes, and aspect ratios of targets, stronger generalization ability without prior knowledge or manual feature engineering, and other advantages. For example, based on CenterNet[30] or YOLOx[31]: compared to other methods that only associate high-scoring detection bboxes, ByteTrack has improved the accuracy of data association by incorporating low-scoring detection bboxes into the process[7]. Also integrated with YOLOx[31], Yang et al. designed a Cascaded-Buffered Intersection over Union (C-BIoU) tracker to solve the performance reduction of traditional methods with the interference of unreliable appearance features and irregular motions. Specifically, in C-BIoU's matching stage, instead of using IoU as the above methods, this tracker utilized BIoU and set a proportion to expand the size of the matching bbox, thereby increasing the chances of properly matching. Additionally, a cascaded matching approach is adapted to prevent BIoU from expanding arbitrarily and causing mismatching. Based on these to address the issues of tracking failures caused by non-overlapping detections and tracks of identical objects in adjacent frames, as well as mismatches between detection and tracks due to inaccurate motion estimation[35]. Compared to the optimization of C-BIoU in an IoU matching way, based on IoU, Aharon et al. introduced a new tracker, BoT-SORT, which attempted to adapt a simple fusion method of IoU and Re-ID's cosine distance, to establish a stronger correlation between detections and tracks in data association. along with the better bboxes positions obtained by adding camera-motion compensation features and a more accurate Kalman filter[39] state vector, achieving better performance of MOT[36]. Also based on "SORT-like" methods as BoT-SORT[36], to tackle the phenomenon of DeepSORT's[6] under-performance relative to recent existing advanced methods, StrongSORT enhances its tracking capabilities through the equipped modules such as the Appearance-Free Linking model (AFLink) and the Gaussian Smoothed Interpolation algorithm (GSI)[8]. Similarly focused on SORT[5], OC-SORT attempted to optimize the motion model and emphasized the critical role of detection information (i.e., observation) in recovering lost trajectories and reducing the accumulation of Kalman filter errors during occlusion periods, thus improving the overall robustness of MOT in occlusion or non-linear motion scenarios[37]. Furthermore, Deep OC-SORT achieved performance improvement by introducing dynamic visual appearance and camera motion compensation upon OC-SORT[37], as well as the addition of adaptive weighting factors during the trajectory association stage to balance the weights between motion and appearance models in constructing a reasonable cost distance[38].

Tracking-by-detection paradigm could take advantage of the superior of object detection to obtain more detailed information about targets and employ these to assist effectively tracking in occlusion, small objects, or blurred scenes. However, in this two-stage paradigm, the detect-first-then-track execution logic would result in the over-reliance of tracking performance on the detector's ability, and increases the computational complexity.

B. Joint Detection and Tracking

With the rapid advancement of multi-task learning in deep learning, there is increasing interest in one-shot MOT research where the core concept is to perform object detection and re-identification embedding simultaneously in a unified architecture with the aim of reducing inference time. Also according to the used backbone detector, the recent Joint-Detection-and-Tracing (JDT) methods can be presented in two-stage, one-stage, and anchor-free perspectives.

1) Joint Two-stage Detection and Tracking: D & T is one of the pioneering approaches that adopted the JDT paradigm to

TABLE I
SUMMARY OF TBD RELATED PAPERS

Algorithm	Year	Detection	Dataset	MOTA(%)
SORT[5]	2016	Faster-RCNN	MOT15, MOT16	33.4, 59.8
DeepSORT[6]	2017	Faster-RCNN	MOT16	61.4
[33]	2019	Faster-RCNN	MOT16, MOT17	58.6, 58.8
MAT[34]	2022	Faster-RCNN	MOT16, MOT17	70.5 69.5
ByteTrack[7]	2021	YOLOX	MOT17, MOT20	80.3, 77.8
StrongSORT[8]	2021	YOLOX	MOT17, MOT20	78.3, 72.2
C-BIou[35]	2022	YOLOX	MOT17, DanceTrack	81.1, 91.6
BoT-SORT[36]	2022	YOLOX	MOT17, MOT20	80.6, 77.8
OC-SORT[37]	2023	YOLOX	MOT17, MOT20	78.0, 75.5
DEEP OC-SORT[38]	2023	YOLOX	MOT17, MOT20	79.4, 75.6

optimize MOT performance through mutual feedback between detection and tracking[41]. Following the JDT paradigm, Tracktor++ additionally integrated a Siamese CNN-based re-ID network and a motion model based on Constant Velocity Model (CMC), Constant Velocity Assumption (CVA) to enhance MOT performance, where CMC was used to address occlusion and camera jitter, and CVA was designed for fast-moving target in low frame rate scenarios in object tracking[42]. Though the importance of motion information, based on a Siamese-like architecture, Tobias et al. created a quasi-dense tracking (QDTrack) to achieve tracking in scenes with significant appearance changes and dense targets only by appearance information effectively utilizing[43]. In contrast to the aforementioned papers' focus on motion or appearance features extraction, Yu et al. mainly considered the feature representation and proposed ReleationTrack. In their scenario, for a better feature presentation, ReleationTrack addressed the issue of feature dependence in the JDT paradigm by introducing Global Context Disentangling (GCD) to decouple the extracted features into task-specific features. additionally, ReleationTrack identified the problem of neglecting global semantic relevance in existing ReID features and designed the Guided Transformer Encoder (GTE) module to learn more globally aware ReID features[44]. In any case, reasonable feature acquisition or effective feature representation would lead to the more efficient development of the current MOT framework.

2) Joint One-stage Detection with Tracking: With the introduction and optimization of the one-stage detectors, they were also employed in JDT paradigm for further improvement of MOT performance. For instance: Wang et al. established the Joint-detection-with-tracking (JDE) paradigm to integrate object detection and Re-ID feature extraction in a shared model, their method reduced the inference time and provided a benchmark of JDT[45]. Following with JDE[45], CSTrack[46] constructed a Cross-correlation network (CCN) to focus on the problem of over-competition between detection and ReID, and the performance of both branches was enhanced by cross-relating the two branches with the utilization of an attention mechanism, the Scale-Aware-Attention-Network (SAAN) is also developed to effectively prevent semantic misalignment and improve the association capability of ID embeddings, finally achieving superior MOT performance. As further, CSTrackV2[47] re-stored the misclassified targets caused by the detector through the temporal clue for secondary inspection, based on which to effectively recover missed targets and obtain smoother track trajectories. Comparatively, AttTrack adopted the dual-model structure that utilized an attention mechanism to transfer knowledge from the complex network (teacher model) to the lightweight network (student model) for MOT acceleration. This approach further enhancing the tracking accuracy through an interleaved model, which leverages updated predictions from the teacher model as prior knowledge to assist the student model[48].

Also focusing on studying the complexity and invalidity of TBD paradigm, but unlike the above optimizations, RetinaTrack[49] mainly considered the feature representation. Specifically, Lu et al. added post-FPN prediction layers in utilizing the RetinaNet[28], which has demonstrated advantages in small object detection, to obtain instance-level features for each target by splitting different-size features corresponding to different anchor points in advance, by continuously matching the anchor points with instance-level features to solve the problem of incorrect target matching (ID Switch) caused by occlusion.

In contrast to the above methods, which only implemented detection and embedding within a unified framework and actually still relied on two-stage tracking of detection+embeding and data association, Chained-Tracker[50] transformed the data association into a regression task of paired detection bboxes within two frames, and by leveraging the feature enhancement of the joint attention module, Chained-Tracker

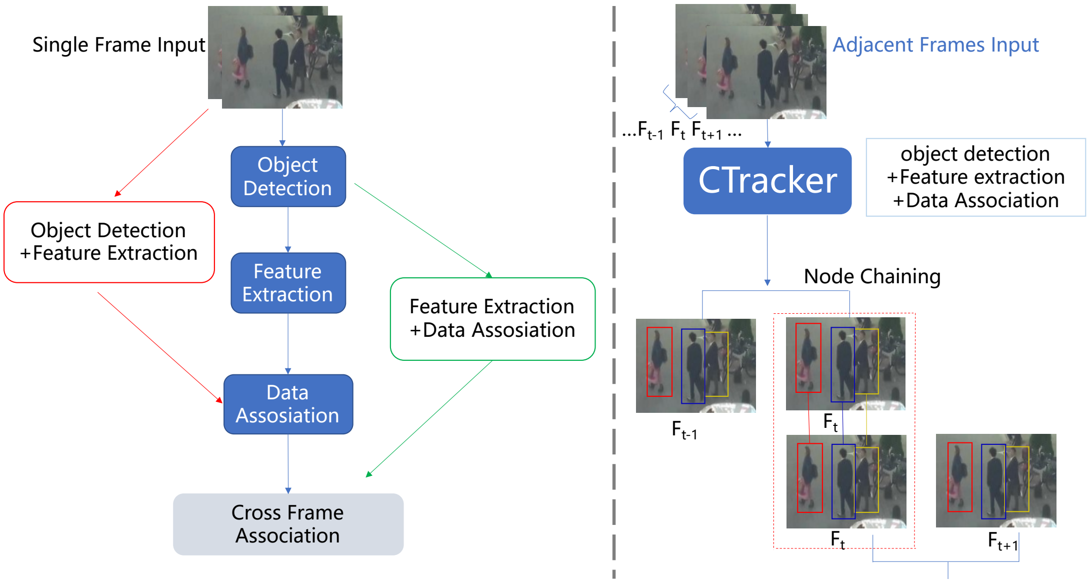


Fig. 2. Comparison between Chained-Tracker and Traditional MOT algorithm.

achieves a full end-to-end implementation of JDT, thereby further improving the performance of MOT[50], the comparison between Chained-Tracker and traditional MOT algorithm is illustrated in Fig. 2.

3) Joint Anchor-free Detection with Tracking: Though the one-shot trackers exhibit superior performance, the anchor-based architecture integrated detectors such as YOLO have been found not ideal for learning ReID features. Despite the good results, but also led to a relatively larger number of ID switches. Therefore, tracking with anchor-free detectors has been conducted to address the computational complexity and inefficiency issues for the JDT paradigm. Examples include:

Following JDE, FairMOT[51] proposed an anchor-free approach to tackle the issue of anchor confusion that plagues anchor-based detectors, which impeded the learning of appropriate Re-ID information and identification of the same target. Specifically, the authors integrated the anchor-free detector CenterNet[30] and Re-ID module into two homogeneous branches, this innovation helped to alleviate the problem of over-competition between detection and ReID, resulting in a better balance between the two and the acquisition of high-quality ReID embedding. Contrasting to the aforementioned methods that optimized the branches in JDT, CorrTracker[52] utilized self-supervised learning for training the local correlation module to model the relationships between the target and its surrounding environment, in this way, to strengthen the model's ability in discriminating similar objects, ultimately improving its recognition performance in complex scenarios. Similarly, to improve the discriminability of target features and solve the issue of missed low-confidence detection caused by occlusion or blur, SGT[53] adopted GNN in FairMOT[51], and represented the detection, connection, and relationship features as nodes, edges, edge features. With the aid of edge and edge features, high-order relationship features are generated by aggregating adjacent detection features and relationships with the current, based on which to enhance the discriminability of detection and recover missed detection. Furthermore, Ren et al. considered the unreliability of coarse-grained global object representation generated according to bbox or center features which were in the presence of occlusion and other interfering factors. FineTrack was designed to adopt a comprehensive approach that describes the object's appearance from both global and local perspectives. Concretely, FineTrack proposed the Flow Alignment FPN (FAFPN) and Multi-head Part Mask Generator (MPMG) to explore multi-dimensional fine-grained feature representation, Among which, FAFPN corrected misalignment and aggregated multi-scale features, MPMG focused on extracting fine-grained representations of the aligned feature map[54].

The JDE-followed solutions mentioned above all demonstrated superior performance. However, JDE paradigm only implements a unified architecture of detection and Re-ID embedding, which is still a two-stage approach that involves detection + embedding and data association. For this, Tokmakov et al. thought of a thoroughly one-stage network, CenterTrack, that combined detection and data association in a more meaningful way. In this model, the effective migration of the detection and data association were undertaken by the utilization of CenterNet[30]. In detail, CenterTrack used the CenterNet[30] to locate the center of the target as a simplified target representation, based on this, model took the current and previous frame image pairs as detector inputs, combined with the previous frame's trajectory heatmap, CenterTrack[55] could predict the center points, object box size, and offset of the targets in the current frame, especially for occluded targets. After that, the greedy matching would be performed relying only on the predicted offset and the distance between center points detected in the previous frame to solve all

TABLE II
SUMMARY OF JDT RELATED PAPERS

Algorithm	Year	Detection	Dataset	MOTA(%)
D&T[41]	2017	R-FCN	ImageNet VID	74.2%(mAP)
Tracktor++[42]	2019	Faster-RCNN	MOT17	61.9
QDTrack[43]	2022	Faster-RCNN	MOT17	64.6
RelationTrack[44]	2022	Faster-RCNN	MOT16, MOT17, MOT20	75.6, 73.8, 67.2
JDE[45]	2019	YOLOv3	MOT16	64.4
CSTrack[46]	2020	YOLOv5	MOT16, MOT17, MOT20	75.6, 74.9, 66.6
CSTrackV2[47]	2021	YOLOv3	MOT16, MOT17, MOT20	76.4, 76.3, 70.7
AttTrack[48]	2022	YOLOv5	MOT15, MOT17	52.9,4 3.6
RetinaTrack[49]	2020	RetinaNet	MOT17, Waymo	39.19, 44.92
ChainedTrack[50]	2020	RetinaNet	MOT16, MOT17	67.6, 66.6
FairMOT[51]	2021	CenterNet	MOT16, MOT17, MOT20	74.9,73.7, 61.8
CorrTracker[52]	2021	CenterNet	MOT16, MOT17, MOT20	76.6, 76.5, 65.2
SGT[53]	2022	CenterNet	MOT16, MOT17, MOT20	76.8, 76.5, 72.8
FineTrack[54]	2023	YOLOX	MOT17, MOT20	80.0, 77.9
CenterTrack[55]	2020	CenterNet	MOT17, KITTI	67.8, 89.44

targets effectively location in each frame and the association of the correct target over time, thus generating trajectories with minimal local input.

Compared to TBD, JDT paradigm could reduce the cases of missed and false detections, improving the accuracy and stability of MOT, and importantly, more suitable for real-time implementation. However, drawbacks still exist, such as the over-competition between detection and tracking[56, 57], the need for sufficient training with dataset preparation, and the performance bottleneck faced with occlusion or motion blur. All of these need further exploration.

IV. MOT BENCHMARKS AND APPLICATION

A. Benchmarks

In recent years, numerous MOT public datasets have been released by universities, companies, and research teams, their published annotated videos along with unified hypotheses, annotations, and evaluation tools ensure the feasibility of validating MOT schemes and promote the development of MOT. Specifically, the development of MOT datasets is shown in Fig. 3.

B. Application

With the development of MOT, it can be integrated in many fields, including but not limited to:

(1) Military Filed

As the deployment of unmanned aerial systems (UAS)[58] continues to rise, the demand for effective MOT in aerial surveillance is increasing. Leveraging MOT can enable precise positioning and tacking of enemies and military weapons, thereby significantly enhancing combat efficiency.

(2) Video Surveillance

As one of the primary supporting technologies in video surveillance[59], MOT can track and analyze the trajectory of pedestrians, vehicles, and other objects in specific areas, providing various intelligent services such as monitoring abnormal behaviors of pedestrians or vehicles[60], and warning of safety hazards. This technology has played a significant role in enhancing the effectiveness of video surveillance.

(3) Autonomous Driving

MOT is a critical component in the development of autonomous driving technology, it enables autonomous vehicles to effectively predict the subsequent movement of pedestrians, vehicles, or others based on their current trajectories and assist plan their driving paths accordingly. This allows for accurate avoidance of obstacles in the surroundings, preventing collisions and achieving safe, advanced autonomous driving[61].

(4) Medical Diagnosis

MOT can also be used in medical image analysis to perform the tasks such as cell tracking[62], blood vessel tracking[63], neuron tracking[64], and more. This technology can assist doctors in tracking diseases such as tumors and vascular lesions, ultimately enhancing the accuracy of medical diagnoses.

(5) Motion Analysis

MOT is still employed in motion analysis, particularly in sports[65], by tracking and capturing information on athletes' positions and movements, this technology can assist in analyzing and tracking athletes; actions and performance, aiding in performance evaluation and the development of effective

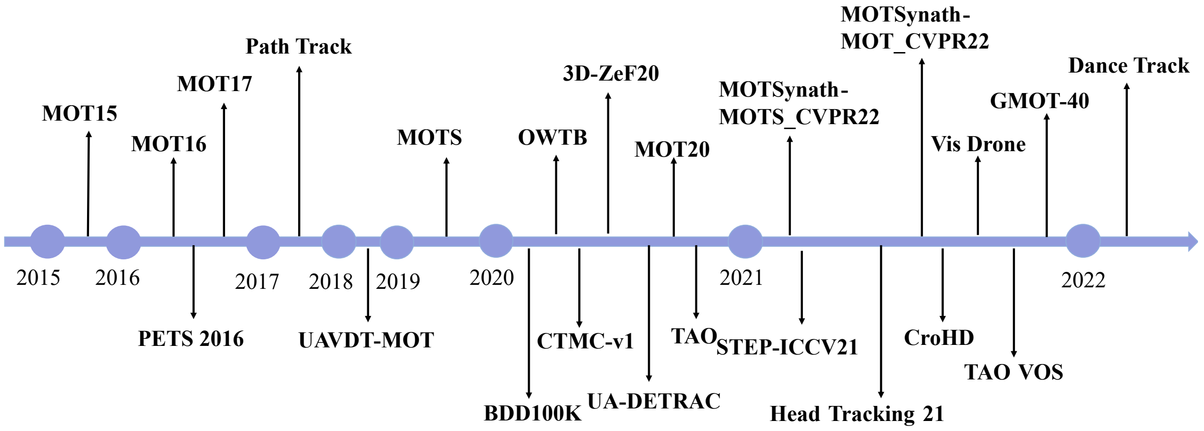


Fig. 3. Development of MOT Dataset.

training plans. Ultimately, leading to improved training outcomes.

V. Conclusions

According to the analysis of the lastest research, this paper particularly focuses on discussing the continuous optimization strategies of MOT around the development of object detection at each stage. Explored the relevant algorithms and characteristics of TBD and JDT methods separately. In addition, we also introduced datasets and application areas related to multi-target tracking. Finally, we hope to provide useful references for researchers in this field.

References

[1] Zheng Tang et al. “Single-Camera and Inter-Camera Vehicle Tracking and 3D Speed Estimation Based on Fusion of Visual and Semantic Features”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2018, pp. 108–1087. DOI: 10.1109/CVPRW.2018.00022.

[2] Jingxuan Hao et al. “A Review of Target Tracking Algorithm Based on UAV”. In: *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*. 2018, pp. 328–333. DOI: 10.1109/CBS.2018.8612263.

[3] Hou-Ning Hu et al. “Joint Monocular 3D Vehicle Detection and Tracking”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 5389–5398. DOI: 10.1109/ICCV.2019.00549.

[4] Cyril Robin and Simon Lacroix. “Multi-robot target detection and tracking: taxonomy and survey”. In: *Autonomous Robots* (2015), pp. 7 29–760. DOI: 10.1007/s10514-015-9491-7.

[5] Alex Bewley et al. “Simple online and realtime tracking”. In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, pp. 3464–3468. DOI: 10.1109/ICIP.2016.7533003.

[6] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. “Simple online and realtime tracking with a deep association metric”. In: *2017 IEEE International Conference on Image Processing (ICIP)*. 2017, pp. 3645–3649. DOI: 10.1109/ICIP.2017.8296962.

[7] Yifu Zhang et al. *ByteTrack: Multi-Object Tracking by Associating Every Detection Box*. 2022. arXiv: 2110.06864 [cs.CV].

[8] Yunhao Du et al. “StrongSORT: Make DeepSORT Great Again”. In: *IEEE Transactions on Multimedia* (2023), pp. 1–14. DOI: 10.1109/TMM.2023.3240881.

[9] Sankar K. Pal et al. “Deep learning in multi-object detection and tracking: state of the art”. In: *Applied Intelligence* 51.9 (2021), pp. 6400–6429. ISSN: 1573-7497. DOI: 10.1007/s10489-021-02293-7.

[10] Gioele Ciaparrone et al. “Deep learning in video multi-object tracking: A survey”. In: *Neurocomputing* 381 (2020), pp. 61–88. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2019.11.023. URL: %5Curl%7Bhttps://www.sciencedirect.com/science/article/pii/S0925231219315966%7D.

[11] Yan Dai et al. “A survey of detection-based video multi-object tracking”. In: *Displays* 75 (2022), p. 102317. ISSN: 0141-9382. DOI: 10.1016/j.displa.2022.102317. URL: %5Curl%7Bhttps://www.sciencedirect.com/science/article/pii/S0141938222001354%7D.

[12] Xiaoding Yuan et al. “Robust Instance Segmentation through Reasoning about Multi-Object Occlusion”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 11136–11145. DOI: 10.1109/CVPR46437.2021.01099.

[13] P. Viola and M. Jones. “Rapid object detection using a boosted cascade of simple features”. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.CVPR*. Vol. 1, pp. I–I. ISBN: 1063-6919. DOI: 10.1109/CVPR.2001.990517.

[14] N. Dalal and B. Triggs. “Histograms of oriented gradients for human detection”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Vol. 1. 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.

[15] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. “A discriminatively trained, multiscale, deformable part model”. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587597.

[16] David G. Lowe. “Distinctive Image Features from ScaleInvariant Keypoints”. In: *International Journal of Computer Vision* (2004), pp. 91–110. DOI: 10.1023/b:visi.0000029664.99615.94.

[17] Ross Girshick et al. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 580–587. DOI: 10.1109/CVPR.2014.81.

[18] Kaiming He et al. “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.9 (2015), pp. 1904–1916. DOI: 10.1109/TPAMI.2015.2389824.

[19] Ross Girshick. “Fast R-CNN”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1440–1448. DOI: 10.1109/ICCV.2015.169.

[20] Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1137–1149. DOI: 10.1109/TPAMI.2016.2577031.

[21] Chenchen Zhu, Yihui He, and Marios Savvides. “Feature Selective Anchor-Free Module for Single-Shot Object Detection”. In: *2019 IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVP R)*. 2019, pp. 840–849. DOI: 10.1109/CVPR.2019.00093.

[22] Zhaowei Cai and Nuno Vasconcelos. “Cascade R-CNN: Delving Into High Quality Object Detection”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6154–6162. DOI: 10.1109/CVPR.2018.00644.

[23] Joseph Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.

[24] Wei Liu et al. “SSD: Single Shot MultiBox Detector”. In: *Computer Vision (ECCV) 2016*. Springer International Publishing, 2016, pp. 21–37. DOI: 10.1007/978-3-319-46448-0_2.

[25] Joseph Redmon and Ali Farhadi. “YOLO9000: Better, Faster, Stronger”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6517–6525. DOI: 10.1109/CVPR.2017.690.

[26] Joseph Redmon and Ali Farhadi. *YOLOv3: An Incremental Improvement*. 2018. arXiv: 1804.02767 [cs.CV].

[27] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. *YOLOv4: Optimal Speed and Accuracy of Object Detection*. 2020. arXiv: 2004.10934 [cs.CV].

[28] Tsung-Yi Lin et al. “Focal Loss for Dense Object Detection”. In: *201 7 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2999–3007. DOI: 10.1109/ICCV.2017.324.

[29] Hei Law and Jia Deng. “CornerNet: Detecting Objects as Paired Keypoints”. In: *International Journal of Computer Vision* 128.3 (2020), pp. 642–656. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01204-1.

[30] Kaiwen Duan et al. “CenterNet: Keypoint Triplets for Object Detection”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 6568–6577. DOI: 10.1109/ICCV.2019.00667.

[31] Zheng Ge et al. *YOLOX: Exceeding YOLO Series in 2021*. 2021. arXiv: 2107.08430 [cs.CV].

[32] Zhihong Sun et al. “A Survey of Multiple Pedestrian Tracking Based on Tracking-by-Detection Framework”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.5 (2021), pp. 1819–1833. DOI: 10.1109/TCSVT.2020.3009717.

[33] Guillem Brasó and Laura Leal-Taixé. “Learning a Neural Solver for Multiple Object Tracking”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 6246–6256. DOI: 10.1109/CVPR42600.2020.00628.

[34] Shoudong Han et al. “MAT: Motion-aware multi-object tracking”. In: *Neurocomputing* 476 (2022), pp. 75–86. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2021.12.104. URL: %5Curl%7Bhttps://www.sciencedirect.com/science/article/pii/S0925231221019627%7D.

[35] Fan Yang et al. “Hard to Track Objects with Irregular Motions and Similar Appearances? Make It Easier by Buffering the Matching Space”. In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 4788–4797. DOI: 10.1109/WACV56688.2023.00478.

[36] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. *BoT-SORT: Robust Associations Multi-Pedestrian Tracking*. 2022. arXiv: 2206.14651 [cs.CV].

[37] Jinkun Cao et al. *Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking*. 2023. arXiv: 2203.14360 [cs.CV].

[38] Gerard Maggiolino et al. *Deep OC-SORT: Multi-Pedestrian Tracking by Adaptive Re-Identification*. 2023. arXiv: 2302.11813 [cs.CV].

[39] Du Yong Kim and Moongu Jeon. “Data fusion of radar and image measurements for multi-object tracking via Kalman filtering”. In: *Information Sciences* 278 (2014), pp. 641–652. ISSN: 0020-0255. DOI: 10.1016/j.ins.2014.03.080. URL: %5Curl%7Bhttps://www.sciencedirect.com/science/article/pii/S0020025514003715%7D.

[40] H. W. Kuhn. “The Hungarian method for the assignment problem”. In: *Naval Research Logistics Quarterly* 2.1-2 (1955), pp. 83–97. DOI: 10.1002/nav.3800020109. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800020109. URL: %5Curl%7Bhttps://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109%7D.

[41] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. *Detect to Track and Track to Detect*. 2018. arXiv: 1710.03958 [cs.CV].

[42] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. “Tracking Without Bells and Whistles”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 941–951. DOI: 10.1109/ICCV.2019.00103.

[43] Tobias Fischer et al. *QDTrack: Quasi-Dense Similarity Learning for Appearance-Only Multiple Object Tracking*. 2022. arXiv: 2210.06984 [cs.CV].

[44] En Yu et al. “RelationTrack: Relation-aware Multiple Object Tracking with Decoupled Representation”. In: *IEEE Transactions on Multimedia* (2022), pp. 1–1. DOI: 10.1109/TMM.2022.3150169.

[45] Zhongdao Wang et al. *Towards Real-Time Multi-Object Tracking*. 2020. arXiv: 1909.12605 [cs.CV].

[46] Chao Liang et al. “Rethinking the Competition Between Detection and ReID in Multiobject Tracking”. In: *IEEE Transactions on Image Processing* 31 (2022), pp. 3182–3196. DOI: 10.1109/TIP.2022.3165376.

[47] Chao Liang et al. *One More Check: Making ”Fake Background” Be Tracked Again*. 2021. arXiv: 2104.09441 [cs.CV].

[48] Keivan Nalaie and Rong Zheng. “AttTrack: Online Deep Attention Transfer for Multi-object Tracking”. In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 1654–1663. DOI: 10.1109/WACV56688.2023.00170.

[49] Zhichao Lu et al. “RetinaTrack: Online Single Stage Joint Detection and Tracking”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 14656–14666. DOI: 10.1109/CVPR42600.2020.01468.

[50] Jinlong Peng et al. *Chained-Tracker: Chaining Paired Attentive Regression Results for End-to-End Joint Multiple-Object Detection and Tracking*. 2020. arXiv: 2007.14557 [cs.CV].

[51] Zhang et al. “FairMOT: On the Fairness of Detection and Re-identifica tion in Multiple Object Tracking”. In: *International Journal of Computer Vision* (2021), pp. 3069–3087. DOI: 10.1007/s11263-021-01513-4.

[52] Ying Wang et al. “Learning Correlation for Online Multiple Object Tracking”. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 4713–4717. DOI: 10.1109/ICASSP43922.2022.9746986.

[53] Jeongseok Hyun et al. “Detection Recovery in Online Multi-Object Tracking with Sparse Graph Tracker”. In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 4839–4848. DOI: 10.1109/WACV56688.2023.00483.

[54] Hao Ren et al. *Focus On Details: Online Multi-object Tracking with Diverse Fine-grained Representation*. 2023. arXiv: 2302.14589 [cs.CV].

[55] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. *Tracking Objects as Points*. 2020. arXiv: 2004.01177 [cs.CV].

[56] Jieming Yang et al. “Online multi-object tracking using multi-function integration and tracking simulation training”. In: *Applied Intelligence* (2021), pp. 1268–1288. DOI: 10.1007/s10489-021-02457-5.

[57] Haidong Wang et al. “JDAN: Joint Detection and Association Network for Real-Time Online Multi-Object Tracking”. In: *ACM Trans. Multimedia Comput. Commun. Appl.* 19.1s (Feb. 2023). ISSN: 1551-6857. DOI: 10.1145/3533253.

[58] Giancarmine Fasano et al. “Sense and avoid for unmanned aircraft systems”. In: *IEEE Aerospace and Electronic Systems Magazine* 31.11 (2016), pp. 82–110. DOI: 10.1109/MAES.2016.160116.

[59] Yuhao Luo et al. “Pedestrian tracking in surveillance video based on modified CNN”. In: *Multimedia Tools and Applications* (2018), pp. 24041–24058. DOI: 10.1007/s11042-018-5728-8.

[60] Kuei-Hsiang Chao and Pi-Yun Chen. “An Intelligent Traffic Flow Control System Based on Radio Frequency Identification and Wireless Sensor Networks”. In: *International Journal of Distributed Sensor Networks* 10.5 (2014), p. 694545. ISSN: 1550-1329. DOI: 10.1155/2014/694545.

[61] Ming Gao et al. “Multiple object tracking using a dual-attention network for autonomous driving”. In: *IET Intelligent Transport Systems* (2020), pp. 842–848. DOI: 10.1049/iet-its.2019.0536.

[62] Yoones Imani et al. “A new method for multiple sperm cells tracking”. In: *Journal of medical signals and sensors* 4.1 (2014), pp. 35–42.

[63] Dengqiang Jia and Xiahai Zhuang. “Learning-based algorithms for vessel tracking: A review”. In: *Computerized Medical Imaging and Graphics* 89 (2021), p. 101840. ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2020.101840. URL: %5Curl%7Bhttps://www.sciencedirect.com/science/article/pii/S089561112030135X%7D.

[64] Linus Manubens-Gil et al. “BigNeuron: a resource to benchmark and predict performance of algorithms for automated tracing of neurons in light microscopy datasets”. In: *Nature Methods* (2023). ISSN: 1548-7105. DOI: 10.1038/s41592-023-01848-5.

[65] Sungwon Moon et al. “A comparative study on multi-object tracking methods for sports events”. In: *2017 19th International Conference on Advanced Communication Technology (ICACT)*. 2017, pp. 883–885. DOI: 10.23919/ICACT.2017.7890221.

An Enhanced Topic Modeling Method in Educational Domain by Integrating LDA with Semantic

Ruofei Ding , Pucheng Huang, Shumin Chen, Jiale Zhang, Jingxiu Huang, Yunxiang Zheng✉

School of Educational Information Technology, South China Normal University, China

dr.zheng.scnu@hotmail.com

Abstract— **With the development of online courses, students' discussion texts in online forums and communication groups are increasing. Teachers can use these texts to monitor student learning so that they can adapt the pace of instruction accordingly. And textual topics, as the important information of the text, can be extracted from the text by topic modeling. Currently, a Latent Dirichlet Allocation (LDA) method has been used to identify the critical main topics discussed by students. However, LDA is based on word frequency and ignores semantic information. In this study, we propose a model for fusing semantic information into LDA. To verify the validity of our model, we collected two MOOC datasets for testing and conducted an ablation study using Silhouette Coefficient value and Calinski-Harabasz score as the criterion. The results show that our method is scientifically feasible and better than LDA in the field of educational topic modeling. Thus, our method is able to perform topic modeling more accurately compared to LDA. It can be used by teachers to automatically analyze large amounts of student discussion data to guide personalized learning paths.**

Keywords— **Topic Modeling, Online Discussion, Text Mining, Machine Learning, LDA**

I. Introduction

In today’s world, a large number of texts containing student discussions can be found on online education platforms. These discussion texts serve as reflections of students' discussions on various topics, offering valuable insights for teachers to interpret learning outcomes. But analyzing textual data of student discussions without automation has become an almost impossible task. With the help of topic modeling, it is possible to automatically cluster different discussion texts together.

Latent Dirichlet Allocation (LDA), one of the existing topic models, can determine the main topic of students’ discourse. LDA was created based on Bayesian probability and is widely used in the field of education because of its reasonable inference processes and hypothetical designs, combined with its requirement for only a few parameters to configure.

LDA is always used to summarize the keywords and the trends of topics in recent research in education, and is also used to improve the quality of personalized resource recommendations.

Odden used LDA to perform topic analysis on research published in the Physics Education Research Conference Proceedings from 2001 to 2018[1]. They aimed to analyze the trends of research topics and identify the topics which have received consistent attention. Through this research, they supposed that LDA was expected to help in educational research literature, referring that the analysis was “quantitative, independent, and replicable”. Gurcan Fatih et al. used N-gram modeling together with LDA to study e-learning articles related to the COVID-19, and found the trends of research and popular research issues during that time[2].

Due to the capability of LDA to assist in constructing interest models for learners and computing user preferences, some researchers utilized LDA to optimize personalized resource recommendations in online education. Lin Qi et al. achieved such optimization using LDA[3]. Peng Jiang et al. combined LDA with Artificial Neural Networks for intelligent user recommendation of online video courses[4]. Wei Kuang et al. also utilized LDA to construct user interest models and proposed a resource recommendation method for e-learning systems[5].

The widespread application of LDA models in the online education domain can be attributed to its capacity to enhance the quality of feedback loops and its advantages over some other methods. For example, Chai et al.[6] introduced a method that uses LDA to detect topics in online course feedback. The method can present the topics of feedback to the teachers in the form of word clouds and analyze the relationship between the feedback and various factors such as students’ grade, satisfaction and learning outcomes. Deepak and Shobha[7] used LDA to address the issue of identifying students who fail to complete assigned tasks within the given time in an online learning system. They employed LDA to cluster texts and learners, and the results showed that it achieved significant performance compared to other existing algorithms.

However, this does not imply that LDA is the optimal solution. There are still certain limitations in LDA. Li et al. pointed out that the LDA model fails to use semantic information to enhance feature representation, which may impact the results of semantic analysis. Grootendorst identified a limitation of the LDA model. It ignores the semantic between words due to its use of a bag-of-words representation, which leads to the result that the texts may not be represented accurately[8]. Tajbakhsh, Mir Saman also indicated that LDA disregards the semantic relationships between words in short text clustering[9].

To solve the problem of LDA lacking semantic information, it may be necessary to combine other methods with LDA to further represent the semantics, thus improving the performance of the model. This viewpoint was also shown in the article by Ekinci Ekin et al. They argued that traditional topic models have a significant limitation in which they cannot capture topics related to semantics. Furthermore, they emphasized the crucial role of semantic inference in topic modeling[10]. There are current studies indicating that semantic information can indeed affect the effectiveness of topic modeling. Grootendorst[8] found that it has a better performance in coherence of the result than LDA when using BERTopic for dynamic topic modeling. The topic coherence score is also significantly higher when using Word2vec combined with LSA instead of PLSA[11]. In their analysis of online discourse related to the Hong Kong extradition bill incident, Xu[12] found that there is a better topic relevance when combining LDA with BERT, with a 35.7% enhancement compared to using LDA only.

In text clustering analysis, Li[13] used Word2Vec in combination with LDA for topic modeling and clustering analysis of academic article abstracts. The results showed that their approach achieved approximately a 9% higher accuracy compared to using LDA only. Similarly, George and Sumathy[14] used BERT in combination with LDA for topic modeling and clustering analysis of the open dataset CORD-19, finding that their approach performed at least 10% better than using LDA only.

It can be seen that incorporating semantic information into topic modeling can significantly enhance its performance, enabling researchers to conduct more in-depth analysis of the results of topic modeling. In the current education domain, there is still limited research on combining semantic information with LDA for topic modeling and text clustering analysis. Our research aims to propose a semantic-fusioned LDA topic modeling algorithm for topic modeling and clustering of educational texts.

II. Methodology

A. Data Preparation

To prove the stability and reliability of our method, we collected two datasets (DATASET 1 and DATASET 2) for the years 2018 and 2022 from the course "Instructional Design Principles and Methods" on the China University MOOC Platform. It was a 15-week introductory Educational technology course. It provided learners with course materials, lecture videos, reading materials, and test questions, as well as forums to support peer interactions. Both datasets were obtained from the interactive forum where students and teachers engaged in discussions. Each row of the two datasets contains the question, the student's account and the student's answer. DATASET 1 consists of eight topics with 1397 rows of raw text. DATASET 2 consists of five topics with 758 rows of raw text. We pre-processed DATASET 1 and DATASET 2 by removing duplicates and blacks, and ended up with 1343 left in DATASET 1 and 703 left in DATASET 2.

B. Text representation with semantic

Text should first be transformed into a suitable representation before it can be used as data[15]. The representation determines the effectiveness in natural language processing(NLP) tasks. In the early days, researchers commonly used one-hot coding and TF-IDF coding, but both of them could only represent limited information. With the development of deep learning, the representation of text has shifted from discrete words to continuous vectors.

Continuous n-dimensional vectors can capture semantics. Word2vec[16] and Glove[17] are pre-trained word embedding models that are frequently used to convey semantics through a continuous vector. However, they are static in capturing lexical dimensions and neglect variations of semantic, so they can't represent long-term dependencies between words. In order to better represent semantics, ELMo[18] and BERT[19] have emerged. They generate dynamic word vectors for all words based on context. But ELMo employs a recurrent neural network(RNN), thus ELMo has shortcomings in learning long-term dependencies. In contrast, BERT is based on multi-head attention. So BERT is good at resolving long-term dependencies of text and can therefore represent semantic information of longer texts.

BERT was trained on the BooksCorpus dataset (800 million words) and text passages from the English Wikipedia. BERT can be used on unannotated data directly from a pre-trained model, or it can be fine-tuned for task-specific data. The most common variants of BERT are Roberta[20], DistilBERT[21], XLNet[22], ALBERT[23] and ERNIE 2.0[24]. Among them, ERNIE adapted the MASK disambiguation technique and was trained on a Chinese corpus. Therefore, ERNIE has significant improvements in Chinese NLP tasks. In this paper, we use the pre-trained ERNIE to generate embedding features for each text.

C. Proposed methodology

In this paper, we present a method (Figure 1) to LDA topic modeling that incorporates semantics in the educational domain. First of all, we used LDA for topic modeling in DATASET 1 and DATASET 2, then we obtained probability vectors (PVs) of the text belonging to each topic. Next, considering ERNIE's strengths in Chinese text， we use it to obtain sentence embedding (SEs) containing semantic information. Then, we combined PVs and SEs to obtain "Topic - Semantics" vectors (TSVs), a type of non-linear data. Later, we used ISOMAP[25] to perform dimensionality reduction on the TSVs and obtain DTSVs. Finally, we used the K-means algorithm to cluster the DTSVs. While K-means chooses centroids randomly before clustering, once the centroids are poorly chosen, it may lead to unsatisfactory clustering results. Thus, we used the Particle Swarm Optimisation (PSO) algorithm to optimize K-means.

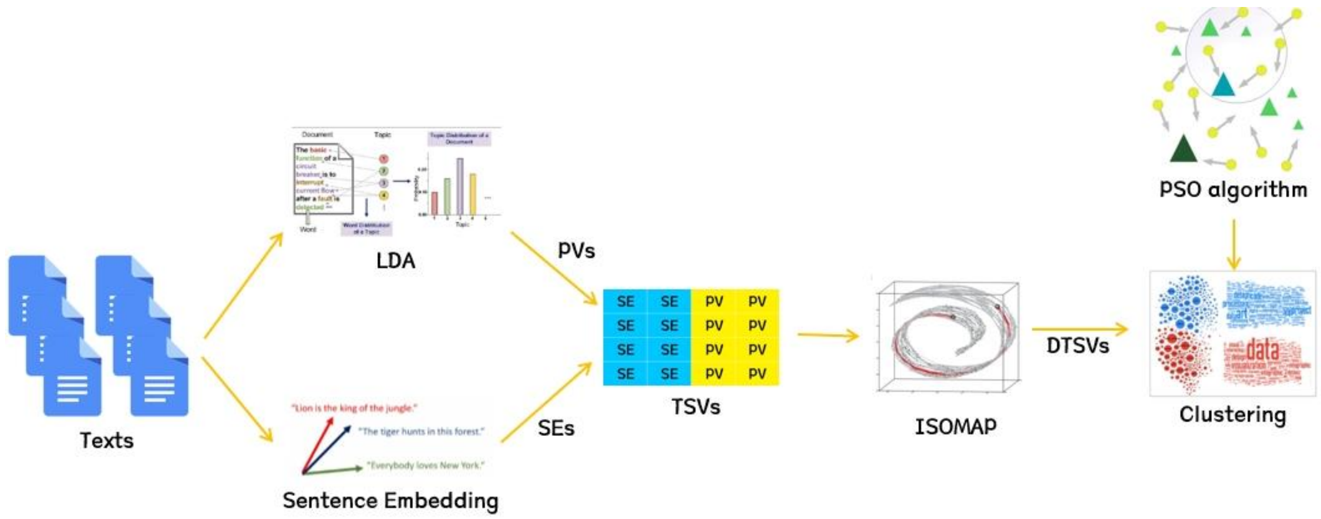

Document
Topic
Topic Distribution of a Document
Word
Word Distribution of a Topic
LDA
PVs
Texts
"Lion is the king of the jungle."
"The tiger hunts in this forest."
"Everybody loves New York."
Sentence Embedding
SEs
SE SE PV PV
SE SE PV PV
SE SE PV PV
SE SE PV PV
TSVs
ISOMAP
DTSVs
PSO algorithm
Clustering

Figure 1. A topic modeling method for LDA incorporating semantics in the educational domain

PSO was invented by James Kennedy and Russell Eberhart inspired by the regularity of the foraging behavior of birds[26]. The algorithm works by initializing a flock of birds randomly over the searching space, where each bird is referred to as a ‘‘particle’’.

Consider that a set of ‘‘particles’’ fly with a certain velocity algorithm and move to find the global best position in an iterative process. At each iteration of the algorithm, the velocity vector for each particle is modified based on three parameters: the particle momentum（The current speed of the particle）, the best position reached by the particle and that of all particles up to the current stage.

The positions and velocities of the particles are calculated using equation (1) and equation (2).

$$x_i = x_i + v_i \qquad (1)$$

$$v_i = w \times v_i + c_1 \times rand() \times (pbest_i - x_i) + c_2 \times rand() \times (gbest_i - x_i) \qquad (2)$$

Where i is the hyperparameter representing the total number of particles. x_i is the current position of the particle. v_i is the directed velocity of the particle, representing the memory term（momentum). w is the learning rate, indicating the efficiency of the particle swarm learning after each iteration. rand() is a random number between (0,1). $pbest_i$ represents the current local searched optimum position searched by the particle. $gbest_i$ represents the current searched optimum position of the swarm.$c_1 \times rand() \times (pbest_i - x_i)$ and $c_2 \times rand() \times (gbest_i - x_i)$ represent the particle pi 's cognitive and the global cognitive of all particles respectively.

PSO continuously adjusts the distance between the initial centroid and the global optimal centroid by continuous iteration. Using the outcome of the PSO as the initial centroids for K-means can effectively improve the result of K-means as these centroids are close to the global optimal centroids. Equation (3) is used to evaluate the clustering effect in each iteration.

$$F(x) = \sum_{i=1}^{k} (C_{labels=i} - centroids_i)^2 \qquad (3)$$

Where $C_{labels=i}$ represents the set of vectors that are labeled with i after the KMEANS clustering in the current state. The value of F(x) represents the total sum of squared distances between each label and the vectors belonging to that label. A smaller value of F(x) indicates a better clustering result for K-means, and indicates that the current position of the particle is more optimal.

D. Evaluation Metrics

To test our model, the Calinski-Harabasz (CH) Score and the Silhouette Coefficient (SC) were used as criteria to evaluate the result.

The CH score is the ratio of inter-cluster distance to intra-cluster distance and is defined as follows:

$$CH = \frac{(\sum_{k=1}^{K} n_k ||\mu_{c_k} - \mu||_2^2)(N-K)}{(\sum_{k=1}^{K}\sum_{i=1}^{n_k} ||x_i - \mu_{c_k}||_2^2)(K-1)} \qquad (4)$$

Where is the number of members in cluster, is the capacity of the dataset, indicates the number of clusters and represents the centroid of the dataset. The range of score is (0, +∞). The higher the value of the CH index is, the better the clustering validity is, that is, clusters are primely separated from each other and are distinctly preferable.

The SC evaluates the effect of clustering through cohesion and separation.The SC defined as:

$$SC = \frac{\sum_{i=1}^{N} \frac{b-a}{max(a,b)}}{N} \qquad (5)$$

Where a is the average distance from this sample to other samples in the same cluster, b is the average distance from this sample to all samples in the nearest neighboring cluster, N is the number of clustered. The range of SC is [-1, 1]. If the SC is close to -1, it indicates poor clustering and there are many samples that should be grouped in the neighboring cluster. If the SC is close to 0, it indicates that there are large areas of overlap between clusters. If the SC is close to 1, it indicates good clustering.

III. Result

To compare our method with LDA in terms of performance improvement, the number of a priori topics was adjusted to the number of topics in the original dataset. Specifically, in DATASET 1, we set the number of a priori topics for LDA to 8. In DATASET 2, the number of a priori topics is set to 5. Such a setup can better evaluate the performance of the LDA model and our proposed method on different datasets while ensuring fairness. Figure 2 and Figure 3 are word cloud results of our method for some topics examples in DATASET 1 and DATASET 2.In the three topic samples of DATASET 1, the main topics student concern about were "Instruction, Design", "Student, Instruction", and "Design, Curriculum". In the three topic samples of DATASET 2, they were "Analysis, Evaluation", "Instruction, Design", and "Training, Analysis". It can be seen from the results above that students focus on different topics when facing different topic samples, which leads to the discrepancy on the topics among these samples.

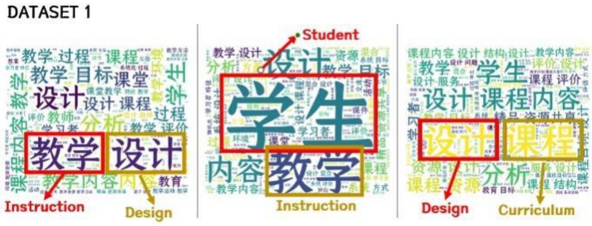


Figure 2. Some examples of topic word clouds from DataSet 1

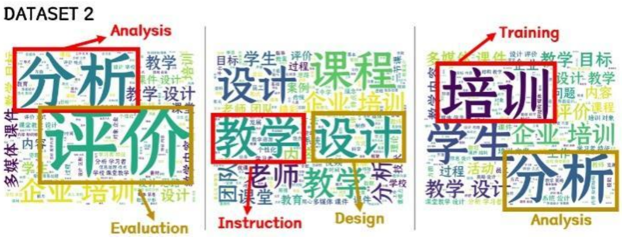


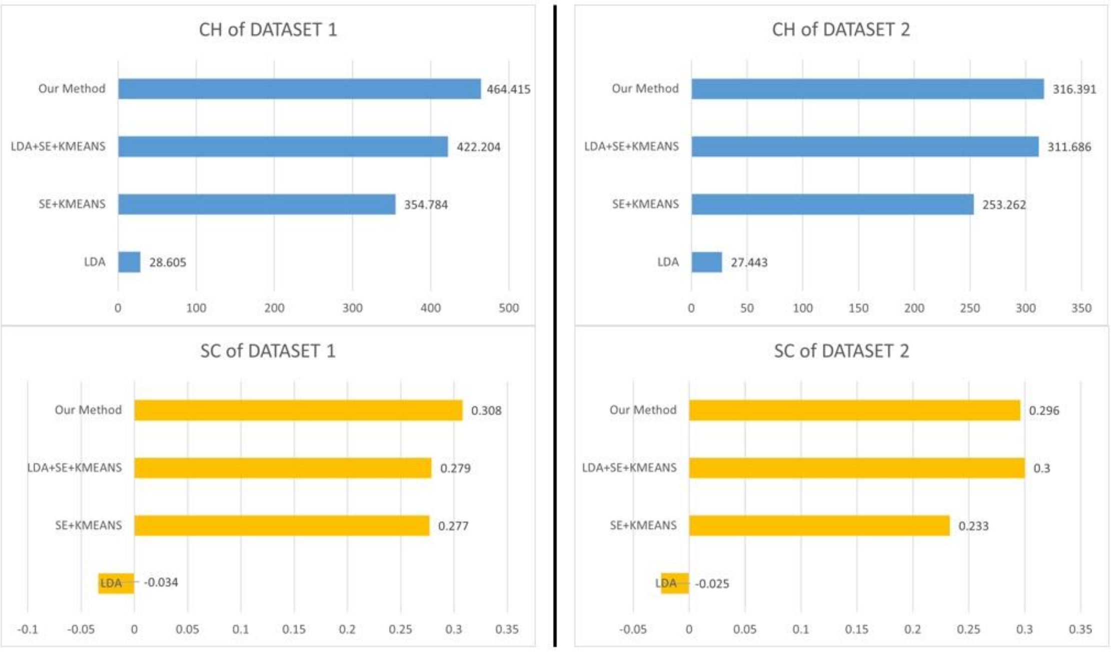
Figure 3. Some examples of topic word clouds from DataSet 2

To further analyze the scientific validity of our proposed method and the performance improvement of our method, we also conducted an ablation study using CH score and SC value. Figure 4 shows the results on two datasets, comparing the CH scores and SC values obtained by clustering using LDA, Sentence Embedding (SE)+KMEANS, LDA+SE+KMEANS, and our method, respectively.

We can see that our method shows superior performance in all models. Specifically, on DATASET 1, our model achieves the highest performance on both metrics. While on DATASET 2, compared to the next best performing LDA + SE + KMEANS model, our model has a higher CH score but a slightly lower SC value. We hypothesized that this may be due to the fact that the topics in DATASET 2 are very different from each other and the topics are not tightly structured internally. To verify our hypothesis, we visualized our data.

Figure 5 and Figure 6 show the clustering results of our method and other methods on DATASET 1 and DATASET 2. Clearly, the data and clustering results of DATASET 1 are closer than those of DATASET 2, which confirms our hypothesis.

Figure 4. CH score and SC value of two datasets with different methods



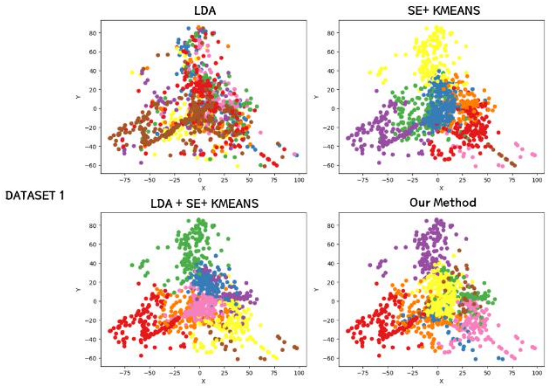


Figure 5. Visualization of clustering for DATASET 1

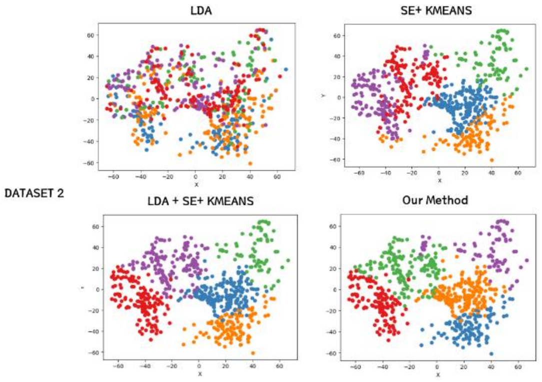


Figure 6. Visualization of clustering for DATASET 2

When comparing results on DATASET 1 and DATASET 2 by using our method, our model displays significantly greater metric differences on DATASET 1 when compared to the second-best LDA + SE + KMEANS method. Therefore, this result implies that our method is more effective in dealing with datasets with more clustered topic distributions. The PSO algorithm played a key role in achieving this result by optimizing the initial centroid selection of K-means to better handle densely distributed data. This also confirms the power of PSO for nonlinear optimization problems.

IV. Conclusion

In this study, we proposed a model for fusing semantic information with LDA. Specifically, we collected two MOOC datasets for testing and conducted an ablation study using SC value and CH score as the criterion. The results showed that our method is scientifically feasible and better than LDA in the field of educational topic modeling.

The innovation of our method is to incorporate semantic information into the LDA topic model and apply it to education. We validated the feasibility and effectiveness of the method in terms of performance. In future research, we will investigate whether the topic model incorporating semantic information can reflect students' cognition and analyze the results at a fine-grained level.

Acknowledgment

This research was funded by the National Natural Science Foundation of China: 62207015, and the Humanities and Social Sciences Youth Foundation of the Chinese Ministry of Education: 22YJC880021.

References

[1] T. O. B. Odden, A. Marin, and M. D. Caballero, “Thematic analysis of 18 years of physics education research conference proceedings using natural language processing,” *Phys. Rev. Phys. Educ. Res.*, vol. 16, no. 1, p. 010142, Jun. 2020, doi: 10.1103/PhysRevPhysEducRes.16.010142.

[2] F. Gurcan, G. G. M. Dalveren, and M. Derawi, “Covid-19 and E-Learning: An Exploratory Analysis of Research Topics and Interests in E-Learning During the Pandemic,” *IEEE Access*, vol. 10, pp. 123349–123357, 2022, doi: 10.1109/ACCESS.2022.3224034.

[3] Q. Lin, S. He, and Y. Deng, “Method of personalized educational resource recommendation based on LDA and learner’s behavior,” *International Journal of Electrical Engineering & Education*, p. 0020720920983511, Jan. 2021, doi: 10.1177/0020720920983511.

[4] P. Jiang, Y. Feng, C. Niu, and Y. Dai, “Study of intelligent recommendation for online video courses,” in *2021 IEEE 5th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC)*, Oct. 2021, pp. 1290–1294. doi: 10.1109/ITNEC52019.2021.9587262.

[5] W. Kuang, N. Luo, and Z. Sun, “Resource recommendation based on topic model for educational system,” in *2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference*, Aug. 2011, pp. 370–374. doi: 10.1109/ITAIC.2011.6030352.

[6] S. Unankard and W. Nadee, “Topic Detection for Online Course Feedback Using LDA,” in *Emerging Technologies for Education*, E. Popescu, T. Hao, T.-C. Hsu, H. Xie, M. Temperini, and W. Chen, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 133–142. doi: 10.1007/978-3-030-38778-5_16.

[7] N. A. Deepak and N. S. Shobha, “Analysis of Learner’s Behavior Using Latent Dirichlet Allocation in Online Learning Environment,” in *Computational Methods and Data Engineering*, V. Singh, V. K. Asari, S. Kumar, and R. B. Patel, Eds., in Advances in Intelligent Systems and Computing. Singapore: Springer, 2021, pp. 231–242. doi: 10.1007/978-981-15-7907-3_18.

[8] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure.” arXiv, Mar. 11, 2022. doi: 10.48550/arXiv.2203.05794.

[9] M. S. Tajbakhsh and J. Bagherzadeh, “Semantic knowledge LDA with topic vector for recommending hashtags: Twitter use case,” *Intelligent Data Analysis*, vol. 23, no. 3, pp. 609–622, Jan. 2019, doi: 10.3233/IDA-183998.

[10] E. EKİNCİ and S. OMURCA, “NET-LDA: a novel topic modeling method based on semantic document similarity,” *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 4, pp. 2244–2260, Jan. 2020, doi: 10.3906/elk-1912-62.

[11] S. Kim, H. Park, and J. Lee, “Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis,” *Expert Systems with Applications*, vol. 152, p. 113401, Aug. 2020, doi: 10.1016/j.eswa.2020.113401.

[12] X. Tan, M. Zhuang, X. Lu, and T. Mao, “An Analysis of the Emotional Evolution of Large-Scale Internet Public Opinion Events Based on the BERT-LDA Hybrid Model,” *IEEE Access*, vol. 9, pp. 15860–15871, 2021, doi: 10.1109/ACCESS.2021.3052566.

[13] C. Li *et al.*, “LDA Meets Word2Vec: A Novel Model for Academic Abstract Clustering,” in *Companion Proceedings of the The Web Conference 2018*, in WWW ’18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, Apr. 2018, pp. 1699–1706. doi: 10.1145/3184558.3191629.

[14] L. George and P. Sumathy, “An integrated clustering and BERT framework for improved topic modeling,” *Int. j. inf. tecnol.*, vol. 15, no. 4, pp. 2187–2195, Apr. 2023, doi: 10.1007/s41870-023-01268-w.

[15] K. Babić, S. Martinčić-Ipšić, and A. Meštrović, “Survey of Neural Text Representation Models,” *Information*, vol. 11, no. 11, Art. no. 11, Nov. 2020, doi: 10.3390/info11110511.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space.” arXiv, Sep. 06, 2013. Accessed: Sep. 28, 2023. [Online]. Available: http://arxiv.org/abs/1301.3781

[17] J. Pennington, R. Socher, and C. Manning, “GloVe: Global Vectors for Word Representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[18] M. E. Peters *et al.*, “Deep contextualized word representations.” arXiv, Mar. 22, 2018. Accessed: Sep. 28, 2023. [Online]. Available: http://arxiv.org/abs/1802.05365

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv, May 24, 2019. doi: 10.48550/arXiv.1810.04805.

[20] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” arXiv, Jul. 26, 2019. doi: 10.48550/arXiv.1907.11692.

[21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.” arXiv, Feb. 29, 2020. doi: 10.48550/arXiv.1910.01108.

[22] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019. Accessed: Sep. 28, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

[23] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.” arXiv, Feb. 08, 2020. doi: 10.48550/arXiv.1909.11942.

[24] Y. Sun *et al.*, “ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, Art. no. 05, Apr. 2020, doi: 10.1609/aaai.v34i05.6428.

[25] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, “A Global Geometric Framework for Nonlinear Dimensionality Reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000, doi: 10.1126/science.290.5500.2319.

[26] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in *Proceedings of ICNN’95 - International Conference on Neural Networks*, Nov. 1995, pp. 1942–1948 vol.4. doi: 10.1109/ICNN.1995.488968.

Session 2C: Security & Blockchain 2

Chair: Prof. Michael Angelo Brogada , Bicol University, Philippines

1 Paper ID: 20240459, 137~142

Intelligent Anomaly Detection System Based on Ensemble and Deep Learning

Dr. Babu Baniya, Mr. Thomas Rush,

Bradley University. USA

2 Paper ID: 20240064, 143~146

A Private Blockchain System based on Zero Trust Architecture

Prof. Yao-Chung Chang, Prof. Yu-Shan Lin, Dr. Hsin-Te Wu, Prof. Arun Kumar Sangaiah,

National Taitung University. Taiwan

3 Paper ID: 20240031, 147~151

Novel Design of Blockchain based IIoT Framework for Smart Factory

Ms. Ahyun Song, Prof. Euiseong Seo, Prof. Heeyoul Kim,

Sungkyunkwan University. Korea(South)

4 Paper ID: 20240467, 152~157

A Reference Implementation of Blockchain Interoperability Architecture Framework

Mr. Harish V, Ms. Swathi R, Mr. Satyanarayana N,

CDAC. India

5 Paper ID: 20240034, 158~162

Multiple Merged Structures Based on Image Recognition for Converting Application of Natural Language Artificial Intelligence Service

Prof. Hui-Yu Huang, Mr. Ming-Hsun Tsai, Dr. Ming Shen Jian,

National Formosa University. Taiwan

6 Paper ID: 20240079, 163~167

Pipeline Based Genetic Algorithm for Patient Scheduling in Hospital Outpatient Department and Laboratory

Dr. Ming Shen Jian, Mr. Cheng-He Wang, Mr. Wei-Siou Wu, Mr. Tzu-Wei Hunag,

National Formosa University. Taiwan

Intelligent Anomaly Detection System Based on Ensemble and Deep Learning

Babu Kaji Baniya
Dept. of Computer Science & Information Systems
Bradley University
Peoria, IL, United States
bbaniya@fsmail.bradley.edu

Thomas Rush
Dept. of Computer Science & Information Systems
Bradley University
Peoria, IL, United States
trush@mail.bradley.edu

***Abstract*—The ubiquity of the Internet plays a pivotal role in connecting individuals and facilitating easy access to various essential services. As of 2022, the International Telecommunication Union (ITU) reports that approximately 5.3 billion people are connected to the internet, underscoring its widespread coverage and indispensability in our daily lives. This expansive coverage enables a myriad of services, including communication, e-banking, e-commerce, online social security access, medical reporting, education, entertainment, weather information, traffic monitoring, online surveys, and more. However, this open platform also exposes vulnerabilities to malicious users who actively seek to exploit weaknesses in the virtual domain, aiming to gain credentials, financial benefits, or reveal critical information through the use of malware. This constant threat poses a serious challenge in safeguarding sensitive information in cyberspace. To address this challenge, we propose the use of ensemble and deep neural network (DNN) based machine learning (ML) techniques to detect malicious intent packets before they can infiltrate or compromise systems and applications. Attackers employ various tactics to evade existing security systems, such as antivirus or intrusion detection systems, necessitating a robust defense mechanism. Our approach involves implementing an ensemble, a collection of diverse classifiers capable of capturing different attack patterns and better generalizing from highly relevant features, thus enhancing protection against a variety of attacks compared to a single classifier. Given the highly unbalanced dataset, the ensemble classifier effectively addresses this condition, and oversampling is also employed to minimize bias toward the majority class. To prevent overfitting, we utilize Random Forest (RF) and the dropout technique in the DNN. Furthermore, we introduce a DNN to assess its ability to recognize complex attack patterns and variations compared to the ensemble approach. Various metrics, such as classification accuracy, precision, recall, F1-score, confusion matrix are utilized to measure the performance of our proposed system, with the aim of outperforming current state-of-the-art intrusion detection systems.**

***Index Terms*—cybersecurity, deep neural network, ensemble, generalizing**

I. Introduction

Internet connectivity has brought about a tremendous transformative impact in various domains, encompassing communication, information sharing, and the provision of goods and services. According to statistics from the ITU, approximately 5.3 billion individuals were connected to the internet in 2022, as illustrated in Figure 1 [1]. This figure reflects the incremental growth of internet users from 2005 to 2018, with a substantial 24% increase since 2019, particularly following the onset of the pandemic. With the rapid expansion of internet access comes numerous advantages, but it also introduces formidable cybersecurity challenges. In the virtual environment, just like in the physical world, individuals with malicious intentions are consistently active around the clock [2]. To safeguard computer systems and network resources from unauthorized access and protect critical information and user credentials, several robust cybersecurity measures are employed. These measures include the implementation of firewalls, data encryption, various authentication techniques, antivirus software, and intrusion detection systems, among others [3]. These security practices are pivotal in defending against cyber threats, although they do not provide absolute guarantees of protection for computer systems and networks. Cybersecurity experts emphasize that cyberattacks are concerted efforts aimed at undermining the fundamental principles of confidentiality, integrity, and availability (CIA) within computer systems [4], [5].

Each cyber attack has unique sophisticated technique that causes the severe flaw of security measure (tools) in detection (before compromise the system). For example, denial of service (DoS) attack prevents the legitimate user for accessing the network and host computer, distributed denial of service (DDoS) attacks accomplish by flooding the ACK to target system/network using different sources to make service unable to user/s, and malware, characterized as a malevolent piece of software, is meticulously crafted to inflict harm upon computers, networks, and manipulate user data [6], [7]. This category encompasses an array of malicious entities such as computer viruses, worms, trojan horses, ransomware, spyware, and other insidious code [8], [9].

A low-footprint attack aims to minimize traces and evade remaining undetected for as long as possible, allowing attackers to achieve their goals with a reduced risk of being discovered. Many research studies and innovative ideas have already been put forward to develop an intelligent Intrusion Detection System (IDS) as a solid line of defense against low-footprint attacks. The IDS is classified into two major categories: Misuse-based Intrusion Detection System (MDS) and Anomaly-based Intrusion Detection System (ADS) [5], [10], [11]. MDS monitors network traffic or host traces to match observed behaviors against known threats and their indicators of compromise (IoCs), such as malicious network attacks, file hashes, byte sequences, etc. Although it provides higher detection rates and lower false positive rates (FPRs), it cannot identify zero-day attacks [6] or even variants of existing attacks. Moreover, it requires significant effort and expertise to frequently update the threat, involving a set of rules for each attack type [12]. On the other hand, an ADS creates a legitimate profile of network or host events and, using learning algorithms, detects any deviation from it as an anomaly. As it can detect both existing and new attacks, including zero-day attacks, and unlike MDS, does not require effort to generate rules or search for known IoCs. It simply identifies out-of-ordinary patterns better to trigger alerts than MDS when its detection method is well designed [3].

Despite the unweighted advantages of ADS, it encounters several challenges in terms of its applications. These challenges include dealing with dynamic environments since systems and networks evolve continuously, requiring constant updates and baseline monitoring [13]. Another challenge is scalability, as it may struggle to effectively monitor large and complex networks. This is because networks consist of various components, software, and platforms, each handling significant data volumes, high data rates, and a wide variety of dimensionality, making it more difficult for ADS to operate efficiently [3]. The backbone of ADS techniques typically includes ML, data mining, statistical models, fuzzy sets, knowledge bases, and various other methods and tools to detect and identify anomalies in network and system behavior [14]–[16]. These techniques are the fundamental building blocks of ADS.

UNSW-NB15 encompasses nine distinct types of cyber attack classes, each exhibiting unique attack patterns. Recognizing that a single classifier may struggle to effectively capture all nine patterns, including one for the normal class, we adopt an ensemble approach as best practice. An ensemble combines multiple classifiers, leveraging the strengths of each; if one classifier fails to grasp a particular attack pattern, others may fill the gap, enhancing network defense and safeguarding critical information. Although dataset comprises numerous features, their significance in detecting attack patterns varies. To address this variability, we incorporate a feature selection algorithm to identify the most relevant features for our feature pool [18], [19]. Additionally, we also introduce DNNs for their prowess in learning complex patterns and representations from input features. DNNs introduce non-linearities through activation functions, enabling them to model intricate relationships and capture dependencies within the features. Their ability to autonomously discover relevant features enhances pattern recognition. Thus, we employ DNNs for the detection of both normal and attack patterns, including the nine sub-categories.

Following the structured organization of the paper, Section II delves into the exploration of the anomaly-based method, providing a comprehensive understanding of this approach. In Section III, we conduct a meticulous description of the dataset used in the study. Moving forward, Section IV investigates the experimental results of both RF and DNN classifiers, presenting key findings and engaging in a thorough discussion. Finally, Section V encapsulates the paper, offering a succinct yet insightful conclusion that summarizes key takeaways and outlines potential directions for future work.

II. Anomaly-based Intrusion Detection

Anomaly-based scenarios present a multitude of challenges. Firstly, the uneven distribution of samples among classes, where one class significantly outweighs the other, may introduce a bias toward the majority of samples, posing a hurdle for effective machine learning models. The imbalance in the UNSW dataset is notably evident, as highlighted in Table I. Class samples manifest unexpected variations, exemplified by the Normal class with the highest number of samples at 37,000 for training and 56,000 for testing, while Worms exhibit lower numbers at 44 for training and 130 for testing. Intruders continually evolve their techniques to circumvent existing security measures, presenting a challenge for traditional machine learning models, which may struggle to adapt without substantial modification, updates, or immediate human intervention. Additionally, the identification of relevant attack features, as shown in Figure 4 (the vertical axis represents the selected features, and the horizontal axis represents the feature importance), is crucial for precise anomaly detection in complex and high-dimensional datasets. To tackle this challenge, we introduce a feature selection technique to identify pertinent features from a feature pool [17]. Furthermore, we employ an ensemble approach, specifically random forest, for complex pattern recognition. The introduction of deep neural networks further enhances detection capabilities, collectively addressing the multifaceted challenges in anomaly misuse detection.

Due to the substantial variation in class samples within the UNSW-NB15 dataset [18], we addressed the imbalance by augmenting the size of the minority class through the generation of new instances. While this strategy enhances the model's ability to learn the minority class pattern, it also poses the risk of overfitting. To mitigate this risk, we carefully chose RF as an ensemble learning algorithm. RF constructs multiple decision trees and combines their predictions, offering resilience against overfitting compared to individual decision trees. Similarly, to counteract overfitting in DNNs, we implemented dropout as a regularization technique. Dropout functions by randomly deactivating a fraction of neurons in a layer during training, preventing the co-adaptation of hidden units and promoting independence among neurons. This dual approach contributes to a more robust and generalizable model.

TABLE I: UNSW-NB15 training and testing samples distribution in each class.

Class		Training samples	Testing samples
Normal		37,000	56,000
Attack	Generic	18871	40,000
	Exploits	11,132	33,393
	Fuzzers	6,062	18,184
	DoS	4,089	12,264
	Reconnaissance	3,496	10,491
	Analysis	677	2,000
	Backdoor	583	1,746
	Shellcode	378	1,133
	Worms	44	130

III. Dataset Description

We employed the well-known intrusion detection UNSW-NB15 dataset to evaluate the performance of our proposed system. This dataset comprises normal and attack categories, with a particular focus on nine distinct sub-categories of attacks, namely: Backdoors, DoS, Exploits, Fuzzers, Reconnaissance, Shellcode, Analysis, Generic, and Worms shown in Table II. The dataset is further divided into a training set and a testing set, with the training set comprising 82,332 samples and the testing set containing 175,341 samples, resulting in a total of 257,673 data samples. It's worth noting that both the training and testing datasets are imbalanced (in term of classes: normal and attack, and corresponding sub-classes samples of attack), and we presented their distribution in percentages using pie charts in Figure 2.

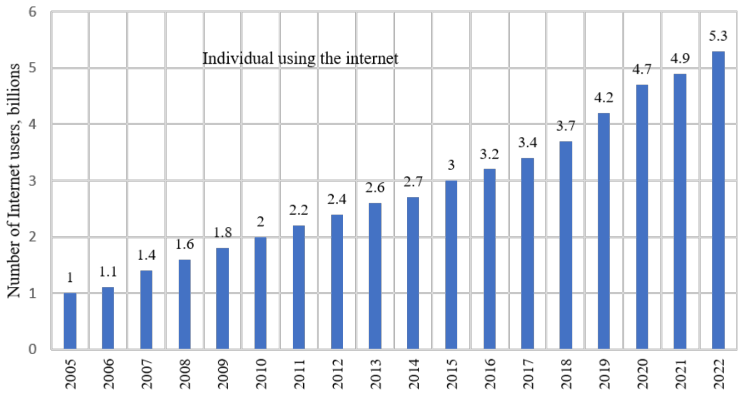


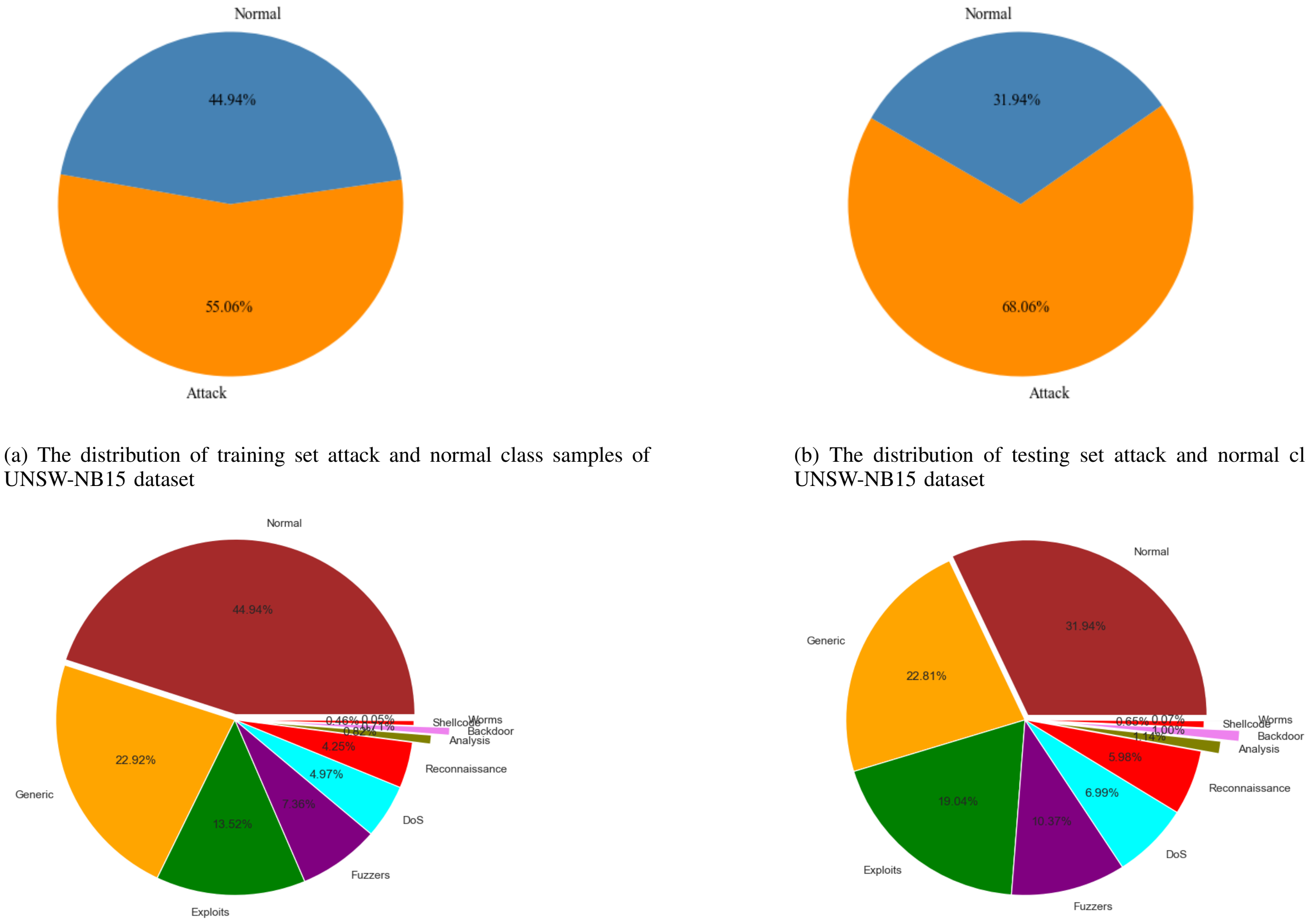
Fig. 1: It shows the increasing trend of internet users in the world's population in each year since 2005, the vertical axis presents the number of internet users in billion (around 66% of the world population using the internet) and horizontal axis represents year.

TABLE II: UNSW-NB15 dataset contains the nine different sub-types of attacks and their corresponding description [18].

Attack types	Description
Worms	They are self-replicating malicious software programs that can spread across computer networks and systems without any user intervention.
Shellcode	It is designed to be injected into a target system to run specific commands and scripts, providing unauthorized access to the system.
Reconnaissance	The preliminary phase of an attack where an attacker gathers information about an entry point of vulnerable target system or network and this information is used for preparation of future attack.
Generic	A variety of different attack types that do not fit into the other. categories.
Exploits	The pieces of code that take advantage of vulnerabilities or weaknesses in system to gain unauthorized access.
DoS	It attacks disrupt the normal function of a system or network and makes the service unavailable to legitimate users.
Fuzzers	Launch attacks by sending random data to a system, assessing its resilience, and identifying vulnerabilities.
Analysis	Attack involves system analysis to identify weaknesses and potential targets for exploitation.
Backdoors	An unauthorized or hidden access point is created within a system or software, allowing attackers to gain access even after security measures have been implemented.

IV. Experimental Results and Discussions

The assessment of our proposed system's performance involves a comprehensive analysis utilizing various metrics, such as accuracies, precision, recall, F1-score, and the receiver operating characteristic (ROC) curve. To ensure the resilience and consistency of the anomaly-based intrusion detection system, we conducted a thorough evaluation using RF and DNN classifiers. The outcomes of these classifiers were meticulously recorded and calculated based on specific formulas using four different terms: True Positive (TP), which represents instances when the system correctly detects anomalies in the dataset; True Negative (TN), denoting cases where the system correctly identifies the absence of anomalies; False Positive (FP), indicating instances where the system wrongly detects anomalies in the absence of risk in the dataset; and False Negative (FN), representing cases where the system fails to detect anomalies when the risk is present in the dataset. High precision and recall values signify the proposed model's accuracy in predictions while minimizing the omission of true positive instances, showcasing its ability to generalize effectively to unseen instances



(a) The distribution of training set attack and normal class samples of UNSW-NB15 dataset

(b) The distribution of testing set attack and normal class samples of UNSW-NB15 dataset

(c) Normal and nine attack sub-categories samples distribution of UNSW-NB15 training set in pie chart

(d) Normal and nine attack sub-categories samples distribution of UNSW-NB15 testing set in pie chart

Fig. 2: Distribution of UNSW-NB15 dataset samples to their corresponding classes and sub-classes.

of the minority class. F1-score, considering both FP and FN, proves valuable for assessing the overall performance of the model on an imbalanced dataset like UNSW-NB15. Additionally, we visualize the ROC to gauge the model's ability to distinguish between positive and negative instances across varying probability thresholds. A high area under the curve signifies the model's effective discrimination between classes, a crucial aspect in evaluating its performance across different decision thresholds.

1) Accuracy: it estimates the ratio of risk recognized of the entire conditions (cases). If accuracy is higher, the machine learning model is better.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

$$F1 - score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4}$$

The confusion matrix for the UNSW-NB15 dataset, employing a RF classifier, reveals the model's effectiveness in distinguishing between Normal and Attack categories shown in Table III. With a high true positive count of 54,733 for Normal instances, the model excels in correctly identifying genuine Normal instances. However, a false negative count of 1,267 indicates instances where the model misclassifies actual Normals as Attacks. On the Attack side, the model correctly identifies 103,424 instances but erroneously classifies 15,917 instances as Normal. The overall accuracy stands at 90.20%, signifying the proportion of correct classifications. The F1-score, a harmonized measure of precision and recall, is robust at 90.45%. The precision of 91.98% underscores the accuracy of Normal predictions among instances classified as positive, while the recall of 90.20% indicates the model's capability to capture most actual Normal instances. This suggests that the RF classifier exhibits strong performance in classifying instances within the UNSW-NB15 dataset, achieving a balanced and accurate prediction of Normal and Attack categories.

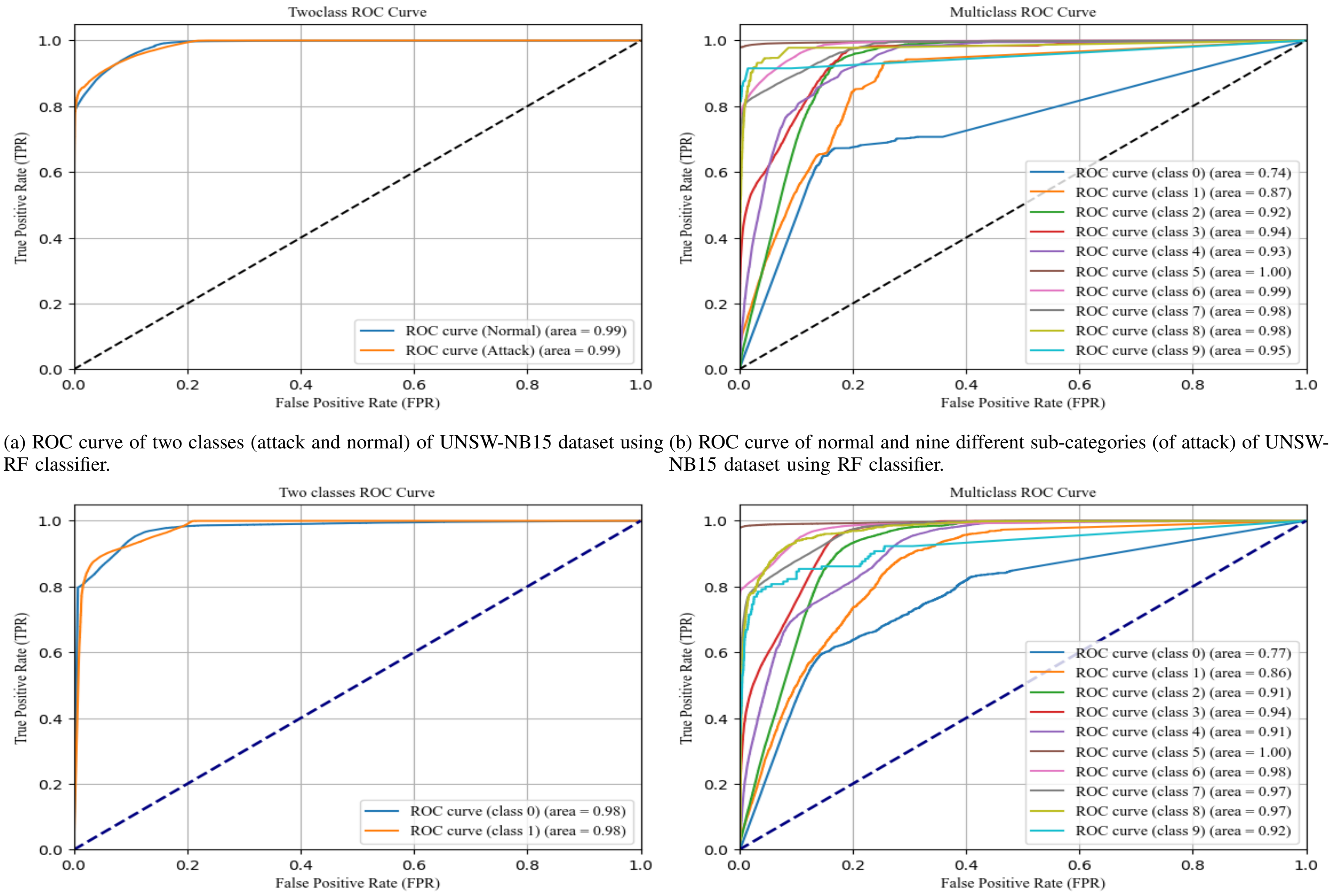
TABLE III: Confusion matrix of UNSW-NB15 dataset of Normal and Attack categories by using RF classifier.

	Normal	Attack
Normal	**54,733**	1,267
Attack	15,917	**103,424**

TABLE IV: Confusion matrix of UNSW-NB15 dataset of Normal and Attack categories by using DNN classifier

	Normal	Attack
Normal	**54,878**	1,122
Attack	19,305	**100,036**

The confusion matrix for the UNSW-NB15 dataset, employing a DNN algorithm, provides insights into the model's performance in distinguishing between Normal and Attack categories shown in Table IV. In the Normal class, the model achieves a high true positive count of 54,878 instances, indicating its ability to accurately identify genuine Normal instances. However, a false negative count of 1,122 suggests instances where the model misclassifies actual Normals as Attacks. On the Attack side, the model correctly identifies 100,036 instances but misclassifies 19,305 instances as Normal. The overall accuracy stands at 88.35%, representing the proportion of correct classifications.



(a) ROC curve of two classes (attack and normal) of UNSW-NB15 dataset using RF classifier.

(b) ROC curve of normal and nine different sub-categories (of attack) of UNSW-NB15 dataset using RF classifier.

(c) ROC curve of two classes (attack and normal) of UNSW-NB15 dataset using DNN classifier.

(d) ROC curve of normal and nine different sub-categories (of attack) of UNSW-NB15 dataset using DNN classifier.

Fig. 3: ROC curve of two classes (Normal and Attack), and nine different attack sub-categories of UNSW-NB15 dataset using RF and DNN.

TABLE V: Confusion matrix of UNSW-NB15 dataset of normal and all sub-categories of attack by using RF classifier.

	Analysis	Backdoor	DoS	Exploits	Fuzzers	Generic	Normal	Reconnaissance	Shellcode	Worms
Analysis	**43**	103	1124	17	1	105	601	1	5	0
Backdoor	37	**226**	1,144	105	17	74	114	13	14	2
DoS	295	761	**7,974**	1,368	92	734	864	50	123	3
Exploits	381	911	10,067	**18,143**	262	1,089	1,850	505	154	31
Fuzzers	43	105	1140	272	**2,578**	128	13,832	11	73	2
Generic	4	8	262	288	23	**39,318**	84	4	8	1
Normal	0	1	18	398	605	4	**54,912**	47	15	0
Reconnaissance	47	163	1,340	802	24	130	226	**7,744**	14	1
Shellcode	0	0	34	127	25	14	262	21	**649**	1
Worms	0	0	2	51	1	2	5	2	0	**67**

TABLE VI: Confusion matrix of UNSW-NB15 dataset of normal and sub-categories of attack by using DNN classifier.

	Analysis	Backdoor	DoS	Exploits	Fuzzers	Generic	Normal	Reconnaissance	Shellcode	Worms
Analysis	**46**	48	1,275	56	1	0	521	11	42	0
Backdoor	37	**80**	1,326	77	31	2	60	73	51	9
DoS	313	356	**9,131**	1,095	94	20	578	229	432	16
Exploits	416	499	12,708	**14,964**	514	137	1,858	1,426	692	179
Fuzzers	57	54	1,458	232	**5,068**	194	10,429	361	317	14
Generic	15	11	357	275	41	**39,165**	73	29	25	9
Normal	8	2	181	311	1,961	20	**53,247**	181	84	5
Reconnaissance	49	63	1,573	299	44	1	425	**7,959**	78	10
Shellcode	1	0	17	56	34	2	157	155	**708**	3
Worms	0	0	7	52	3	2	7	4	5	**50**

The F1-score, a balanced measure of precision and recall, is robust at 88.63%. The precision of 90.93% emphasizes the accuracy of Normal predictions among instances classified as positive, while the recall of 88.35% indicates the model's capability to capture most actual Normal instances. This suggests that the DNN algorithm exhibits strong performance in classifying instances within the UNSW-NB15 dataset, achieving a balanced and accurate prediction of Normal and Attack categories.

The evaluation of multiclass intrusion detection performance, as indicated by the confusion matrix for the UNSW-NB15 dataset using the RF classifier, provides valuable insights into the classification accuracy of various classes shown in Table V. Notably, the classes Analysis, Backdoor, and Exploits exhibit a high degree of overlap with the DoS class, implying shared characteristics among these categories. Additionally, these classes show misclassifications with the Normal class, indicating similarities in their features. The Fuzzers class, in

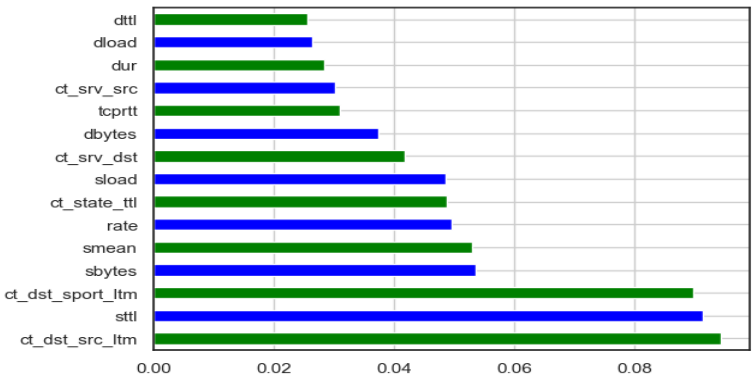


Fig. 4: It shows the feature ranking and their corresponding importance using RF of UNSW-NB15 dataset.

particular, demonstrates a pronounced overlap with the Normal class compared to other classes in the dataset. While the Worms class has relatively few samples, it shows an overlap with the Exploits class. On the contrary, the remaining classes are accurately classified by the RF classifier. The overall classification accuracy metrics are as follows: accuracy (75.17%), F1 score (72.93%), precision (77.45%), and recall (75.17%). These metrics collectively reflect the model's ability to correctly classify instances across different classes, with a notable focus on its accuracy and precision in handling the unique characteristics of each class. Almost the similar result obtained from DNN classifier and results shown in VI.

The Table VII presents a comparative analysis of the performance measures between RF and DNN classifiers, assessing their effectiveness in both two-class and multiclass scenarios. In the two-class classification, RF outperforms DNN across various metrics. RF achieves a higher accuracy (90.20%) compared to DNN (88.35%), and a superior F1-score (90.45%) compared to DNN (88.63%). The precision of RF (91.98%) also exceeds that of DNN (90.93%). Moving to the multiclass setting, RF maintains its dominance, exhibiting a higher accuracy (75.17%), F1-score (72.93%), and precision (77.44%) compared to DNN (74.33%, 73.07%, and 77.86%, respectively). These results emphasize the robust performance of RF in both two-class and multiclass classification scenarios, highlighting its efficacy in accurately classifying instances across various metrics. Our overall results are also highly competitive with the baseline model [3] which originally collected the UNSW-NB15 dataset.

TABLE VII: Different performance measures of RF and DNN classifiers in %.

Classifier	Accuracy	F1-score	Precision	Recall
RF (Two-class)	90.20	90.45	91.98	90.20
DNN (Two-class)	88.35	88.63	90.93	88.35
RF (Multiclass)	75.17	72.93	77.44	75.17
DNN (Multiclass)	74.33	73.07	77.86	74.33

ROC curve is a graphical representation of a classifier's performance across various threshold settings shown in Figure 3. It illustrates the trade-off between true positive rate (TPR) and false positive rate (FPR) at different classification thresholds. Area under the ROC curve measures the model's ability to distinguish between classes shown in Figure 3a. An area of 0.99 for both the "Normal" and "Attack" classes indicates very high performance in terms of classification. An AUC of 0.99 suggests that the RF classifier has an excellent ability to separate between the "Normal" and "Attack" classes, showcasing strong performance in terms of true positive rate and false positive rate. Higher AUC values generally indicate better model performance. Similarly, we presented the area under the ROC curve of multiclass (Normal and 9 different sub-categories attack) RF classifier in Figure 3b. We also plotted the area under ROC curve for both two class and multiclass DNN classifier, and shown in Figure 3c and Figure 3d.

V. Conclusions

We introduced an ADS utilizing RF and DNN classifiers to identify diverse anomaly and normal patterns. Both classifiers effectively distinguish various intruder patterns, and we explored crucial attack features for precise anomaly detection in complex, high-dimensional datasets. To address this challenge, we introduced a feature selection technique to identify pertinent features and minimize computational complexity. With multiple subcategories of attacks, each with distinct characteristics, RF demonstrated the ability for complex pattern recognition. Additionally, DNN was implemented to further enhance detection capabilities, collectively addressing multifaceted challenges in anomaly detection. The highly imbalanced UNSW-NB15 dataset prompted us to implement oversampling, carefully designing RF and DNN to prevent overfitting. We evaluated performance using diverse metrics, including classification accuracy, class label accuracy using the confusion matrix, precision, recall, F1-score, and ROC curve. The consistent and convincing results obtained from both classifiers underscore the effectiveness and reliability of the proposed method.

Our immediate plan involves implementing time-based features for intrusion detection to precisely detect evolving intrusion patterns over time, enhancing host or network computer security.

Acknowledgment

This material is based on work supported by the National Science Foundation under Grant Number HBCU-EiR-2101181. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

References

[1] Source link: https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

[2] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, and S. Venkatraman, "Robust Intelligent Malware Detection Using Deep Learning," in IEEE Access, vol. 7, pp. 46717-46738, 2019.

[3] N. Moustafa, J. Slay, and G. Creech, "Novel Geometric Area Analysis Technique for Anomaly Detection Using Trapezoidal Area Estimation on Large-Scale Networks," in IEEE Transactions on Big Data, vol. 5, no. 4, pp. 481-494, 1 Dec. 2019.

[4] R. Heady, G. F. Luger, A. Maccabe, and M. Servilla, "The Architecture of a Network Level Intrusion Detection System," Albuquerque, NM, USA: Dept. Comput. Sci., College Eng., Univ. New Mexico, 1990.

[5] N. Moustafa, and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in Proc. Military Commun. Inf. Syst. Conf., pp. 1–6, 2015.

[6] T. Giannetsos and T. Dimitriou, "Spy-sense: Spyware tool for executing stealthy exploits against sensor networks," in Proc. 2nd ACM Workshop Hot Topics Wireless Netw. Secur. Privacy, pp. 7–12, 2013.

[7] Y. Ye, Tao Li, D. Adjeroh, and S. S. Iyengar, "A Survey on Malware Detection Using Data Mining Techniques," ACM Comput. Surv. 50, 3, Article 41, May 2018.

[8] M. Ligh, S. Adair, B. Hartstein, and M. Richard, "Malware Analyst's Cookbook and DVD: Tools and Techniques for Fighting Malicious Code," Wiley Publishing, 2010.

[9] M. Sikorski, and A. Honig, "Practical Malware Analysis: The Hands-On Guide to Dissecting Malicious Software," Computers and Security, v6, pp. 802-803, 2012.

[10] W. Lee, S. J. Stolfo, and K. W. Mok, "A data mining framework for building intrusion detection models," in Proc. IEEE Symp. Secur. Privacy, pp. 120–132, 1999.

[11] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maci-a-Fern-andez, and E. V-azquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," Comput. Secur., vol. 28, no. 1, pp. 18–28, 2009.

[12] Z. Tan, A. Jamdagni, X. He, P. Nanda, and R. P. Liu, "A system for denial-of-service attack detection based on multivariate correlation analysis," IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 2, pp. 447–456, Feb. 2014.

[13] G. Creech and J. Hu, "A semantic approach to host-based intrusion detection systems using contiguousand discontiguous system call patterns," IEEE Trans. Comput., vol. 63, no. 4, pp. 807–819, Apr. 2014.

[14] W. Hu, J. Gao, Y. Wang, O. Wu, and S. Maybank, "Online adaboost-based parameterized methods for dynamic distributed network intrusion detection," IEEE Trans. Cybern., vol. 44, no. 1, pp. 66–82, Jan. 2014.

[15] A. Jamdagni, Z. Tan, X. He, P. Nanda, and R. P. Liu, "RePIDS: A multi tier real-time payload-based intrusion detection system," Comput. Netw., vol. 57, no. 3, pp. 811–824, 2013.

[16] M. Almseidin and S. Kovacs, "Intrusion detection mechanism using fuzzy rule interpolation," arXiv preprint arXiv:1904.08790, 2019.

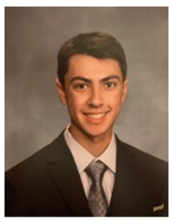
[17] B. K. Baniya, "Intrusion Representation and Classification using Learning Algorithm," 2022 24th International Conference on Advanced Communication Technology (ICACT), PyeongChang Kwangwoon-Do, Korea, Republic of, pp. 279-284, 2022.

[18] N. Moustafa, "Designing an online and reliable statistical anomaly detection framework for dealing with large high-speed network traffic," Diss. University of New South Wales, Canberra, Australia, 2017.

[19] B. K. Baniya and E. Z. Gnimpieba, "The Effectiveness of Distinctive Information for Cancer Cell Analysis Through Big Data," In: Arai, K., Kapoor, S. (eds) Advances in Computer Vision. CVC 2019. Advances in Intelligent Systems and Computing, vol 944. Springer, Cham.

[20] B. K. Baniya and J. Lee, "Importance of audio feature reduction in automatic music genre classification," Multimed Tools Appl 75, 3013–3026 (2016). https://doi.org/10.1007/s11042-014-2418-z

Babu Kaji Baniya holds a B.E. degree in Computer Engineering from Pokhara University, Nepal, which he obtained in 2005. He further pursued his education and completed an M.E. and Ph.D. in Department of Computer Science and Engineering from Chonbuk National University, Republic of Korea in 2015. Following his doctoral studies, he gained valuable experience as a postdoctoral researcher in the Department of Computer Science and Biomedical Engineering at the University of South Dakota. He then served as an assistant professor in the Department of Computer Science and Digital Technologies at Grambling State University, Louisiana. Currently, he holds the position of assistant professor in the Department of Computer Science and Information Systems at Bradley University, located in Peoria, Illinois. Throughout his career, he has taught a wide range of Computer Science courses at both the graduate and undergraduate levels. His research interests span several key areas, including audio signal processing, information retrieval, cybersecurity, bioinformatics, Big Data, and machine learning. A specific focus of his research involves the application of machine learning and deep learning algorithms in securing the Internet of Medical Things (IoMT). He is also IEEE member.



Thomas Rush is currently pursuing a B.S. degree in Computer Science with a concentration in Data Science at Bradley University in Peoria, Illinois. His research experience includes working as a Research Assistant to Dr. Baniya at Bradley University, where he focused on classifying Android malware using a variety of machine learning models and deep learning techniques. Thomas has applied these skills in practical settings, having completed coursework in Data Science, Machine Learning, Data Mining, and Artificial Intelligence at Bradley University. Additionally, he contributed to data analytics for the Bradley University Men's Soccer Team through his senior capstone project.

A Private Blockchain System based on Zero Trust Architecture

Yao-Chung Chang [a], Yu-Shan Lin [b], Arun Kumar Sangaiah[c] and Hsin-Te Wu [a]

[a] Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan

[b] Department of Information Science and Management Systems, National Taitung University, Taitung, Taiwan

[c] Department of International Graduate School of Artificial Intelligence, National Yunlin University of Science and Technology, Yunlin, Taiwan

Email: ycc.nttu@gmail.com, ysl.nttu@gmail.com, arunks@yuntech.edu.tw, wuhsinte@nttu.edu.tw

Abstract— **During the outbreak of COVID-19, many enterprises massively created Virtual Private Networks (VPNs) for companies to cooperate; however, these accounts lacked efficient management after the epidemic, leading to data leakage or the suffering of malicious attacks. Consequently, several firms have started to build private blockchains for data preservation and verification. Private blockchains are usually built internally in companies; once a firm's internal network opens for massive external logins, many servers and private blockchains cannot work properly to protect data. In general, private blockchains store critical personal information or relevant confidential data; once a private blockchain opens to external access, all the information in the network nodes will be exposed, making private them lose their protection functions.**

This paper proposes a security mechanism using a private blockchain system based on zero trust architecture. The zero trust architecture tracks every user's network conditions and analyses whether their behaviors are authorized. Additionally, the system utilizes micro-segmentation to divide the private blockchain, preventing the system from malicious attacks. The proposed system employs user multi-factor authentication to identify users, and the zero trust architecture tracks and analyzes if users' behaviors are reasonable. This method effectively ensures corporate networks' security and enables private blockchain to filter legal and authorized users to access and verify.

Keywords— **Zero Trust Architecture, Virtual Private Networks, Blockchain**

I. INTRODUCTION

Despite the rapid development of technology and the Internet, corporate network architectures are mostly conventional types. However, many technological companies require information access across external and internal networks, making corporate networks even more unsafe. Moreover, many employees and partner suppliers have started working remotely, and firms massively create VPN accounts to maintain their business operations, lowering corporate networks' protection. As a result, once hackers access a company's internal network through VPNs to attack or tap, they will cause network paralysis or data breaches of various vital servers. Private blockchains protect networks internally, preventing external users from accessing or adding data. When the internal protection fails, external users can access information on private blockchains freely. As companies usually build private blockchains for storing confidential data, the function of a private blockchain can avoid servers' misfunctions that stop the users who genuinely need to access important data. The setup of private blockchains allows specific groups of users to access data; therefore, it is necessary to develop more authentication methods and protection measures to secure the safety of private blockchains.

This research presents a private blockchain mechanism based on zero trust architecture using an intrusion-detection system and firewall. We also utilize the intrusion-detection system to construct the micro-segmentation mechanism, which will separate high-risk equipment and make malicious attacks fail to penetrate the separated parts. The operation of micro-segmentation prevents unauthorized users from accessing the specific equipment. Meanwhile, this study further analyzes user behaviors to judge the rationality of each user; when discovering unusual actions, the intrusion-detection system will terminate the user's access. The multi-factor authentication mechanism used in this research verifies if a user possesses the authorization to access the private blockchain. The experiment proves that the proposed approach of this article can enhance the safety of private blockchains; in other words, the research implements network attacks using popular techniques to confirm whether the attacks invade the separated equipment. Consequently, the experimental result has proven that the proposed method is effective.

II. RELATED WORK

Literature [1] demonstrates a method to store patients' electronic medical records and the Internet of bodies' data in blockchains, resolving the massive storage issue of medical data; furthermore, blockchains possess the features of verification and distribution, which can ensure data safety and completeness. In Literature [2], the article presents a lightweight blockchain mechanism that employs a chameleon hash function to verify user identities and data completeness; the function allows users to generate a secret key freely between themselves and verify data completeness directly. The experimental result of the research proves that the approach has reached a level of data security with lightweight encryption and decryption. On the other hand, Literature [3] constructs a blockchain system to deliver employees' performance; the system establishes records for employees to prove their monthly workloads and confirms credibility through data authentication. In the future, enterprises can evaluate and conduct employees' job appraisals by reviewing their historical workload records. The eVoting system presented in Literature [4] suggests a secured technique to evaluate voters' identities; afterward, blockchains in the system, which have features of better immutability and completeness, will process the voting, achieving the goal of an eVoting system. Meanwhile, the traceable agricultural products system developed in Literature [5] combines blockchain technology to tackle the conventional supply chain issues of agricultural products effectively; additionally, the system designs a user-friendly interface for farmers to input data easily. Literature [6] aims to improve the source verification of vaccines; the system uses blockchains to inspect vaccine distribution, and users can review relevant information through the system. Finally, Literature [7] builds a peer-to-peer energy trading system; the immutability characteristic of blockchains enables users to process peer-to-peer transactions without seeking a third-party certification.

III. THE PROPOSED SYSTEM

A. System Model

In our research, the zero trust architecture developed by an intrusion-detection system and a firewall is utilized to establish a security mechanism using private blockchains, as shown in Figure 1. Firstly, the firewall divides equipment into different parts and verifies identities; afterward, the intrusion-detection system creates Virtual Local Area Networks (VLANs) for separation, preventing malicious users from accessing critical equipment. Meanwhile, the intrusion-detection system will also track users' online behaviors; when detecting unusual behaviors, the system temporarily suspends the user's action. Users' unusual behaviors vary; for example, when a user's login location changes dramatically within one hour, it probably means the user has experienced hackers' invasion and conducted attacks in many countries using IP hopping. The proposed design also contains a multi-factor authentication mechanism, using various identity authentication methods to verify legal users. For instance, users can access private blockchains to conduct further actions when confirmed and authorized; if they fail to pass the verification three times, the system will lock their IP locations. This study aims to create a complete network security architecture, making private blockchains safer and avoiding data leakage.

B. Zero Trust Architecture

The next-generation firewall can operate on the first layer of the open systems interconnection model, namely the physical layer, without needing to change the setting of other adjacent network equipment; moreover, the firewall works on the connection modes of the first through the third layers. As shown in Figure 2, the setup of the next-generation firewall is the same as setting up a core switch; the next-generation firewall distributes network interfaces, and the cables connect to related office equipment, such as network switches and wireless network devices. Unlike a core switch in a conventional network setup, where server cables can connect directly to a core switch, the average port cost of the next-generation firewall is higher. Furthermore, because the port numbers are fewer after transferring to the next-generation firewall, several servers have to share 1 Gbps of bandwidth. Hence, engineers must consider whether the network bandwidth is sufficient to fulfill business needs. The study enhances the transmission bandwidth using the Link Aggregation Control Protocol standard to enrich access efficiency, reserving the original architecture with the defined VLAN and allowing the next-generation firewall to distribute the routing among VLANs.

C. Multi-factor authentication

The multi-factor authentication mechanism used in this study primarily employs the APP-ID technique developed by Palo Alto Networks, a precise traffic identification and classification system. Unlike the conventional fourth-layer firewall, the APP-ID technology does not rely on port connection or communication protocols to identify applications but utilizes various mechanisms to check them. Such a technique provides visibility, operational status, behavior traits, and relevant risk information regarding the applications, preventing the possible methods of evading inspection used by hackers. Furthermore, the system builds security rules by application types to initiate, inspect, or block unnecessary applications, as shown in Figure 3. Apart from the APP-ID technique, the proposed mechanism further checks users' personal information using different authentication methods, such as birthdays and phone numbers; users who pass the authentication can access the private blockchain system.

IV. EXPERIMENT RES

The experimental environment is set in an intranet; we employ three devices to attack the system and test if the micro-segmentation can effectively reduce attack ranges and risks. Additionally, our research further connects the system to Palo Alto Networks, simulating hackers' attacks on the Internet to evaluate the firewall performance on threat defense. Finally, we conduct lateral movement to the data center; Table 1 presents the attack approaches utilized in this study. In sum, our research implemented various malicious attacks and found that all the attack ranges were within the areas of micro-segmentation, proving that the proposed method can effectively reduce attack ranges.

Table 1、Experiment type

Attack model	coping strategies
Antivirus/anti-malware	Firewall IPS Intrusion Prevention
Network intrusion prevention	Firewall IPS Intrusion Prevention
Restrict web-based content	Firewall IPS Intrusion Prevention
Endpoint behavior prevention	Firewall IPS Intrusion Prevention
Implement prevention	Firewall IPS Intrusion Prevention
Password policy	Multi-factor authentication
Network traffic filtering	Firewall IPS Intrusion Prevention
Restrict access to resources over the network	Firewall IPS Intrusion Prevention

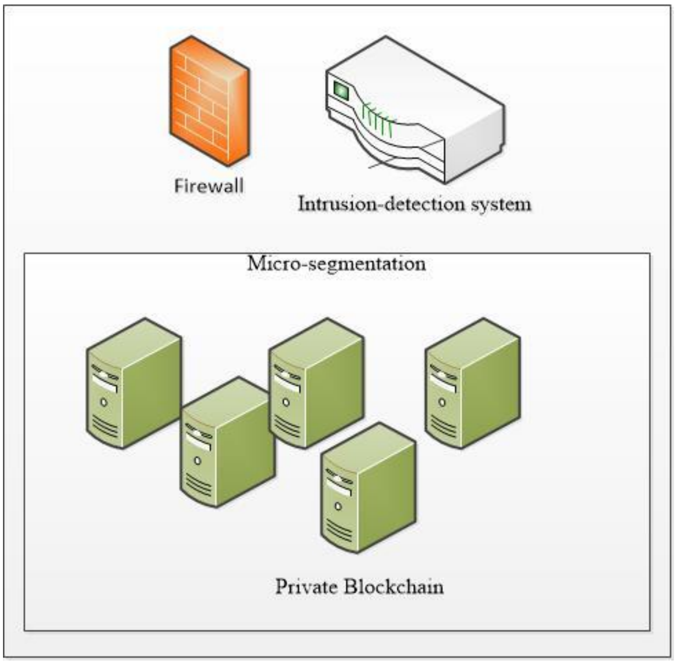


Figure 1. System Model

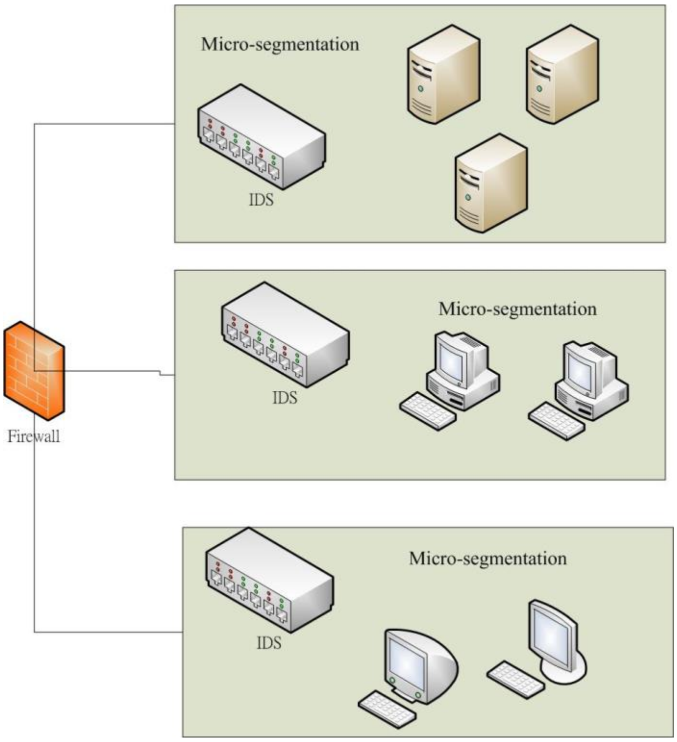


Figure 2. Zero Trust Architecture

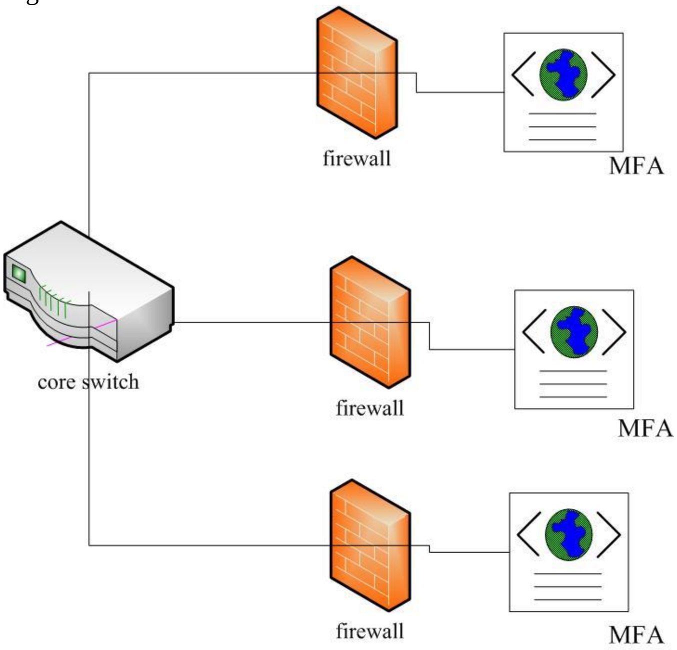


Figure 3. Multi-factor authentication

V. Conclusions

With the advance of emerging technologies, conventional network architectures cannot fulfill today's technical needs; moreover, those corporate VPN accounts created during the outbreak of COVID-19 might increase the risks of cyberattacks if hackers invade. This article presents a security mechanism using a private blockchain system based on zero trust architecture, utilizing micro-segmentation and user behavior analytics to fight against the vulnerability attacks of emerging technologies. We have also combined multi-factor authentication to verify user legality. The experimental result proves that the proposed method can effectively withstand different cyberattacks, preventing separated equipment from relevant attacks.

Acknowledgment

This paper was supported by the National Center for High-Performance Computing and National Science and Technology Council, Taiwan, under grant no. NSTC 111-2410-H-143 -023, NSTC 111-2222-E-224 -004, NSTC 111-2221-E-143 -006 -MY2, NSTC 111-2410-H-143 -007, NSTC 112-2622-H-143 -001 and NSTC 111-2622-H-006 -004.

References

[1] J. Liu et al., "Conditional Anonymous Remote Healthcare Data Sharing Over Blockchain," in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 5, pp. 2231-2242, May 2023, doi: 10.1109/JBHI.2022.3183397.

[2] C. Xu, Y. Qu, T. H. Luan, P. W. Eklund, Y. Xiang and L. Gao, "A Lightweight and Attack-Proof Bidirectional Blockchain Paradigm for Internet of Things," in IEEE Internet of Things Journal, vol. 9, no. 6, pp. 4371-4384, 15 March15, 2022, doi: 10.1109/JIOT.2021.3103275.

[3] F. Wilhelmi, S. Barrachina-Muñoz and P. Dini, "End-to-End Latency Analysis and Optimal Block Size of Proof-of-Work Blockchain Applications," in IEEE Communications Letters, vol. 26, no. 10, pp. 2332-2335, Oct. 2022, doi: 10.1109/LCOMM.2022.3194561.

[4] F. D. Giraldo, B. Milton C. and C. E. Gamboa, "Electronic Voting Using Blockchain And Smart Contracts: Proof Of Concept," in IEEE Latin America Transactions, vol. 18, no. 10, pp. 1743-1751, October 2020, doi: 10.1109/TLA.2020.9387645.

[5] A. Tharatipyakul and S. Pongnumkul, "User Interface of Blockchain-Based Agri-Food Traceability Applications: A Review," in IEEE Access, vol. 9, pp. 82909-82929, 2021, doi: 10.1109/ACCESS.2021.3085982.

[6] L. Cui, Z. Xiao, F. Chen, H. Dai and J. Li, "Protecting Vaccine Safety: An Improved, Blockchain-Based, Storage-Efficient Scheme," in IEEE Transactions on Cybernetics, vol. 53, no. 6, pp. 3588-3598, June 2023, doi: 10.1109/TCYB.2022.3163743.

[7] J. Abdella, Z. Tari, A. Anwar, A. Mahmood and F. Han, "An Architecture and Performance Evaluation of Blockchain-Based Peer-to-Peer Energy Trading," in IEEE Transactions on Smart Grid, vol. 12, no. 4, pp. 3364-3378, July 2021, doi: 10.1109/TSG.2021.3056147.

Yao-Chung Chang (M’03) received the Ph.D. degree from National Dong Hwa University, Hualien, Taiwan, in 2006.

He is a Professor of the Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan. His primary research interests include intelligent communication System, IoT, and cloud computing.

Dr. Chang is a recipient of the subsidization program in universities for encouraging exceptional talent, Ministry of Science and Technology, Taiwan

Yu-Shan Lin received the Ph.D. degree from National Sun Yat-sen University, Kaohsiung, Taiwan, in 2006.

She is a Professor of the Department of Information Science and Management Systems, National Taitung University, Taitung, Taiwan. Her research interesting areas include Digital Learning, Information Technology Education, Marketing Management, Internet Marketing, and Tourism Marketing. Dr. Lin had the honor to get the Subsidy for College and University Research Rewarding from Ministry of Science and Technology (MOST).

Prof. Arun Kumar Sangaiah received his Ph.D. from School of Computer Science and Engineering, VIT University, Vellore, India. He is currently a Full Professor with National Yunlin University of Science and Technology, Taiwan. He has published more than 300 research articles in refereed journals (IEEE TII, IEEE TITS, IEEE TNSE, IEEE TETCI, IEEE SysJ, IEEE SensJ, IEEE IOTJ, ACM TOSN) 11 edited books, as well as 1 patents (held and filed) and 3 projects, among one funded by National Science and Technology Council (NSTC), Taiwan, Ministry of IT of India and few international projects (CAS, Guangdong Research fund, Australian Research Council) cost worth of 500000 USD. Dr. Sangaiah has received many awards, Yushan Young Scholar, Clarivate Highly Cited Researcher (2021,2022), Top 2% Scientist (Standord Report-2020,2021,2022), PIFI-CAS fellowship, Top-10 outstanding researcher, CSI significant Contributor etc. Also, he is responsible for Editor-in-Chief, and Associate Editor of various reputed ISI journals. Dr. Sangaiah is a visiting scientist (2018-2019) with Chinese Academy of Sciences (CAS), China and visiting researcher of Université Paris-Est (UPEC), France (2019-2020) and etc. His Google Scholar Citations reached 19800+ with h-index: 78 and i10-index: 298.

HSIN-TE WU is an Associate Professor of Department of Computer Science and Information Engineering from National Taitung University, Taiwan. He has served as Associate Editor for International Journal of Big Data and Analytics in Healthcare. He has served as Special Issue Guest Editor of Journal of Supercomputing and IET Networks. His research interests include computer networks, wireless network, speech compression, network security, blockchain and Internet of things.

Novel Design of Blockchain based IIoT Framework for Smart Factory

Ahyun Song*, Euiseong Seo*, Heeyoul Kim**

* Department of Computer Science and Engineering, Sungkyunkwan University, Republic of Korea

**Division of Computer Science and Engineering, Kyonggi University, Republic of Korea

fialle@g.skku.edu, euiseong@skku.edu, heeyoul.kim@kgu.ac.kr

***Abstract*— A smart factory is an advanced manufacturing system that utilizes various cutting-edge technologies such as IIoT (Industrial Internet of Things), big data, AI(Artificial Intelligence), blockchain to automate and optimize production processes. While there is a growing demand for smart factories in recent times, the adoption and proliferation of these facilities have been delayed due to concerns about security, the reliability of collected information, and challenges in management and control. Moreover, traditional centralized smart factory systems pose a risk of operational downtime in the event of failures or attacks since a central server controls the entire system. Therefore, we propose the design of a secure, transparent, and reliable blockchain-based IIoT framework for smart factories. The framework consists of three layers: the blockchain core layer, the blockchain operation layer, and the IIoT service layer. The IIoT service layer plays a crucial role in providing various services essential for the advancement of smart factories, utilizing blockchain technology. Our proposed framework combines IIoT and blockchain technologies to leverage the advantages of decentralization, trust, security, transparency, data management, and traceability.**

***Keywords*— Blockchain, Industrial Internet of Things(IIoT), Smart Factory, Framework, Decentralization**

I. Introduction

With the advancement of various technologies and innovative ideas such as IIoT, big data, AI, edge and cloud computing, and blockchain, the modern industry is evolving to a new level [1]. Smart factories represent an advanced manufacturing system that harnesses these latest technologies to automate and optimize production processes. Smart factories are a core concept driving the digital transformation of manufacturing industries, enhancing productivity, improving quality, and conserving resources and time.

Despite the growing demand for smart factories driven by the development of IIoT technology and intelligent information technology, their construction and proliferation have been delayed due to security concerns, reliability issues of collected information, and challenges in management and control. In the existing centralized smart factory system, a central server controls the entire system, so there is a risk of operation interruption due to failures and attacks. In particular, recent studies have shown that IIoT cloud platforms for smart factories are semi-trusted third parties, and there is a security risk in managing smart factory information [2, 3]. Therefore, to promote the activation of the smart factory industry, a trusted platform is needed to ensure the management of IIoT devices and the reliability of collected information.

Blockchain is a distributed ledger management technology that stores data in a chain-like structure of blocks connected in a peer-to-peer-based distributed data storage environment. Blockchain technology has the potential to provide innovative solutions in various industrial sectors, including reliable data exchange, smart contracts, security, and logistics management in a decentralized environment [4]. The combination of IIoT and blockchain technology in the smart factory domain can enhance safety, efficiency, and transparency, bringing innovation to various processes such as manufacturing process optimization, quality management, and distribution management.

In this paper, we design the architecture of a blockchain-based IIoT framework for smart factories. The proposed framework consists of three layers: the blockchain core layer, the blockchain operation layer, and the IIoT service layer. The blockchain core layer and blockchain operation layer provide essential technologies for the operation of the blockchain system and functionalities required for blockchain system operation in the IIoT environment. The IIoT service layer, a critical component, utilizes blockchain to offer various services necessary for the advancement of smart factories, including IIoT device management, smart data management, and automated control services.

Firstly, the IIoT device management service, based on blockchain, provides efficient and reliable authentication of IIoT devices, verification and management of the status information of IIoT devices, firmware updates, and history management. Secondly, the smart data management service shares data in a decentralized manner, without the need for a central DB server, leveraging blockchain technology. The irreversibility of blockchain ensures data integrity collected from IIoT devices and transparency in the data processing process. Additionally, we facilitate smart factory advancement by offering data through open APIs for machine learning and big data analysis of the collected data. Lastly, the automated control service utilizes smart contracts to automate manufacturing processes within the smart factory, inspect the manufacturing status, and provide control over the manufacturing process.

The rest of this paper is organized as follows. In Section II, we provide a brief overview of blockchain and IIoT, and review relevant research on blockchain based smart factories. Section III offers a detailed explanation of our proposed blockchain-based IIoT Framework. In Section IV, we discuss the challenges that arise when practically applying IIoT and blockchain technology to smart factories. Finally, Section V concludes the paper.

II. Related work

A. Blockchain & IIoT & Smart factory

IIoT refers to Internet of Things (IoT) technology used in industrial environments [5]. IIoT connects various devices and sensors in manufacturing, production, and industrial sectors to collect data, which can be transmitted to central servers or the cloud. This enables real-time monitoring, remote operation, and maintenance in industrial environments, leading to increased production efficiency and improved product quality. The IIoT architecture is composed of three layers: the physical layer, communication layer, and application layer [6]. The physical layer consists of physical devices such as sensors, manufacturing equipment, and data centers. The communication layer uses network technologies like wireless sensor networks, 5G, and others to integrate various devices from the physical layer. These two layers together form cyber-physical systems (CPS) to support the development of applications like smart factories [7].

Blockchain is a distributed ledger management technology where data in block form is connected in a chain within a peer-to-peer based distributed data storage environment. This technology ensures data safety and integrity and utilizes peer-to-peer networks for data sharing and verification [8]. In blockchain, anyone can read the data, but tampering with it is difficult, ensuring the reliability of the data. This is a crucial characteristic of distributed ledger management, which guarantees the trustworthiness of data. Nodes participating in the blockchain store and share the same records, and any changes to the data require consensus among the nodes. Therefore, as the number of participating nodes increases, blockchain management becomes more complex, but decentralization is reinforced. Decentralization is one of the core features of blockchain, enabling the provision of services in entirely different forms compared to traditional centralized services.

Blockchain technology has the potential to offer innovative solutions in various industries, such as reliable data exchange, smart contracts, security, and logistics management, within this decentralized environment [4]. Leveraging these core features of blockchain, it is possible to implement innovative technologies and services, such as data management, enhanced security, real-time tracking, and automation, in smart factories and manufacturing environments.

B. Blockchain based Smart Factory

[3] proposes a solution that combines Blockchain technology with IIoT architecture for smart factories. The proposed architecture consists of five layers: the sensing layer, the management hub layer, the storage layer, the firmware layer, and the application layer. The Sensing layer includes various types of sensors and at least one microcomputer, which collect information from various devices and preprocess the collected data. The Management Hub layer integrates and coordinates various devices and responds to user requirements in real time. The Storage layer functions as a data center, specifically designed for storing data using a private blockchain. Additionally, the proposed architecture enhances data security and privacy by employing a whitelist mechanism and asymmetric encryption mechanism.

According to [2], while there are many mature IIoT cloud platforms for smart factories, securely storing and sharing smart factory data on these platforms remains a challenge. To address this issue, [2] proposes a blockchain-based security access scheme for IIoT in smart factories that supports traceability and revocability. In this proposed scheme, the blockchain performs unified identity authentication and stores all public keys, user attribute sets, and revocation lists.

[9] presents a blockchain system that considers both security and efficiency for power-constrained IIoT devices in smart factories. The proposed credit-based consensus mechanism can reduce the power consumption of honest devices. Additionally, the proposed data authority management method protects data privacy without affecting system performance.

III. Blockchain based IIoT Framework

The combination of IIoT and blockchain technology enhances safety, efficiency, and transparency in the field of smart factories, bringing innovation to various application areas such as manufacturing process optimization, quality management, and supply chain management. A more detailed explanation of the key advantages resulting from the combination of IIoT and blockchain technology is as follows:

- **Decentralization:** Blockchain utilizes a distributed network instead of centralized systems. This means that data and processes within smart factories are not controlled from a central authority but are distributed and managed among multiple nodes. This enables the storage and management of reliable, distributed data and enhances resilience to device failures or faults. Important decisions and transactions can be executed without the need for central intermediaries, thereby strengthening trust.
- **Trust and Security:** Blockchain technology enhances data integrity and security. It provides assurance that data blocks on the blockchain are not altered or tampered with, ensuring the trustworthiness of the data. Storing data related to manufacturing processes and product information on the blockchain within a smart factory helps prevent tampering or fraudulent activities,

ensuring product safety. Additionally, encryption techniques can further enhance data security.

- **Transparency:** Blockchain provides transparency of data, allowing all relevant parties to access and track transactions. Recording all data related to the manufacturing process within a smart factory on the blockchain offers transparency regarding production stages, product origins, quality control, and distribution. This fosters greater trust between producers and consumers and provides information about product safety and quality.
- **Data Management and Traceability:** Blockchain facilitates data tracking and management across all stages, from raw materials to the final product. This enables capabilities such as tracking defective products, recall management, and optimizing production processes. Manufacturers can monitor the production and distribution processes of their products in real-time and take prompt action in case of issues.

In this paper, we design the architecture of a blockchain-based IIoT service framework for smart factories. As shown in Figure 1, the proposed framework operates on top of a communication layer where IIoT devices are interconnected and aims to provide services that support the operation, control, and data analysis of the higher-level smart factory. This framework is internally composed of three main layers:

- **Blockchain Core Layer:** This layer provides the core technologies required for the operation of the blockchain system. It consists of 1) cryptographic algorithms including hash functions, elliptic curve cryptosystems, and ECDSA electronic signatures, 2) consensus algorithms suitable for IIoT environments, 3) transaction management for the structure, storage, and verification of transactions, and 4) ledger management components including technologies such as ledger storage and merkle tree.
- **Blockchain Operation Layer:** This layer offers functionalities necessary for operating the blockchain system within the IIoT environment. It encompasses smart contracts, which are programs executed on the blockchain, governance mechanisms that determine policies and rules for blockchain operation, and components responsible for authenticating and controlling access to participating nodes in the blockchain.
- **IIoT Service Layer:** This layer provides blockchain based IIoT services for smart factories. It includes services for managing IIoT devices that comprise the smart factory, a smart data management service for processing data collected from IIoT devices, and an automated control service for controlling the manufacturing process of the smart factory.

In particular, the IIoT Service Layer is a crucial layer for delivering various services essential for advancing smart factories using blockchain. Here are detailed descriptions of the IIoT services provided in this layer:

IIoT Device Management

- **Device authentication service:** In a smart factory environment, due to the participation of a large number of IIoT devices, authentication is essential to verify the legitimacy of devices and defend against the infiltration of malicious devices by attackers. It is difficult to apply traditional authentication methods to heterogeneous IIoT devices with different hardware, communication methods, and security policies. Authentication services based on blockchain can provide efficient and highly reliable authentication for IIoT devices.
- **Firmware update service:** The firmware of IIoT devices needs to be updated frequently, and there is a high probability that an attacker's malware will be secretly propagated and infiltrated into the device during this process. To prevent such attacks, it is

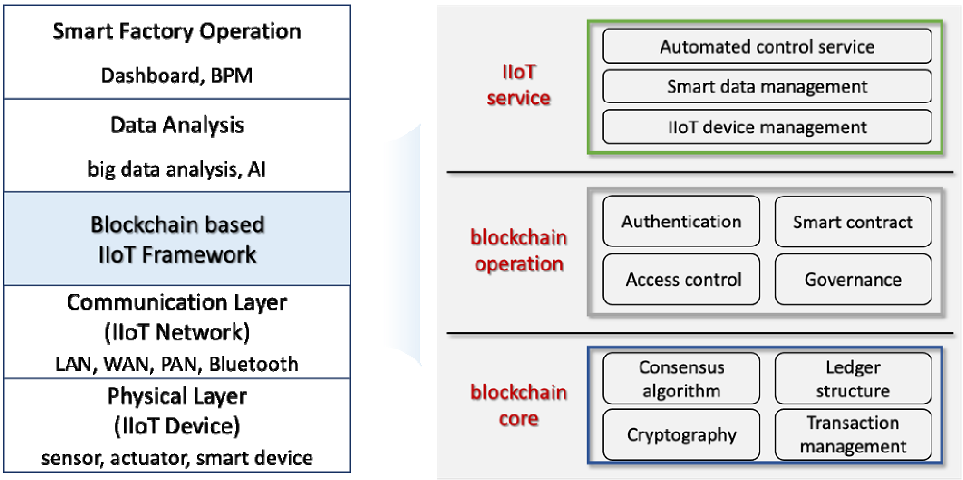


Figure 1 Proposed Blockchain based IIoT Framework for Smart Factory

necessary to ensure the safety of the update process and firmware, and existing methods have difficulty securing centralized update servers. Blockchain based firmware update services have the advantage of providing firmware integrity and reliability, and transparently tracking and verifying update history.

- **Device status managing service:** Smart factories require management technology that can efficiently maintain a large number of IIoT devices of various types. In particular, management technology that can collect and analyze the overall life cycle status information from device installation to operation status check, location movement, malfunction detection, and removal is required. The blockchain-based device status managing service secures the reliability of device management and enables efficient device status information sharing.

Smart Data Management

- **Data sharing service:** Various types of data, such as sensing information from IIoT devices and information about the progress of manufacturing processes, are generated during manufacturing processes. Depending on the nature of the information, it needs to be shared between factory managers and devices. Data sharing service, based on blockchain, enables decentralized data sharing without the need for a central database server. Furthermore, it facilitates information sharing and collaboration between multiple smart factory hubs.
- **Data validation service:** Data collected through IIoT devices is susceptible to errors at each stage of collection, transmission, and storage, and there is a high risk of malicious manipulation to disrupt manufacturing processes. To prevent these issues, it is essential to ensure the integrity of collected data and the transparency of data processing. The data validation service, based on the irreversibility of blockchain, verifies data to guarantee its reliability.
- **Open data service:** Collected data needs to be supplied to external entities to assist in making policy decisions regarding smart factory operations while ensuring that sensitive manufacturing information is not leaked. The open data service provides the functionality to supply collected data to the external world through open APIs. In particular, it supports the enhancement of smart factories by providing clean data to upper layers such as machine learning and big data analysis.

Automated Control Service

- **Automated manufacturing process service:** This service provides automation of manufacturing processes within the smart factory using smart contracts. It supports the definition, creation, deployment, and execution of smart contracts tailored to the manufacturing processes defined through analysis of the smart factory's manufacturing environment and requirements. It ensures that predefined processes are automatically executed when specific conditions are met. Additionally, it leverages blockchain technology to enable fault-tolerant progress of manufacturing processes even in the presence of certain failures.
- **Manufacturing control service:** This service offers functionalities to inspect the manufacturing status within the smart factory and control the manufacturing processes. It verifies the environmental information of the smart factory and the information related to manufacturing processes. It facilitates easy process control for administrators and provides access control and security policy management to prevent the leakage of confidential information in the manufacturing process. Furthermore, when integrated with financial services, this service can serve as a foundation for the end-to-end integration of processes, from product contracts and orders to production, distribution, and payments.

IV. Discussion

Despite the various advantages of combining IIoT and blockchain technology, there are challenges when it comes to practical implementation in smart factories.

- **Data Privacy:** There can be conflicts between the transparency of blockchain and data privacy. Blockchain's nature of publicly sharing data may make it challenging to protect sensitive production data and personal information in smart factories. Finding a way to manage data access rights while maintaining data integrity is necessary. Additionally, while transactions in blockchain are processed anonymously, in some cases, specific data can be linked to certain companies or devices, posing privacy concerns. Striking a balance between anonymity and identifiability is essential.
- **Energy Efficiency:** Blockchain can consume a significant amount of energy due to its complex encryption and consensus processes. On the other hand, IIoT devices often operate in low-power environments, necessitating the minimization of power consumption. Therefore, implementing blockchain in smart factories should consider energy efficiency.
- **Throughput Restrictions:** Most blockchains have limitations on transaction processing speed, making it challenging to handle large-scale real time data processing in smart factories. Especially in large-scale smart factory environments, scalability issues in the blockchain network may arise. Therefore, effective strategies are required to manage numerous IIoT devices and large datasets.
- **Security:** Some blockchain networks face risks of malicious activities like 51% attacks, which can impact the security of smart factory data. Additionally, smart contracts may be exposed to threats due to code bugs or security vulnerabilities, which need to be considered.

Taking these challenges into account, we have designed and proposed a secure and practical blockchain-based IIoT framework in Section III. As future tasks, we aim to conduct research on technologies that can realistically implement the proposed framework and develop a blockchain-based IIoT system for smart factories. The combination of blockchain and IIoT is a crucial factor driving innovation in the smart factory sector, and by overcoming these challenges, we can establish a safe and efficient production environment.

V. Conclusions

In this paper, we designed the architecture of a secure, transparent, and reliable blockchain-based IIoT framework for smart factories. The proposed framework consists of three layers: the blockchain core layer, the blockchain operation layer, and the IIoT service layer. The blockchain core layer and blockchain operation layer provide essential technologies for the operation of the blockchain system and functionalities required for blockchain system operation within the IIoT environment. The IIoT service layer is a crucial layer that leverages blockchain to offer various services necessary for the enhancement of smart factories, including the authentication of trustworthy IIoT devices, verification and management of the status information of IIoT devices, transparent and secure sharing and verification of information collected through IIoT devices, and automated control services for manufacturing processes within smart factories. The combination of IIoT and blockchain technology offers a variety of advantages, including decentralization, trust, security, transparency, data management, and traceability.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No. 2020R1A6A1A03040583)

References

[1] D. Miller, "Blockchain and the internet of things in the industrial sector," IT professional, vol. 20, no. 3, pp. 15-18, 2018.

[2] K. Yu, L. Tan, M. Aloqaily, H. Yang, and Y. Jararweh, "Blockchain-enhanced data sharing with traceable and direct revocation in IIoT," IEEE transactions on industrial informatics, vol. 17, no. 11, pp. 7669-7678, 2021.

[3] J. Wan, J. Li, M. Imran, and D. Li, "A blockchain-based solution for enhancing security and privacy in smart factory," IEEE Transactions on Industrial Informatics, vol. 15, no. 6, pp. 3652-3660, 2019.

[4] J. Leng et al., "Blockchain-secured smart manufacturing in industry 4.0: A survey," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 1, pp. 237-252, 2020.

[5] T. M. Fernandez-Carames and P. Fraga-Lamas, "A review on the application of blockchain to the next generation of cybersecure industry 4.0 smart factories," IEEE Access, vol. 7, pp. 45201-45218, 2019.

[6] B. Chen, J. Wan, L. Shu, P. Li, M. Mukherjee, and B. Yin, "Smart factory of industry 4.0: Key technologies, application case, and challenges," IEEE Access, vol. 6, pp. 6505-6519, 2017.

[7] T. Alladi, V. Chamola, R. M. Parizi, and K.-K. R. Choo, "Blockchain applications for industry 4.0 and industrial IoT: A review," IEEE Access, vol. 7, pp. 176935-176951, 2019.

[8] N. Mohamed and J. Al-Jaroodi, "Applying blockchain in industry 4.0 applications," in 2019 IEEE 9th annual computing and communication workshop and conference (CCWC), 2019: IEEE, pp. 0852-0858.

[9] J. Huang, L. Kong, G. Chen, M.-Y. Wu, X. Liu, and P. Zeng, "Towards secure industrial IoT: Blockchain system with credit-based consensus mechanism," IEEE Transactions on Industrial Informatics, vol. 15, no. 6, pp. 3680-3689, 2019.

Ahyun Song received the M.S. degree in Computer Science from KAIST, Korea, in 2005. From 2005 to 2015 she was a manager at Korea Financial Telecommunications & Clearings Institute. Since 2015 she has been a senior manager at Financial Security Institute in Korea. She is pursuing the Ph.D. degree in Computer Science and Engineering at Sungkyunkwan University. Her major research interests include security, blockchain, and DeFi.

Euiseong Seo received his B.S., M.S., and Ph.D. degree in computer science from Korea Advanced Institute of Science and Technology (KAIST) in 2000, 2002, and 2007, respectively. He is currently a professor in Department of Computer Science and Engineering at Sungkyunkwan University, Rep. of Korea. Before joining Sungkyunkwan University in 2012, he had been an assistant professor at Ulsan National Institute of Science and Technology (UNIST), Rep. of Korea from 2009 to 2012, and a research associate at the Pennsylvania State University from 2007 to 2009. His research interests are system software, embedded systems, and cloud computing.

Heeyoul Kim received the B.E. degree in Computer Science from KAIST, Korea, in 2000, the M.S. degree in Computer Science from KAIST in 2002, and the Ph.D. degree in computer science from KAIST in 2007. From 2007 to 2008, with the Samsung Electronics as a senior engineer. Since 2009 he has been a faculty member of Department of Computer Science at Kyonggi University. His major research interests include cryptography, security and blockchain.

A Reference Implementation of Blockchain Interoperability Architecture Framework

Harish V, Swathi R, Satyanarayana N

e-Security Department, Centre for Development of Advanced Computing, Hyderabad, India

vharish@cdac.in, rswathi@cdac.in, nanduris@cdac.in

***Abstract*—In this paper we describe interoperability architecture between blockchain networks. A blockchain network can serve multiple organizations that form as a consortium for dealing with transactions among themselves. However, when multiple blockchain networks are deployed how do we support interoperability and what are the major concerns in this direction have to be studied. Towards this direction, we developed a reference architecture for smooth transmission of transactions between the blockchain networks using custom designed transaction flow mechanism on top of Cactus framework from Linux Foundation.**

Keywords—blockchain, interoperability, Hyperledger fabric, Hyperledger sawtooth

I. Introduction

There is no single blockchain platform which caters to the needs of each and every application domain. Certain blockchain platforms have been designed keeping a particular application domain in view (for e.g., corda is considered to be well suited for financial related applications). Hence there is a need for exploring opportunities and technical feasibilities for providing cross chain support between different domains.

Suppose when existing blockchain based applications prefer to switch to a new or an upgraded blockchain platform there is a need for interoperating with legacy blockchain platforms. In such cases, cross chain or multichain is an option where multiple chains coexist and interoperate with each other corresponding to an application domain. Further when a blockchain network managed by a consortium of organizations want to interoperate with another blockchain administered by a separate consortium of organizations we need effective interoperable blockchain solutions.

Scheme/Platform	**Hash Time Lock Contracts**	**Oracles**	**Notary Scheme**	**Relays/Side Chains**
Polkadot [7]	No	No	No	Yes
Cactus	Yes	No	Yes	No

In literature survey, we can find many different interoperability approaches suggested by researchers [1], [2], [3] and some of them have been implemented also. Each proposed technique of interoperability has its own purpose. For example, Hash Time Lock contracts can be used for confirmation of transaction upon confirming payment between the respective parties.

In our approach, we have taken Cactus [4] an interoperability framework from Hyperledger community as a base code and built up on it a transaction flow model in the context of interacting applications running on different blockchain platforms. Without the transaction flow mechanism in place with the available source code the logic to deal with exchange of transaction information across the blockchain had to be dealt by the cactus middleware. This results in a bottleneck situation as the cactus middleware code has to be redesigned as and when more networks are to be added by integrating necessary logic to deal with transaction information exchange between them. We studied this aspect and have come out with a unique strategy of making cactus middleware responsible for information exchange only based on the pre-define transaction flow information. The transaction flow information can be updated on the fly as and when new networks or applications contexts in which different blockchain networks are to be operated.

II. Proposed Blockchain Interoperability framework

A. Architecture

In the proposed architecture, we support two different types of communication mechanism across the blockchain platforms that needs to be interoperate with each other as follows.

Types of communication:

a. Immediate Confirmation: In this type of communication, a client submits a transaction to one of the Blockchain networks which gets confirmed on that particular Blockchain network. Later, the transaction particulars will be forwarded to another network through middleware. The second blockchain network upon receiving the transaction particulars from the middleware, processes it and confirms the same on its side.

b. Deferred Confirmation: In this type of communication, a client submits a transaction to one of the Blockchain networks which will be forwarded to the other Blockchain network may be after initial processing through a middleware. The second blockchain network upon receiving the transaction particulars from the middleware, processes it and confirms the same on its side and sends back the transaction to its initiator through the same middleware. The first Blockchain network upon receiving the transaction particulars from the second blockchain network, it processes the transaction on its side.

The block level diagram of the Interoperability architecture in the context of National Blockchain Framework (NBF) is as shown in figure 1 below.

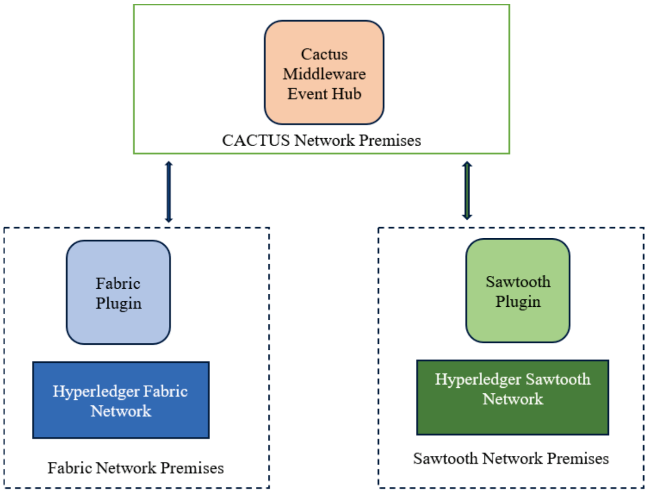

Cactus Middleware Event Hub
CACTUS Network Premises
Fabric Plugin
Hyperledger Fabric Network
Fabric Network Premises
Sawtooth Plugin
Hyperledger Sawtooth Network
Sawtooth Network Premises

Figure 1: Block level diagram of NBF Interoperability Architecture

Fig 1., depicts the block level diagram of NBF Interoperability solution. In this architecture, a fabric network [5] administrator deploys the Fabric plugin network component on their side and a sawtooth network administrator deploys the Sawtooth plugin network component on their side. The cactus-based middleware event hub can be installed on mutually agreed premises.

The respective plugins listen for transaction block events from corresponding blockchain networks and shares them with each other in accordance with the transaction flow logic controlled by the cactus middleware event hub. The transaction block event comprises details about all the transactions hold by a particular block that is generated by the corresponding blockchain network.

The transaction flow logic that is handed over to the cactus middleware event hub is the core component that determines above mentioned types of communication. We will delve in to each component in details in later sections.

Fig 2., below depicts an example deployment scenario of NBF's Interoperability solution between Hyperledger Fabric and Hyperledger Sawtooth [6] networks.

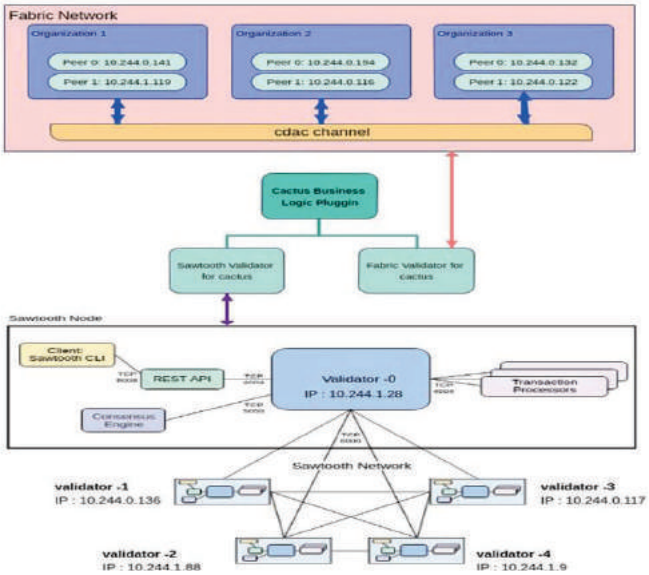


Figure 2.: Example Deployment Scenario of Interoperability Solution

B. *Essential Requirements of Operating Proposed Framework*

Network Perspective:

The administrators of the Fabric, Sawtooth and Cactus Middleware Event Hub components must be knowing about YAML (Yet Another Markup Language) syntax. A detailed tutorial about various directives written in YAML in different configuration files can be found in the developed software.

1) *Fabric Network*

A Hyperledger Fabric network must be in running state and its admin identity credentials from its wallet must be shared with the Fabric plugin administrator by external way.

The fabric plugin component must be configured with fabric network details hence its details must be made available to the Fabric plugin administrator. Ideally, the plugin would be installed on Fabric network side only hence the fabric network administrator only can administer plugin configuration also.

2) *Sawtooth Network*

A Hyperledger Sawtooth network must be in running state and its admin identity credentials from its wallet must be shared with the Sawtooth plugin administrator by external way.

The sawtooth plugin component must be configured with sawtooth network details hence its details must be made available to the Sawtooth plugin administrator. Ideally, the plugin would be installed on sawtooth network side only hence the sawtooth network administrator only can administer plugin configuration also.

3) *Cactus Middleware Event Hub*

The fabric network administrator and sawtooth administrator have to share their respective plugin details with the Cactus Middleware Event Hub administrator. Apart from that the cactus middleware event hub administrator also should get transaction flow details between the applications deployed on Fabric and Sawtooth networks in consultation with the respective network's smart contract developers.

Developer Perspective

1) Chain code / Smart code Developer

The chain code developer has to upgrade their current chain code/smart contract in order to support interoperability as per the tutorial provided in developed software. The tutorial is made available for JavaScript developers.

A sample code snippet in the tutorial being provided through the software is described below.

Let's assume that we have a smart contract with the name electricityAsset.js in which updateMeter(ctx, meterID, owner, usage) is one of the functions. It's code looks like the below.

```
async UpdateMeter(ctx, meterID, usage) {
      let updatedMeter = {};
      updatedMeter.meterID = meterID;
      updatedMeter.usage = usage;
      ctx.stub.putState(meterID,
Buffer.from(JSON.stringify(updatedMeter)));
      return;
   }
```

The code written above is self-explanatory hence no detailed description is provided here. Now, in case, if the application owner would like to interoperate with any other Blockchain network, he/she has to revise the code of UpdateMeter. In the below example, it is envisaged that as and when client application sends its request to the Hyperledger Fabric network specifying the name of chaincode function to be invoked i.e., UpdateMeter_db and an object with request parameter property as a mandatory property and dbserver, nextSeq property as an optional property.

For example,

```
contract.submitTransaction("UpdateMeter_db",{requestparam
:{meterID:101, usage:32});
```

Here, the request parameters that the client used to send in legacy code have been embedded into an object as requestparam. Apart from that two optional properties (dbserver and nextSeq) also can be set to convey to the chaincode that mongo dbserver is required to be used for fetching additional information and nextSeq value that indicates the cactus middleware event hub about its next course of action as defined in the transaction-flow.yaml configuration.

2) User Roles

Hyperledger Fabric Network Administrator – Responsible for monitoring Hyperledger Fabric network and its configuration as well as fabric plugin node monitoring and its configuration.

Hyperledger Sawtooth Network Administrator – Responsible for monitoring and managing Hyperledger Sawtooth network and its configuration as well as Sawtooth Plugin Node monitoring and its configuration.

Cactus Middleware Event Hub Tx Flow Configurator – Responsible for defining tx flow configuration in consultation with Fabric and Sawtooth network's smart contract developers in a specific application context.

Hyperledger Sawtooth Transaction Processor Developer – Responsible for Transaction processor development and deployment of the same on Sawtooth network in a specific application context.

Hyperledger Fabric Chaincode Developer – Responsible for smart contracts development and deployment of the same on Fabric network in a specific application context.

C. *Connecting Blockchain Networks*

The Interoperability solution has been developed for NBF using upgraded Cactus framework mechanism.

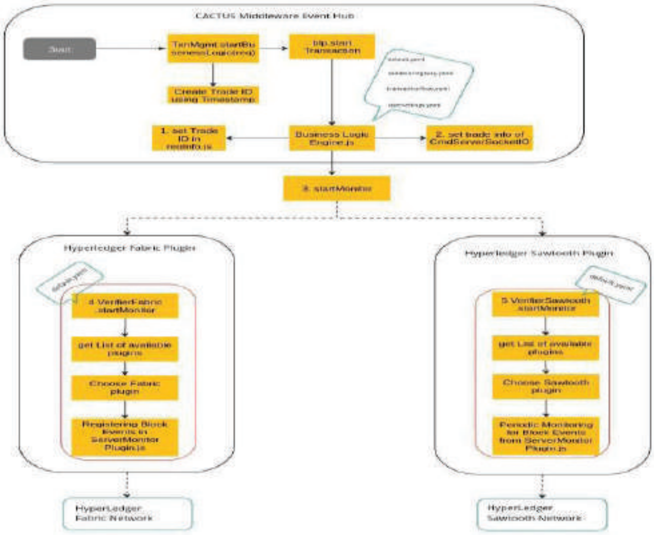

CACTUS Middleware Event Hub
Start
TxnMgmt.startBusinessLogic(req)
blp.start Transaction
Create Trade ID using Timestamp
1. set Trade ID in reqInfo.js
Business Logic Engine.js
2. set trade info of CmdServerSocketIO
3. startMonitor
Hyperledger Fabric Plugin
4. VerifierFabric .startMonitor
get List of available plugins
Choose Fabric plugin
Registering Block Events in ServerMonitor Plugin.js
Hyperledger Sawtooth Plugin
5. VerifierSawtooth startMonitor
get List of available plugins
Choose Sawtooth plugin
Periodic Monitoring for Block Events from ServerMonitor Plugin.js
HyperLedger Fabric Network
HyperLedger Sawtooth Network

Figure 3: Bootstrapping of Fabric and Sawtooth along with Cactus middleware hub

Figure 4., below depicts the initiation of client request from respective Blockchain platform network and their request processing paradigm.

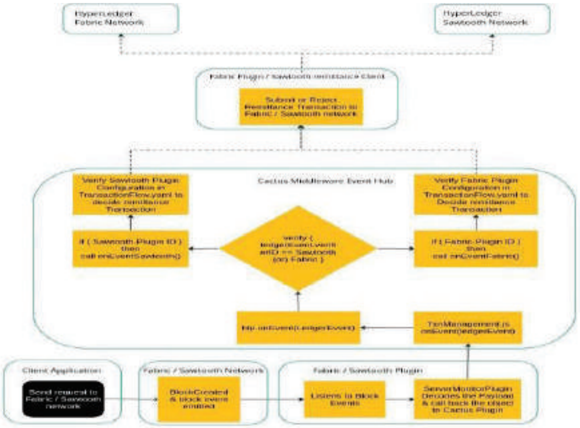
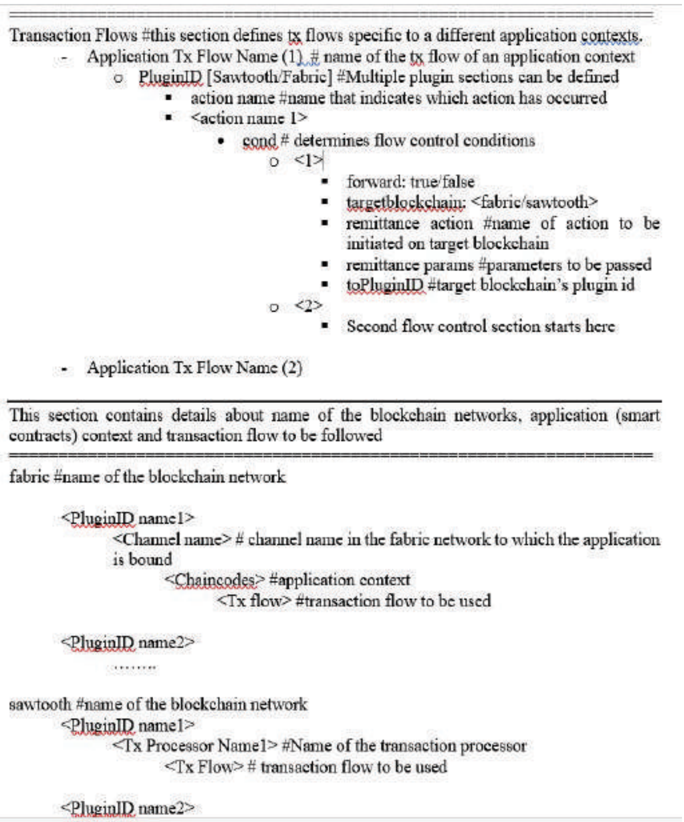

HyperLedger Fabric Network
HyperLedger Sawtooth Network
Fabric Plugin / Sawtooth remittance Client
Submit or Reject Remittance Transaction to Fabric / Sawtooth network
Cactus Middleware Event Hub
Verify Sawtooth Plugin Configuration in TransactionFlow.yaml to decide remittance Transaction
Verify Fabric Plugin Configuration in TransactionFlow.yaml to Decide remittance Transaction
if (Sawtooth Plugin ID) then call onEventSawtooth()
verify (ledgerEvent.verifierID == Sawtooth (or) Fabric)
if (Fabric Plugin ID) then call onEventFabric()
blp.onEvent(LedgerEvent)
TxnManagement.js onEvent(ledgerEvent)
Client Application
Send request to Fabric / Sawtooth network
Fabric / Sawtooth Network
BlockCreated & block event emitted
Fabric / Sawtooth Plugin
Listens to Block Events
ServerMonitorPlugin Decodes the Payload & call back the object to Cactus Plugin

Figure 4. Client request handling by Blockchain networks

D. *Understanding Tx Flow Configuration between Blocckhain platforms*

The transaction-flow.yaml file contains various sections that is defined by the Tx-Flow-Configurator in the following manner.

The below section describes about various transaction flows that can be defined between different blockchain networks
==



```
==================================================================
Transaction Flows #this section defines tx flows specific to a different application contexts.
    -   Application Tx Flow Name (1) # name of the tx flow of an application context
        o   PluginID [Sawtooth/Fabric] #Multiple plugin sections can be defined
            ▪   action name #name that indicates which action has occurred
            ▪   <action name 1>
                •   cond # determines flow control conditions
                    o   <1>
                        ▪   forward: true/false
                        ▪   targetblockchain: <fabric/sawtooth>
                        ▪   remittance action #name of action to be
                            initiated on target blockchain
                        ▪   remittance params #parameters to be passed
                        ▪   toPluginID #target blockchain's plugin id
                    o   <2>
                        ▪   Second flow control section starts here

    -   Application Tx Flow Name (2)
__________________________________________________________________
This section contains details about name of the blockchain networks, application (smart
contracts) context and transaction flow to be followed
==================================================================
fabric #name of the blockchain network

        <PluginID name1>
                <Channel name> # channel name in the fabric network to which the application
                is bound
                        <Chaincodes> #application context
                                <Tx flow> #transaction flow to be used

        <PluginID name2>
                ........

sawtooth #name of the blockchain network
        <PluginID name1>
                <Tx Processor Name1> #Name of the transaction processor
                        <Tx Flow> # transaction flow to be used

        <PluginID name2>
```

In order to support Interoperability between Hyperledger Fabric and Hyperledger Sawtooth networks (please be noted that the solution supports both homogenous and heterogenous network combinations), there is a need to upgrade existing chaincode or transaction processors (a.k.a smart contracts) in the respective networks. Hence, the smart contract developers have to stick to few suggestions as mentioned below.

At present the solution has been tested with extended fabric's chaincode version. As per the Hyperledger Fabric's chaincode syntactical requirements, the chaincode functions can be invoked by passing desired number of request parameters from the client application. However, the chaincode function signature does not allow passing of variable request parameters.

If we look at the transaction-flow.yaml directives mentioned above there is a need for passing additional parameters other than the one submitted by the client applications. Since the present chaincode function signature does not permit or useful in fulfilling this requirement there is a need for upgrading existing chaincodes.

The chaincodes have to be upgraded in such a manner that the function signature in it accepts the JSON object instead of individual parameters. The JSON object format is as follows and the client applications have to pass the JSON object in the prescribed format only.

```
obj:= {
        requestparam : {key1:val1, key2:val2,…..},
        dbobj: {dbserver: ‘mongodb’, data: {<database reply
in json format>}}
        nextSeq: <number [1…]>
}
```

In the above object format, nextSeq property is optional. Usually when the client sends his/her request while initiating transaction, this property would not be included. The chaincode developer need not handle presence or absence of this property in the chaincode functions. However, the chaincode developer MUST handle the other two properties viz., requestparam and dbobj in the chaincode functions. All functions in chaincode MUST adhere to above requirement in order to support Interoperability.

III. USE CASE SCENARIO

This section describes about a fictitious use case scenario related to electricity units consumption and bill generation by two different blockchain networks.

A. *Application Context - Electricity Trade*

In this use case, the fabric network is used for recording client requests about consumed electricity units and corresponding bill details on its side and the sawtooth network is used for recording client requests about electricity units' consumption details on its side. Here, for fabric network client may be end users who is consuming the electricity for domestic or commercial purpose whereas client w.r.t sawtooth network could be the electricity distribution agency.

Fabric clients and Sawtooth clients can hit their respective blockchain networks any time. However, the system should confirm the client request on both networks to meet the specific purpose of information maintained at respective networks.

B. *Transaction Initiation*

The sequence of events that occurs when client applications initiate transactions on either side of blockchain networks would be as follows.

Scenario 1: The electricity distributing agency has sent its request with details such as meter id, consumed units to the sawtooth blockchain network.

Step 1: Client sends his/her request to the sawtooth network.

Step 2: Sawtooth network receives the request from the client, processes it and confirms the transaction on its side.

Step 3: The sawtooth network generates a block event comprising details about confirmed transaction on its side which comprises data about meter id and consumed units information.

Step 4: The block event will be captured by the sawtooth plugin node that is associated with the sawtooth network, which in turn, read out each transaction information from the block event, and sends the transaction particulars to the cactus middleware event hub node.

Step 5: The cactus middleware event hub node upon receiving transaction particulars from the sawtooth plugin, read out the tx flow information from its configuration settings and decides upon whether the request has to be further forwarded to the fabric network or not.

Step 6: In case, the tx flow definition that is available with the cactus middleware event hub finds that the tx has to be forwarded to the fabric network, the cactus middleware event hub forwards the tx information unaltered to the fabric plugin node which in turn send the information to the fabric network to which it is associated.

Step 7: The fabric network upon receiving the tx information, processes it and generates a chaincode event to fetch the unit cost from external data source. The chaincode event will be captured by the fabric plugin node. The chaincode event contains tx data such as meter id, consumed units and data source name that needs to be contacted for further information.

Step 8: The fabric plugin upon receiving the chaincode event finds information about external data source that needs to be contacted for fetching unit cost of electricity. It then connects to the external data source (for example, mongo db) and fetches the required details. The collection name, data source url, user authentication particulars everything can be fetched from its default.yaml file.

Step 9: Upon receiving the data from the external data source, the fabric plugin sends information to the cactus middleware event hub node. The event hub node determines whether the information has to be sent back to the fabric network based on the defined settings in tx flow file.

Step 10: If cactus middleware event hub node finds it that the information has to be sent back to the fabric node from the tx flow configuration file, it forwards the request to the fabric plugin node which in turn forwards the same to its associated fabric network.

Step 11: The fabric network upon receiving the information from its fabric plugin node, processes it, compute the bill based on unit cost information received and confirms the transaction on its side.

Step 12: The fabric plugin node that is associated with the fabric network receives block events with transaction details that are confirmed by the fabric network. The fabric plugin node, then verifies each and every transaction that is found in the block event and forwards the tx particulars to the cactus middleware event hub node.

Step 13: The cactus middleware event hub node decides whether to end the transaction or next course of action based on tx flow configuration settings. In this scenario, it ends further transaction processing as the objective of recording transaction particulars on both the networks is achieved.

Scenario 2: The electricity distribution agency has sent its request with details such as meter id, consumed units to the fabric blockchain network.

All the steps defined in Scenario 1 will be followed in the same manner whenever the client sends its request to the fabric network initially. However, in this case, the transaction will not be confirmed on fabric network initially as it needs to fetch the unit cost before confirming it on its side. Once it receives the unit cost from the external data source i.e., mongo db server, it confirms it on its side and send the transaction particulars to the sawtooth network through cactus middleware event hub.

Scenario 3: Both scenario 1 & 2 have defined immediate transaction confirmation-based model, meaning, in both cases the transaction particulars are forwarded to another block chain network only after they are confirmed on respective blockchain network first.

In scenario 3, we shall brief the steps when deferred tx confirmation approach has to be followed assuming that client had sent his/her request to the Hyperledger Fabric Network initially.

Step 1: Client sends his/her request containing consumed units and meter id to the Hyperledger fabric network.

Step 2: Fabric network sends a chaincode event to the Fabric plugin node along with units consumed and meter id and data source name. In this case, the tx is not yet confirmed on peer ledger.

Step 3: Cactus middleware event hub accesses the tx flow settings, finds that it has to send the information to the Sawtooth network.

Step 4: Sawtooth plugin receives data from cactus middleware event hub node and forwards the same to the associated sawtooth network.

Step 5: Sawtooth network records the transaction event on its network and sends block event about confirmed transaction to the associated sawtooth plugin node.

Step 6: Sawtooth plugin node sends the information received from its associated sawtooth network in turn to the cactus middleware event hub node.

Step 7: Cactus middleware event hub finds from the Tx flow settings that it has to forward the confirmed Tx details to the fabric network through the associated fabric plugin node.

Step 8: Fabric plugin receives the unit cost from the mongo db server and forwards the transaction data along with unit cost to the cactus middleware event hub node.

Step 9: Cactus middleware event hub finds from the tx flow settings that it has to forward the confirmed tx details to the fabric network through the associated fabric plugin node.

Step 10: Fabric network upon receiving data from the associated fabric plugin node process it by generating the bill and confirms the transaction on its side. The fabric plugin will get the block event about the confirmed transaction subsequently from its associated fabric network and forwards the tx detail to the cactus middleware event hub node. At this point of time, the cactus middleware event hub node completes transaction processing when it sees that there is no next action to be performed from the tx flow settings file.

Fig 5, 6 and 7 demonstrate how to configure various parameters that help fabric and sawtooth plugins interconnect with their respective blockchain networks as well as with the cactus middleware hub.

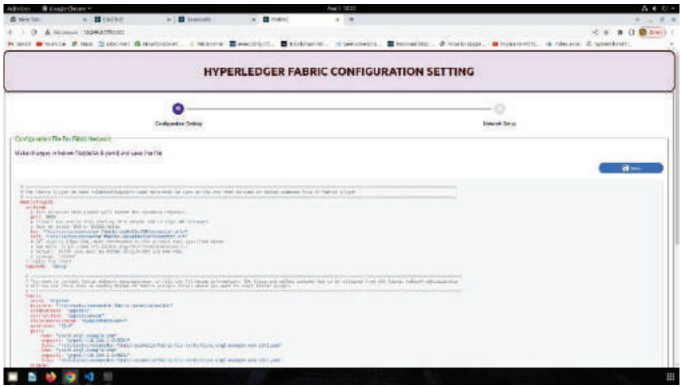


Figure 5: Hyperledger Fabric Plugin Configuration Settings

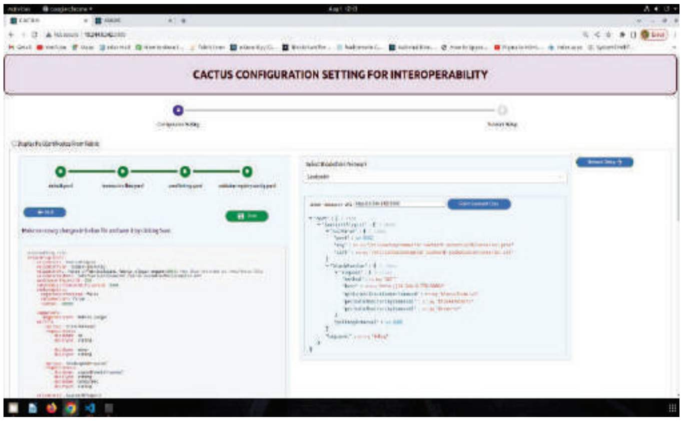


Figure 6: Sawtooth Plugin Configuration

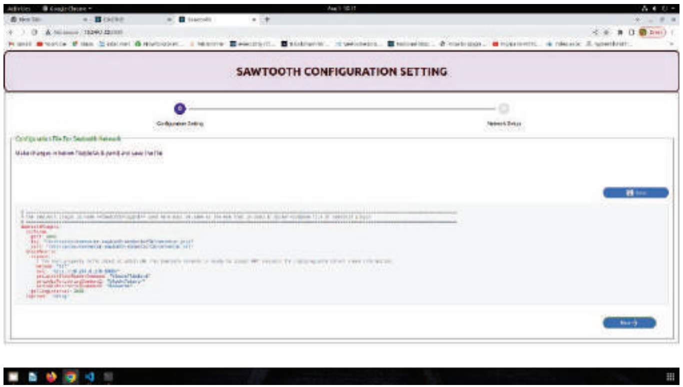


Figure 7: Cactus Plugin Configuration

The plugin configurations can be done respective administrators. The cactus plugin information include tx flow configuration also as it needs to take care of routing decision about incoming transaction events from the respective blockchain netwok

CONCLUSION

In this paper, we highlighted research approaches that are proposed by the researchers and explained about how the Blockchain interoperability solution is supported across multiple blockchain platforms. In our work, we focused more upon intricacies related to the implementation of these approached by creating a layer of transaction flow mechanism as a configurable option based on which cactus middleware takes a decision about how to deal with transactions across the platforms. The proposed architecture has been tested in the LAN environment using sample use case scenarios.

ACKNOWLEDGMENT

The work in this paper has been carried out as part of the National Blockchain Framework initiative supported by the Ministry of Electronics and Information Technology, Government of India.

REFERENCES

[1] Belchior, R., Vasconcelos, A., Guerreiro, S., & Correia, M. (2020). A Survey on Blockchain Interoperability: Past, Present, and Future Trends. *ArXiv*. /abs/2005.14282

[2] Monika and R. Bhatia, "Interoperability Solutions for Blockchain," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, India, 2020, pp. 381-385, doi: 10.1109/ICSTCEE49637.2020.9277054.

[3] S. Zhu, C. Chi and Y. Liu, "A study on the challenges and solutions of blockchain interoperability," in China Communications, vol. 20, no. 6, pp. 148-165, June 2023, doi: 10.23919/JCC.2023.00.026.

[4] Hyperledger Cactus Documentation, Retrieved from "Welcome to Hyperledger Cactus documentation! — Hyperledger Cactus 2.0.0-alpha.2 documentation (hyperledger-cactus.readthedocs.io)" on 23-11-2023

[5] Hyperledger Transaction Flow, Retrieved from "Transaction Flow — hyperledger-fabricdocs main documentation" on 23-11-2023.

[6] Creating a Hyperledger Sawtooth Network, Retrieved on "Creating a Sawtooth Test Network (hyperledger.org)" on 23-11-2023.

[7] Polkadot Blockchain Network, Retrieved from Polkadot: Web3 Interoperability | Decentralized Blockchain on 23-11-2023

Multiple Merged Structures Based on Image Recognition for Converting Application of Natural Language Artificial Intelligence Service

Hui-Yu Huang, Ming-Hsun Tsai, Ming-Shen Jian*

Dept. of CSIE, National Formosa University, Yunlin County, Taiwan 632

hyhuang@nfu.edu.tw, sean408031015@gmail.com, jianms@nfu.edu.tw*

Abstract— **This research provide the Application of Natural Language Artificial Intelligence Service based on the Multiple Merged Structures through the Image Recognition. According to the object and category/class relationships, each object can be implemented corresponding to the on demand defined data structure or category. Based on the recursively object recognition, the information corresponding to the object data structure could be chained. By merging multiple structures from the image recognition, the image can be described by the natural language as the artificial intelligence service.**

Keywords— **Image Recognition, Data Structure, Artificial Intelligence, Natural Language, Application**

I. Introduction

Recently, artificial intelligence (A.I.) services are generally used in industry, education, and even entertainment. To support the A.I. services, huge hardware resources for complex computing are needed. Although there are several popular application programming interfaces (APIs) provided by various companies through the network, some functions would be limited or the additional usage fee would be required depending on the computing missions of A.I.

Generally speaking, the small models of A.I. service could be managed and deployed more easily. However, due to the computing ability of the small models, to compute and deal with the complex scenario or mission would be difficult. However, small models are easily deployed and implemented. Therefore, to integrate or merge multiple models could be a possible solution.

With the limited computing resources, to parallel and distribute the A.I. computing missions into multiple computing cores is a possible method [1]. Although the processing unit cannot provide huge computing resources and performance, by dividing the original computing job into multiple tasks, the small models could still deal with the A.I. computing works. Some systems would deploy the specific A.I. mission or computing service on the individual model. Hence, when the computing job is combined with various A.I. services required, each request could be delivered to the corresponding model.

With the development of A.I. services, to let the A.I. application draw an image even the melody with some text descriptions are popular [2-5]. Service users could provide the text description of some objects or emotions, the A.I. services could provide the possible and various images or music according to the data in the database and the past learning histories. Since the training data for the A.I. services are collected from various users and database, it means that one description would cause various feasible solutions or results.

On the other hand, based on the object recognition from the image processing, to transfer or translate from an image to the text description would also cause the various explanations or descriptions. With the importance, recognition accuracy, the size of the segmentation, etc., of the objects in the image, the description or explanation of the specific image would be more different. Therefore, how to let the A.I. service explains the image more similar to people with the order of importance and the topic of the story would be an issue.

In this research, we propose the recursive object segmentation and recognition procedure, called Multiple Merged Models, for the image description and explanation. In section 2, the related researches and works are introduced. Section 3 presents the proposed system. The verification results are given in section 4. Finally, the section 5 provides the conclusion.

II. Related Works

Recently, the A.I. services could generate the image according to the given sentence or description which is divided into multiple keywords. One keyword would be used to generate the multiple possible images or objects. Finally, by merging these possible images various images corresponding to the description are presented.

To draw an object corresponding to the text description, the features of the specific object should be defined in advance. Traditionally, different object recognitions could be done depending on huge object data training. By labelling the various objects with huge image data, the object recognition and segmentation could be achieved [6].

However, to execute the A.I. services or image recognitions, especially for the data training, huge computing resources are needed. Except the centralized computing resource for the A.I. services, the parallel processing and distributed system based on the modular computing jobs was proposed [1]. Originally, the requested computing job of image recognition or A.I. service would contain multiple purposes or the recognition of various objects. For the parallel processing, the requested computing job could be divided into multiple tasks corresponding to multiple purposes or objects. Each task would be only for a single recognition target or computing process. Finally, the tasks are distributed to the multiple Computing nodes for further processing. Therefore, we can use the distributed system based on multiple computing nodes with less hardware cost instead of centralized A.I. computing.

In addition to the parallel processing and distributed system for computing, to properly describe the recognized objects, even to give the description corresponding to the original images are also the research issues. The first step is to recognize objects included in the images. By using YOLO algorithm and implemented service, users can give the images related to the target objects for training. After training, the YOLO service could identify and recognize the specific target object included in the image. For example, according to the music spectrum or characteristics, the similar music instrument can be detected [3]. In addition, by collecting the image of the target object for YOLO algorithm training without anchor points, the target object can be recognized even under the water with high turbidity [4]. Even when the target object is less different in color from the background, the YOLO algorithm could still recognize the fig fruit [5].

After recognition, the found objects should be shown to the users. Based on the YOLO recognition results, the identified objects can be estimated and feedback to the users through audio [2]. In other words, the objects can be represented as the text plain content. It means that to give the description of the image is possible after object recognition.

III.Proposed System

In this research, suppose that an object can be represented as a data structure. Various attributes included in the data structure are used to describe the corresponding object. Some attributes in the current object data structure can be pointed to another independent data structure as another new object with other new attributes. For example, a person can be an object with various attributes such as clothes, glasses, height, etc. An attribute, take clothes as an example, could be also an object with addition attributes which include color, material, sleeve length, etc. Therefore, similar to the tree structure, an object in the root node could be described and defined by more extending structures or data nodes. The following figure shows the possible data tree based on the data structure.

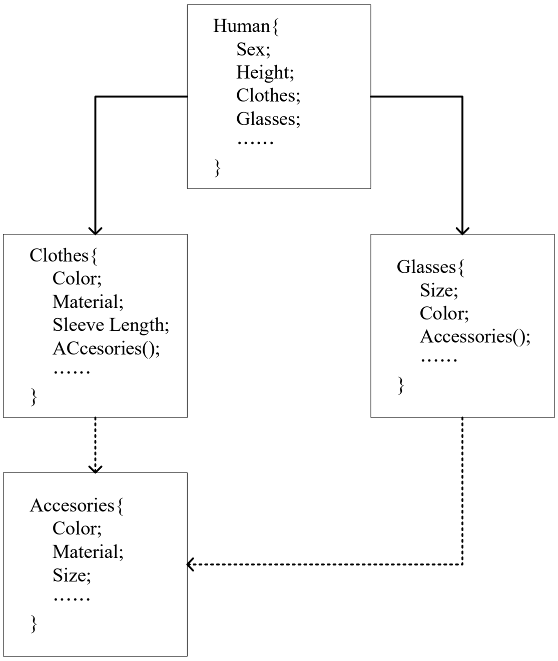


Figure 1. The UML diagram for the relationship between different classes of objects

In this research, some classes of data structure are defined as the root structures or root classes for the root object implementation. When a root object is recognized and segmented, the possible extending object according to the attributes in the parent class would be verified within the parent object segmentation area similar to the zoom in management of the image. Since the object would be only recognized and found within the limited area, the relationship between the parent object and the extending object can be stronger with higher accuracy.

Figure 2 presents the process flowchart. In the beginning, the image file would be reviewed and recognized according to the pre-defined root classes. If there are objects recognized that matching the definition of root classes, these objects are called root objects. In this research, an independent data structure corresponding to the recognized object is implemented. Therefore, one main object would be followed by multiple sub-objects. It also means that the data structure corresponding to an object would be pointed and chained extendedly by other structures when the attributes point to the structure of the sub-objects.

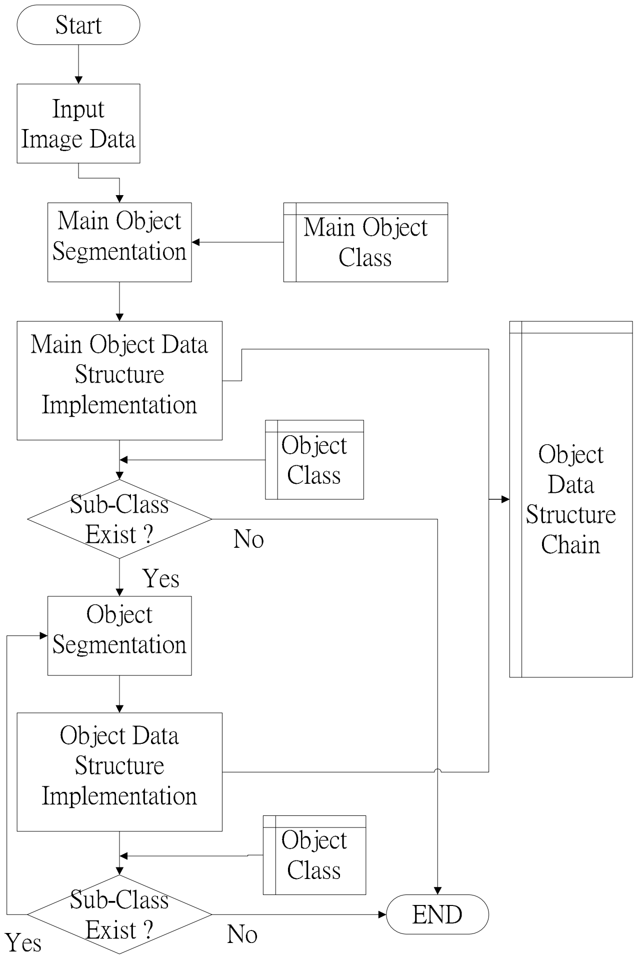


Figure 2. The flowchart of the object recognition and segmentation for the object data structure chain

In other words, according to the attributes included the data structure, the maximum amount of the objects recognized and segmented in the image can be decided.

Therefore, according to the Figure 2, all the found objects would be represented as multiple object data structure chains. Each main object would be extended as the data tree structure or multiple branch chains. Generally speaking, most recognized objects could be presented or expressed as a noun individually. In addition, the attributes of the structure corresponding to individual object are various adjectives such as color, size, etc. Considering the possible verbs related to the specific noun of the object, a suitable description of the current object can be developed.

IV. Verification Results

In this research, the image recognition is implemented based on YOLO v8. When a new image comes, the central of the image will be decided first. Generally speaking, the object in the central location would be the main object or motif of the image. After first round of YOLO, the main object with the biggest size or the nearest position of the image centre can be recognized. Hence, the procedure in Figure 2 can be used for the data information chain development. For example, in the Figure 3, there are 3 persons are recognized with the bigger segment size. Since these 3 persons can be implemented as 3 individual objects according to the same data structure/class called ‘person’, it means that these three objects would have the same importance. Then, according to the data structure of person, there are multiple attributes which may also point to another structure. In this research, shoes, pants, emotions, etc., are the possible attributes which would need the on demand defined data structure for description.

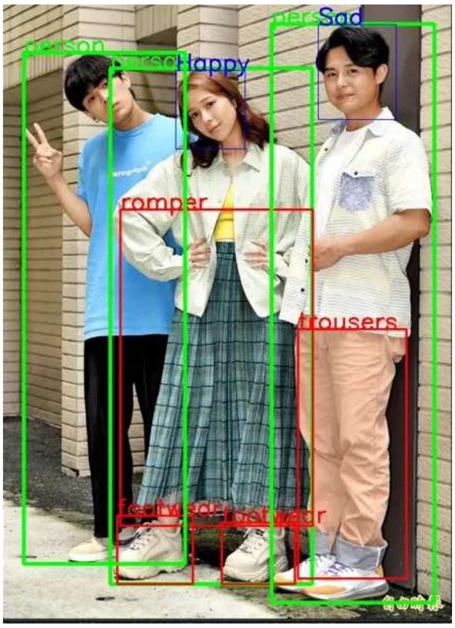


Figure 3. The example of the object recognition according to the proposed system

Then, according to the on demand defined structure corresponding to the results of image recognition, the various recognized objects related to the main object can be found and chained. Table 1 shows the data structure chain corresponding to the individual main object.

TABLE 1. DATA STRUCTURE CHAIN CORRESPONDING TO INDIVIDUAL FOUND MAIN OBJECT

No.	Data Structure Chain	
	Main Object	**Sub-Class / Object**
1	person	[footwear, white] [footwear, white] [romper, black] [Happy]
2	person	
3	person	[trouser, pink] [Sad]

Hence, by merging multiple structures in the individual data structure chain corresponding to the main object, the suitable description can be given. For example, the object number 1 can be described as “in this image, a person who wears white shoes and black romper is happy”.

V. CONCLUSIONS

Based on the proposed merging data structure chain, multiple objects corresponding to the main object can be represented as the normal natural language. In addition, the structures of recognized objects can be correctly connected to the main object structure. In other words, based on the data chain or even the tree structure, more suitable text plain content can be given.

REFERENCES

[1] M. -S. Jian, J. –T. Gui, X. –M. Wu, “Uncentralized Artificial Intelligence Computing Agent with the Distributed Training and Computing Tasks Based on Open Source Cloud Proxy,” CSCE 2023 - ICAI, Las Vegas, USA, p.93. (ISBN:1-60132-518-5)

[2] M. Mahendru and S. K. Dubey, "Real Time Object Detection with Audio Feedback using Yolo vs. Yolo_v3," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 734-740, doi: 10.1109/Confluence51648.2021.9377064.

[3] C. Dewi, R. -C. Chen, Hendry and Y. -T. Liu, "Similar Music Instrument Detection via Deep Convolution YOLO-Generative Adversarial Network," 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), Morioka, Japan, 2019, pp. 1-6, doi: 10.1109/ICAwST.2019.8923404.

[4] X. Wang, X. Jiang, Z. Xia and X. Feng, "Underwater Object Detection Based on Enhanced YOLO," 2022 International Conference on Image Processing and Media Computing (ICIPMC), Xi'an, China, 2022, pp. 17-21, doi: 10.1109/ICIPMC55686.2022.00012.

[5] W. Yijing, Y. Yi, W. Xue-fen, C. Jian and L. Xinyun, "Fig Fruit Recognition Method Based on YOLO v4 Deep Learning," 2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Mai, Thailand, 2021, pp. 303-306, doi: 10.1109/ECTI-CON51831.2021.9454904.

[6] A. Torralba, B. C. Russell and J. Yuen, "LabelMe: Online Image Annotation and Applications," in Proceedings of the IEEE, vol. 98, no. 8, pp. 1467-1484, Aug. 2010, doi: 10.1109/JPROC.2010.2050290.

Hui-Yu Huang received the B.S. degree in Electronic Engineering from Feng Chia University, Taiwan, in 1992, and the M. S. degree in Electrical and Computer Engineering from Yuan Ze Institute of Technology, and the Ph.D degree in Electrical Engineering from National Tsing Hua University, Taiwan, in 1994 and 2002, respectively. Since 2006, she has been with the Department of Computer Science and Information Engineering at National Formosa University in Taiwan, where she is now a professor. Her research interests include computer vision, multimedia security, fuzzy and neural network applications, and content-based image retrieval system. Dr. Huang is a member of the Chinese Association of Image Processing and Pattern Recognition

Ming-Hsun Tsai currently is a master degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to object recognition and segmentation, and image processing. He joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2023.

Ming-Shen Jian was born in Kaohsiung City, Taiwan in 1978. He received the B.S. from the National Chiao Tung University, HsinChu, and Ph.D degrees in Computer Science and Engineering from the National Sun Yat-sen University, Kaohsiung, Taiwan in 2007. From 2018, he was an Associate Professor and director with the National Formosa University Cloud Computing and Intelligent System Laboratory. Currently he is also an IEEE Senior Member. Since 2009, he has been an Assistant Professor with the Computer Science and Information Engineering Department, National Formosa University. He is the author of four books, more than 50 articles, and at least 15 invention patents. His research interests include IOT development and application, Big Data, Optimal Solution, Intelligent System, and Cloud Computing. He was a Secretary of the Taiwan Association of Cloud Computing. Dr. Jian was a recipient of the IEEE sponsored international conference Paper Award in 2016, 2017, and 2018.

Pipeline Based Genetic Algorithm for Patient Scheduling in Hospital Outpatient Department and Laboratory

Ming-Shen Jian*, Cheng-He Wang, Wei-Siou Wu, Tzu-Wei Huang

Dept. of CSIE, National Formosa University, Yunlin County, Taiwan 632

jianms@nfu.edu.tw*, 40843136@gm.nfu.edu.tw, 41043117@gm.nfu.edu.tw, 41043223@gm.nfu.edu.tw

Abstract— **This research propose the pipeline based genetic algorithm for patient scheduling. The time spent in the hospital outpatient department, laboratory, and the idle time caused by the scheduling are considered. According to the subject condition, the optimal solution related to medical service time spent for the patients in different medical departments with single or multiple medical services can be found. Through the cloud computing platform, the developed algorithm can be rapidly used for different requirements from hospitals or patients.**

Keywords— **Genetic Algorithm, Scheduling, Medical Service, Pipeline, Cloud Computing**

I. Introduction

Recent years, due to infectious diseases, such as COVID 19, the workload of the doctors, employees, nurses, etc., in the hospital has increased. Due to the time spent of the inspection corresponding to individual patient would be different. In addition, the human resource in the hospital would also be tight. To maintain the quality of the health and medical care, a suitable work schedule for doctors, nurses, outpatient clinic or operating room, etc., is required.

There were various scheduling algorithms or methods proposed for the nurse scheduling [1-3]. According to the restrictions of the law and the working hours, the total number of nurses required for the hospitalized patients can be decided. However, the scheduling work corresponding to the nursing only focuses on the ward of the hospital and nurses. There are still doctors, inspectors, medical device operators, pharmacists, etc., in the hospital. In addition to the human resources, the physical spaces such as clinic or outpatient department, surgery room, X-ray laboratory, MRI laboratory, etc., are also limited resources. Considering the restrictions of the human resources, physical spaces, time, the various states of the illness corresponding to different patients, etc., how to provide the optimal medical service for patients could be an important issue.

To find the optimal solution, various algorithms were proposed [4-6]. Ant Colony Optimization (ACO) was proposed that emulates the behaviour of the ants finding optimal routes for food. Hence, when given the objective problem, the ACO algorithm could find the optimal solution [4]. Genetic Algorithm (GA) is another popular method to find the optimal solution. Based on the concept of "survival of the fittest", the optimal solution can be found after evolution of many generations 5[]. Simulated Annealing (SA) is a kind of randomized algorithm and proposed to find the near optimal solution within a huge searching space [7]. Since the normal nurse scheduling problem (NSP) was proved as the NP-hard problem, to find the optimal solution instead of the manual operation is the acceptable method.

However, most algorithms for the NP-hard problems would require huge computing resources. Therefore, to achieve the good performance of algorithm computing with enough resources, cloud computing could be the possible solution. Some researches were proposed to find the optimal solution based on the genetic algorithm including Nurse Scheduling Problem (NSP) [1-3,6-7] and Job Shop Scheduling problem (JSP) [4-5,8]. Based on the parallel processing and distributed system, the speed of convergence can be faster than the traditional genetic algorithm executed on the single computer.

Since there exists multiple objective functions corresponding to the human resource, physical medical service places, and patients, the conflicting goals may happen between various optimal solution searching. Therefore, how to deal with the conflict of multiple objects optimal solution searching is an important issue.

In this research, section 2 will introduce the proposed genetic algorithm for the NSP and JSP optimal solution searching. The cloud for rapidly deploying the algorithms is also described. Then, section 3 provides the proposed system for module procedures of genetic algorithm. In addition, the adaptive process for multiple objects optimal solution searching corresponding to the medical service and scheduling in hospital is given. section 4 shows the simulation results. The conclusion is given in section 5.

II. Related Works

The Nurse Scheduling Problem (NSP) and Job Shop Scheduling problem (JSP) were already proposed. Mostly, the optimal solution is searched according to the time spent on the job done by all machines or procedures [5] or the minimized human resources of the nurse for schedule [6]. Suppose that the human resources can satisfy the requirements of the schedule, the research would focus on the quality of the schedule for more fair assignments [7].

Instead of the single objective function being considered for the optimal solution searching, some researchers developed the algorithms for the multiple objectives optimal solution finding [5]. To find the optimal solution within the huge solution space, some researchers proposed the modularized algorithms and procedures for the multiple objectives optimal solution searching [5-6,8]. Then, by rapidly implementing the on demand defined modules, the algorithm can try to find the optimal solution according to various configurations or methods. In other words, based on the cloud computing which owns huge computing resource, the various combinations of the algorithms or methods for the optimal solution searching can be achieved.

In addition, the computing job can be divided into multiple sub-partitions. When an optimal solution searching is required and triggered, the computing data can be separated into the multiple missions of the original job. These missions could be delivered to the computing nodes of the cloud computing environment. In other words, if the data or the mission will be managed and computed by various algorithms or methods, cloud computing environment can execute data or mission corresponding to the algorithm or method individually. Several researches assume that the procedure or process dealing with the data or computing missions can be on demand deployed in the virtual machines of the cloud. Recently, the cloud services based on the container were popular and generally used. Through the same Docker engine, the container with the service program can be deployed directly. Hence, when the services are rapidly and repeatedly required, the on demand configured Docker containers can be built immediately.

Considering the genetic algorithm for the optimal solution searching, there exist several steps of data computing. Excluding the initialization of the chromosomes which are used to represent the possible solutions or sequences, there are steps about selection, crossover, mutation, and survival, which various possible methods were proposed. To select the chromosomes for crossover, there are methods about random selection, roulette wheel selection, elitism, tournament, etc., with different advantages. The crossover method for the gene exchanging and offspring chromosome creating are generally used according to single-point crossover, double-point crossover, multi-point crossover, and uniform crossover. To possibly escape from the local search, the mutation method can be used based on simple mutation or uniform mutation. Finally, when the chromosomes compete each other for survive, the fitness value sort or other conditions for survive can be used. Therefore, various combinations related to different methods of these steps would contribute different performance corresponding to different objective functions for optimal solution searching.

Since the genetic algorithm has multiple steps with various methods, to deploy these methods individually would be the possible solution for various combinations. It means that these methods should be modularized for the computing data exchanging between each other in different steps. To enhance the feasibility of the "ready to use" methods, the container based on Docker engine is the solution. In this research, the genetic algorithm is modularized and divided into multiple containers. The user could select the required methods corresponding to the steps of genetic algorithm for different combinations.

III. Proposed System

Traditionally, the scheduling in the hospital would only focus on the nurse scheduling problem with the fixed time limitation and optimal solution with the balance about human resource and fair schedule. However, due to the time spending of various medical outpatient departments or the medical inspections may be different, the objective function to find the optimal solution of scheduling can be defined as two purposes: 1. shortest time for completing the patients' medical care, or 2. reduce the idle time of all medical outpatient departments and inspections.

In hospital, the specific medical inspection would be needed for multiple medical outpatient departments. For example, thoracic surgery department and orthopaedics department would both require the inspection report from the X-ray laboratory. However, various patients from different departments would cost various time spent in X-ray laboratory. In other words, patients from different medical outpatient departments would affect each other and cause the possible time delay or idle time in waiting the inspection of X-ray. Therefore, more factors such as the priority, quantity of patients, capacity of the laboratory, etc., should be taken into consideration.

Suppose that there are total *K* medical outpatient departments, where the set of these departments can be defined as $\{d_1, d_2, \ldots, d_k\}$. The total patient corresponding to the k^th medical outpatient department can be defined as

$$P(d_k) = \sum_{i=1}^{m} p_i(d_k) \tag{1}$$

where $p_i(d_k) = \begin{cases} 1 & i^{th} \text{ patient is a patitent of } k^{th} \text{ department} \\ 0 & i^{th} \text{ patient is not a patitent of } k^{th} \text{ department} \end{cases}$

and m is the total patients of the hospital.

Considering the real situation of the medical outpatient department, all the patients should seek treatment or doctor in order. Therefore, the patient list in order corresponding to the specific k^th outpatient department can be defined as:

$$P_{Lk}(d_k) = \bigcup_{i=1}^{m} p_i(d_k) \tag{2}$$

Suppose that the i^th patient in the k^th outpatient department would spend time $t_{i,k}$ with the probability $l_{i,k}$ of the further inspection at a laboratory. Assume that the normal time spent in the j^th laboratory corresponding to a patient from the k^th outpatient department is $t_j(k)$. If there exists the time delay or idle time for waiting the j^th laboratory or the medical outpatient department corresponding to the i^th patient, called D_{ij}, then each patient for the complete medical care about the outpatient department and laboratory can be defined as:

$$T_i = \sum_j \sum_{k=1}^{K} [(t_j(k) + D_{ij}) \times l_{i,k} + t_{i,k} \times p_i(d_k)] \tag{3}$$

Therefore, according to eq.3, we can have the first objective function about the total time for an individual patient can be as:

$$\min T_i = \sum_j \sum_{k=1}^{K} [(t_j(k) + D_{ij}) \times l_{i,k} + t_{i,k} \times p_i(d_k)] \tag{4}$$

Since the normal time spent in the jth laboratory is already defined in advance, finding the minimized time according to eq.4 also means that the delay time or idle time should be minimized. In other words, to find the optimal solution for the patient's schedule in the hospital is very important.

However, various patients from different outpatient department would need the inspection report from the same laboratory. To reduce the idle time or delay time in laboratory, considering the normal time spent in the jth laboratory corresponding to a patient from the kth outpatient department and the time delay or idle time for waiting at jth laboratory is important. Hence, the second objective function for the jth laboratory can be defined as:

$$\min \mathrm{T_j} = \sum_{\mathrm{i}} \sum_{\mathrm{k=1}}^{\mathrm{K}} [(\mathrm{t_j(k)} + \mathrm{D_{ij}}) \times l_{\mathrm{i,k}} + \mathrm{t_{i,k}} \times \mathrm{p_i(d_k)}] \quad (5)$$

However, due to the immediate outpatient needs, some patients after the jth laboratory should return to the outpatient department for the immediate inspection report with the probability $\mathrm{R_{j,k}}$. Therefore, even the total time for the ith patient at the kth outpatient department can be defined in advance, the subject condition to the laboratory returned patient should be:

Medical Service for a returned patient should follow the order: 1. outpatient department, 2. Laboratory, and 3. outpatient department.

In other words, the objective function for individual patient defined in eq. 4 should be as:

$$\min \mathrm{T_i} = \sum_{\mathrm{j}} \sum_{\mathrm{k=1}}^{\mathrm{K}} \begin{bmatrix} (\mathrm{t_j(k)} + \mathrm{D_{ij}}) \times l_{\mathrm{i,k}} + \mathrm{t_{i,k}} \times \mathrm{p_i(d_k)} \\ +(\mathrm{t_{i,k}} + \mathrm{D_{ij}}) \times \mathrm{p_i(d_k)} \times R_{\mathrm{j,k}} \end{bmatrix} \quad (6)$$

In this research, the patient is the first important thing for scheduling. Therefore, instead of the nurse scheduling, we would focus on the minimum time of the patient in the hospital for all the required medical services.

However, the physical spaces for the outpatient department are also limited as the subject condition. The law also limits the total work time of the doctor. Therefore, another subject condition should be given as:

Upper bound of the k^{th} outpatient department is T_k

Therefore, the probability of the further medical service such as $l_{\mathrm{i,k}}$ and $\mathrm{R_{j,k}}$ would change the possibility of optimal solution searching. If more patients require the second even third step of the medical service, due to the subject condition, the patient may not complete the medical service in time. Therefore, the objective function for the kth outpatient department should be:

$$\mathrm{T_k} \geq \sum_{\mathrm{i}} \min \mathrm{T_i}$$

$$= \sum_{\mathrm{i}} \sum_{\mathrm{j}} \sum_{\mathrm{k=1}}^{\mathrm{K}} \begin{bmatrix} (\mathrm{t_j(k)} + \mathrm{D_{ij}}) \times l_{\mathrm{i,k}} + \mathrm{t_{i,k}} \times \mathrm{p_i(d_k)} \\ +(\mathrm{t_{i,k}} + \mathrm{D_{ij}}) \times \mathrm{p_i(d_k)} \times R_{\mathrm{j,k}} \end{bmatrix} \quad (7)$$

Since the patient of the outpatient department may need the further inspection at another laboratory, the next patient in the outpatient department could be served in the current department. Therefore, the medical service in this research is defined as the pipeline. Each outpatient department is one individual work section included in the pipeline. Hence, the pipeline sections of ith patient who is a returned patient from the jth laboratory in kth outpatient department can be shown as follows:

k	j	k

If the patient needn't to have the inspection report or not a returned patient, then the pipeline sections can be shown as

k	k	null

If the patient needn't the laboratory, the pipeline sections can be shown as

k	null	null

Therefore, in the kth outpatient department, the possible pipeline for multiple patients can be shown as Figure 1. Suppose that the time $\mathrm{t_j(k)}$ is longer than the time $\mathrm{t_{i,k}}$. Therefore, the section k which represents the time $\mathrm{t_{i,k}}$ of the ith patient in kth outpatient department is shorter or smaller than the section j which indicates the $\mathrm{t_j(k)}$ in jth laboratory from kth outpatient department. Hence, the first patient who is the returned patient should be idle when the third step of the medical service, kth outpatient department, is also occupied by the third patient. When the second patient requires the service of the jth laboratory, additional idle time is needed for waiting the served first patient. Suppose that the third patient needn't to have the third step of the medical service. Then, the total time spent for these three patients would be started from the first patient's first kth outpatient department service to the ended jth laboratory of the third patients. Since the time spent in various laboratories and departments would be different, the additional idle time corresponding to individual patient is also various.

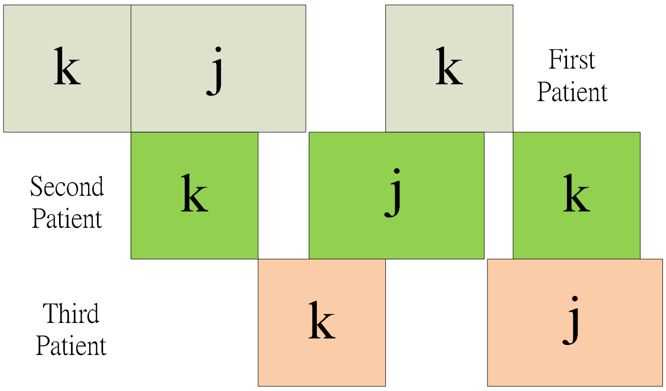


Figure 1. Example of the pipeline corresponding to kth outpatient department and jth laboratory

In this research, we have two coding methods to represent the pipeline into the genetic algorithm chromosome. First, the longest section is used as the time unit for all other sections. Then, all the time sections corresponding to the outpatient departments and laboratories can be the same. Second, crossover can only exchange the whole sections of individual patients. Therefore, only the sort or order of patients would be changed. The required medical service corresponding to each patient is still kept.

IV. Verifications

In this research, we design the scenario which includes 2 independent outpatient departments similar to the "thoracic surgery" and "orthopaedics". These 2 outpatient departments may require the service of X-ray laboratory. Suppose that the probability of a patient about thoracic surgery and orthopaedics for the X-ray laboratory service could be α_{ts} and α_o. In addition, the probabilities of a patient who would be required to back to the related outpatient department after X-ray laboratory are defined as $R_{j,k}$. Then, according to the probability, the possible patients can be divided into 6 types of groups: thoracic surgery for only outpatient, thoracic surgery for both outpatient and X-ray laboratory, thoracic surgery for 2 times outpatient and single time X-ray laboratory, orthopaedics for only outpatient, orthopaedics for both outpatient and X-ray laboratory, and orthopaedics for 2 times outpatient and single time X-ray laboratory.

To process the algorithm, we implement the cloud computing environment based on the Intel CPU Xeon with 16GB memory. The Docker engine v23.0.5 for Docker Desktop application is used.

Suppose that the whole time period for the outpatient department is limited. Then, the probability of α_{ts}, α_o, and $R_{j,k}$ would affect the results. If probability value is high with huge total amount of patients, the optimal solution for all patients having medical service is unavailable due to too much time spending.

In this verification, we assume that the probability $l_{i,k}$ of the patients who require the laboratory checking is defined as: 0.4, 0.6, and 0.8, individually. The probability $R_{j,k}$ of the patients after laboratory is defined as: 0.4, 0.6, and 0.8, individually.

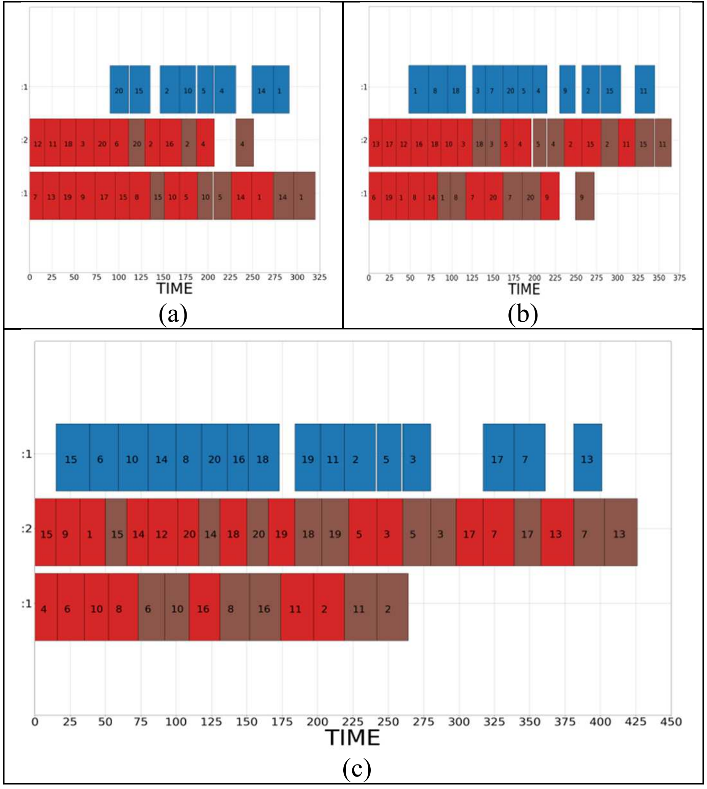


Figure 2. Example of the 20 patients with different probability values corresponding to 2 outpatient and 1 shared laboratory.

According to the figure 2, assume that room 1 is for laboratory, room 2 and 3 are the independent outpatient rooms for different medical clinic. There are total 20 patients. Each patient is assigned the unique ID number. According to the subject conditions, each patient should attend the outpatient first. Then, based on the probability, the patient would have further steps for laboratory or second outpatient. If the patient is first time in the outpatient room, the color in the figure 2 would be red. The laboratory time of the patient is shown in color blue. The time for outpatient and laboratory is randomly decided in the range from 15 to 25. In figure 2 (a), probability $l_{i,k}$ and $R_{j,k}$ are both set as 0.4. Figure 2 (b) shows the verification result with probability $l_{i,k}$ and $R_{j,k}$ which are both set as 0.6. Finally, in figure 2 (c), the probability $l_{i,k}$ and $R_{j,k}$ are both set as 0.8. The verification shows that the proposed system can successfully arrange the schedule of each patient. The progress time of patients, the independent outpatient and laboratory rooms, can be also optimized with the shortest total time spent.

V. Conclusions

The pipeline based genetic algorithm is proposed for dealing with the various medical services required by different patients. By considering the objective function, the minimized time spent for the medical outpatient department can be found. The optimal order or schedule for the patients corresponding to the medical outpatient can be provided.

Acknowledgment

Thanks to Cloud Computing and Intelligent System Lab. In National Formosa University

References

[1] A. Osman, N. Al-Hinai and S. Piya, "Development of Automated Schedule Generator for Nurses in Emergency Department," 2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), Manama, Bahrain, 2019, pp. 1-3, doi: 10.1109/ICMSAO.2019.8880428.

[2] R. Refat, A. Taha and S. Senbel, "A Heuristic Quality-based Nurse Scheduling Algorithm for Emergency Centers," 2014 24th International Conference on Computer Theory and Applications (ICCTA), Alexandria, Egypt, 2014, pp. 50-56, doi: 10.1109/ICCTA35431.2014.9521608.

[3] M. K. Khan, H. Takeuchi and M. Ohki, "An Experimental Application of Multi-Objective Evolutionary Algorithm to Many-Objective Nurse Scheduling for Real General Hospitals," 2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Jeju, Korea (South), 2021, pp. 1-4, doi: 10.1109/ITC-CSCC52171.2021.9501456.

[4] M. huang, D. guo, X. liang and X. liang, "An Improved Ant Colony Algorithm is Proposed to Solve the Single Objective Flexible Job-shop Scheduling Problem," 2020 IEEE 8th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, 2020, pp. 16-21, doi: 10.1109/ICCSNT50940.2020.9305005.

[5] M. -S. Jian, W. -S. Wu, Y. -H. Lin, X. -M. Wu and P. -W. Wang, "Multi-Objective Optimization Based on Adaptive Evolution Algorithm for Disassembly Line Network Problems," 2022 24th International Conference on Advanced Communication Technology (ICACT), PyeongChang Kwangwoon_Do, Korea, Republic of, 2022, pp. 524-533, doi: 10.23919/ICACT53585.2022.9728929.

[6] Ming-Shen Jian, Ming-Sian You, "Cloud Based Hybrid Evolution Algorithm for NP-Complete Pattern in Nurse Scheduling Problem,"

International Journal of Innovation, Management and Technology, Vol. 7, No. 5, pp.234-237, 2016

[7] L. Rosocha, S. Vernerova, R. Verner, “Medical staff scheduling using simulated annealing,” Quality Innovation Prosperity, 19(1), 2015. DOI: 10.12776/qip.v19i1.405

[8] Ming-Shen Jian, Yi-Chen Jhou, Ming-Sian You, “Modular Feedback Assistance Hybrid Evolution Algorithm Based on Cloud Environment for Job Shop Scheduling Problem Optimization,” *ICADIWT2015*, pp. 205-215, Feb. 2015

Ming-Shen Jian was born in Kaohsiung City, Taiwan in 1978. He received the B.S. from the National Chiao Tung University, HsinChu, and Ph.D degrees in Computer Science and Engineering from the National Sun Yat-sen University, Kaohsiung, Taiwan in 2007. From 2018, he was an Associate Professor and director with the National Formosa University Cloud Computing and Intelligent System Laboratory. Currently he is also an IEEE Senior Member. Since 2009, he has been an Assistant Professor with the Computer Science and Information Engineering Department, National Formosa University. He is the author of four books, more than 50 articles, and at least 15 invention patents. His research interests include IOT development and application, Big Data, Optimal Solution, Intelligent System, and Cloud Computing. He was a Secretary of the Taiwan Association of Cloud Computing. Dr. Jian was a recipient of the IEEE sponsored international conference Paper Award in 2016, 2017, and 2018

Cheng-He Wang currently is a master degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to algorithm and scheduling problem. He joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2023.

Wei-Siou Wu currently is a bachelor degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to algorithm, scheduling problem, and system programming. He joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2023

Tzu-Wei Hunag currently is a bachelor degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to algorithm, scheduling problem, and system programming. He joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2023

Session 3A: 6G, Mobile Communication 1

Chair: Prof. Francis C.M. Lau, Hong Kong Polytechnic University, China

1 Paper ID: 20240202, 168~ 173

Traffic Type Recognition in 6G Software-Defined Networking for Telepresence Services

Dr. Artem Volkov, Ms. Varvara Mineeva, Dr. Ammar Muthanna, Prof. Andrey Koucheryavy,

The Bonch-Bruevich Saint Petersburg State Universi. Russia

2 Paper ID: 20240130, 174~182

Microservice-Based Fog Testbed for 6G Applications

Ms. Ekaterina Kuzmina, Ms. Meriem Tefikova, Dr. Artem Volkov, Dr. Ammar Muthanna, Prof. Andrey Koucheryavy,

The Bonch-Bruevich Saint Petersburg State Universi. Russian Federation

3 Paper ID: 20240366, 183~186

Migration routing algorithm for microservice based Fog computing system

Ms. Ekaterina Kuzmina, Ms. Meriem Tefikova, Dr. Artem Volkov, Dr. Ammar Muthanna, Prof. Andrey Koucheryavy,

Department of Telecommunication Networks and Data . Russia

4 Paper ID: 20240241, 187~192

Dual-RIS Assisted 3D Positioning and Beamforming Design in ISAC System

Mr. Dejie Ma, Prof. Zhiquan Bai, Dr. Jinqiu Zhao, Mr. Hao Xu, Mr. Zeyu Liu, Dr. Di Zhou, Prof. Mingyan Jiang, Prof. KyungSup Kwak,

Shandong University. China

5 Paper ID: 20240375, 193~198

Deep Reinforcement Learning Based Beamforming in RIS-assisted MIMO System Under Hardware Loss

Mr. Yuan Sun, Mr. Zhiquan Bai, Ms. Jinqiu Zhao, Mr. Dejie Ma, Ms. Zhaoxia Xian, Mr. KyungSup Kwak,

Shandong University. China

Traffic Type Recognition in 6G Software-Defined Networking for Telepresence Services

Volkov Artem
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
artemanv.work@gmail.com

Varvara Mineeva
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
varvaramineyeva@gmail.com

Ammar Muthanna
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
muthanna.asa@sut.ru

Andrey Koucheryavy
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
akouch@mail.ru

***Abstract*— This paper deals with the problem of traffic typing and telepresence services, presents the results of analysis of existing methods based on DiffServ mechanisms such as Behavior Aggregate, Interface-based, MultiField. An extended traffic typing method based on LSTM networks is presented, a neural network for traffic recognition service in 6G networks is developed, promising directions such as the concept of 2030 networks and telepresence services are discussed, software-defined networking and virtualization of network functions are investigated. In this study, data obtained from an SDN flow table containing information about network traffic characteristics were used to train the ANN. To evaluate the effectiveness of the extended method, a special stand was developed to test and evaluate the quality of traffic typing. The stand includes the necessary hardware and software for conducting experiments and collecting data.**

***Keywords*— 6G networks, traffic typing, telepresence services, SDN multicontroller, software-defined networking, artificial neural network.**

I. Introduction

The Telepresence services are becoming one of the most important services of 6G networks [1,2]. It is worth noting that modern developments already allow the realization of such service models as "telepresence suits". In this regard, an important task of software-defined networking is to manage heterogeneous traffic transmitted over the network. Effective management requires classification of different types of traffic such as voice, video, data, etc. However, with the emergence of new traffic types, including telepresence services, haptic Internet traffic or M2M traffic, there is a need to develop and adapt classification methods for new data types [3,4].

Future networks should be able to provide high data density, more realistic communication and security with robust systems. It should also be noted that as the number of connected devices increases, the amount of data production will significantly increase., so the networks of 2030 will need to handle huge amounts of data. One prominent example is the model of humanoid robots, which are the direct nodes of the network and implement network functions such as data transmission and processing.[5]

Therefore, traffic typing is necessary to ensure guaranteed throughput and congestion management in software-defined networking. This allows for a better user experience and improved data transfer. Typing techniques must be efficient and flexible to handle different types of traffic and adapt to changing network conditions.

Telepresence services are essential for next-generation networks to provide low latency and high data reliability. This is crucial for applications such as autonomous transportation systems, industrial automation, and real-time medical diagnostics. Proper classification and typing of telepresence traffic is essential in ensuring quality of service in these networks.

II. RELATED WORKS

As part of the MEGANET LAB 6G activities, a network architecture model is defined. It encompasses several functional segments, including robotic networks, fog computing, and holography.

A detailed functional diagram is shown in Figure 1.

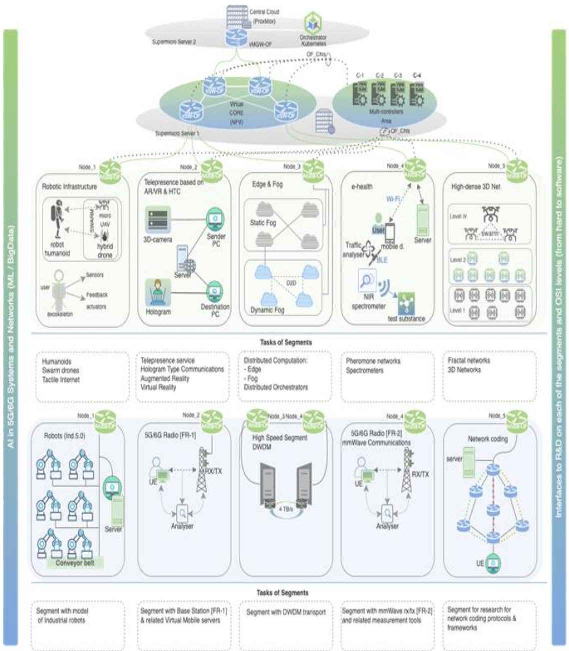


Fig. 1 - Architecture of the model network

According to ITU recommendations, this is a network of networks that provides seamless services regardless of the communication type, access technology, time, or user device. The network includes decentralized computing systems, integrated software-configurable environments represented by software-defined networking with virtual function virtualization, and a distributed software-controlled static and dynamic computing environment. This system operates as a self-evolving, self-regulating, and self-enhancing "living" intellectual entity.[6]

The 6G network will evolve from the 5G/IMT-2020 network, utilizing URLLC technology. Upgrading the 6G network will require point-to-point upgrades and advancements in all network layers, such as base stations, core networks, and end devices. It's crucial that these upgrades don't disrupt service availability, degrade quality, or alter service logic.[6]

Implementing telepresence services is a significant challenge. To manipulate objects in a remote environment, a telepresence system can control the device's movement. More complex alternatives, like reading hand and finger movements with special gloves or capturing body movements with a suit, are also possible. The device reproduces these actions, and the more accurate the movements, the stronger the immersion effect. Teleconferencing services provide new ways for users to interact over the network. Holographic communication has introduced innovative features. Numerous scientific studies, including ITU recommendations, have established a new form of communication known as holographic type communication (HTC).

To achieve this type of communication, the network requires easily configurable and maintainable facilities, which can be facilitated by SDN (Software-defined network) technology. Fig. 1 illustrates a model network that adopts a multi-controller network principle with a software-defined network. Refer to Fig. 2 for this segment.

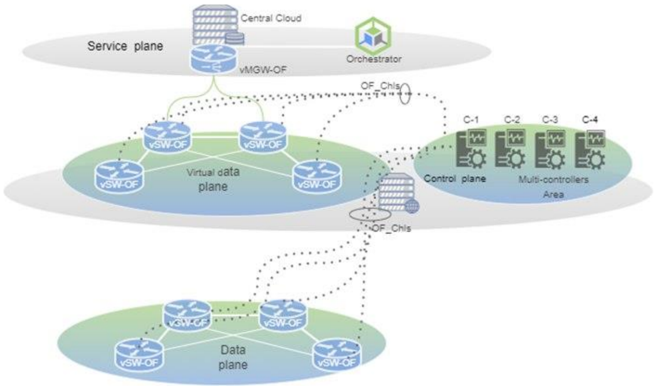


Fig. 2 - Structure of the stand with SDN multicontroller

Recognizing and classifying heterogeneous bulk traffic is a key challenge in designing 6G network traffic. This cannot be achieved using existing traditional methods. Instead, new robust artificial intelligence (AI) methods should be introduced. The algorithm developed to detect and recognize heterogeneous traffic in the underlying network serves as the basis for this approach, which is implemented at the control layer of the Software Defined Networking (SDN) network located in the core network. The algorithm is based on a neural network.[7]

III. PROBLEM STATEMENT

Software-defined networking (SDNs) centralize management and separate control functions from network devices, which allows for rapid and flexible network configuration and operation through software [8,9,10]. With the use of a centralized controller, network administrators can efficiently manage all network devices through software configuration and management, providing greater control and flexibility in network management.

The demand for telepresence services is on the rise. This type of traffic encompasses holographic replicas of individuals, robotic avatars, manipulation devices, and collaborative connections utilized within groups and communities. Telepresence services open up diverse opportunities for applications such as virtual reality, augmented reality, medical simulation, and beyond. High-speed data transfer, minimal delay, and consistent connectivity are essential for delivering an authentic and comprehensive user experience for this data type.

Various machine learning algorithms and methods for traffic typing in software-defined networking were studied in this paper. An investigation and extension of the traffic typing method using LSTM networks on a single SDN multicontroller responsible for network management was conducted. The ANN was trained using data from the SDN flow table containing network traffic information. To evaluate the effectiveness of the extended method, a dedicated stand was developed to test and assess traffic typing quality. The stand is equipped with necessary tools and software for conducting experiments and collecting relevant data.

The experiment produced data and ANN results for the traffic typing task. These results were evaluated based on the study's objectives, while also assessing the quality of traffic typing. Benefits and limitations of the machine learning technique were identified and the outcomes ultimately demonstrate the success of this approach for identifying promising traffic types.

IV. PROPOSED SYSTEM

Analysis of traffic typing methods

There are various methods to address traffic typing issues, and one such solution involves implementing DiffServ mechanisms. DiffServ, or Differentiated Services, is a Quality of Service (QoS) model that allows routers and switches to handle different types of traffic with varying priorities. By using the DiffServ approach, network traffic is classified into different classes based on their configurations, resulting in more efficient network transmission by offering different levels of service to various traffic types. DiffServ is a widely used technique for managing traffic in software-defined networking [11,12].

Based on this model, there are three methods for classifying network traffic:

Behavior Aggregate (BA) - This method utilizes a label in the packet header, such as the IP DSCP field.

Interface-Based - Traffic passing through a designated network interface is assigned to a particular traffic class.

MultiField (MF) is a packet analysis method that evaluates header fields, including IP addresses, ports, and MAC addresses, to attain a comprehensive understanding of network traffic and its behavior.

These traffic classification methods categorize traffic into different classes based on characteristics and labels. This enables network elements to effectively handle varied traffic classes in accordance with their priorities and requirements.

Several neural networks can address traffic typing problems, including recurrent neural networks (RNNs). RNNs analyze traffic with time dependencies and memorize data about past states to inform decisions based on current and previous features. LSTM (Long Short Term Memory) is a prevalent type of RNN that can memorize long-term dependencies in data. Recurrent neural networks with long short-term memory (LSTM) provide the best solution for the traffic typing problem. These networks demonstrate exceptional proficiency in handling time-dependent features. The advantages of using RNNs with LSTM include their ability to scrutinize the sequential structure of data, which is crucial for traffic typing, where packet arrival order holds significant importance. The LSTM cells within the RNN have the capacity to store information in long-term memory while also selecting what to transmit or forget, allowing for effective modeling of temporal dependencies and the capture of significant contextual features within the data. ITU Recommendation Y.3116 details this technique.

Therefore, using recurrent neural networks with LSTM for traffic classification results in high accuracy and robustness in classification. Additionally, it enhances adaptability and processing of intricate temporal data structures, making it the preferred option for this task.

V. PROPOSED ARCHITECTURE

To perform a comprehensive experiment, we developed a testbed architecture that includes the essential equipment and components for collecting and analyzing traffic data. This architecture enables the replication of realistic network conditions and experimentation with various types of traffic.

The booth architecture includes the following major components:

SDN Controller: This component is the centerpiece of the stand and is responsible for network management. The SDN controller is used to manage the switches and collect traffic data.

SDN Switch: There is an SDN switch present in the booth, which performs the function of forwarding traffic in the network. The switch is connected to and receives instructions from the SDN controller.

Data Collection Software: A special software has been used to collect traffic data which makes queries on the north interface of the SDN controller. This software can access the flow table of the SDN switch and extract the necessary information such as byte-count, packet-count and TimeStamp.

The components allow for creating a realistic environment and collecting traffic data within the SDN network. This data will be used to train the neural network and categorize traffic in the future. Refer to Figure 3 for the stand architecture.

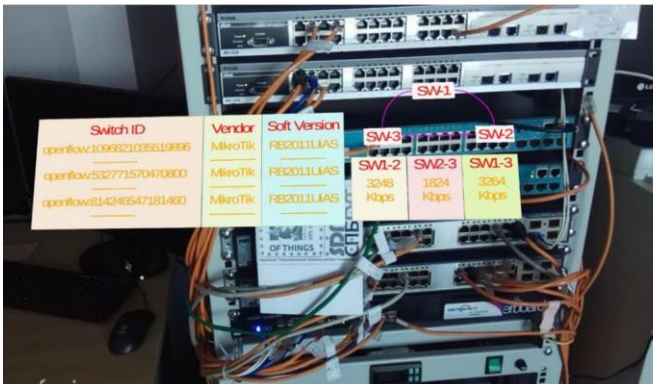


Fig 3 - Stand architecture

The revised version of the testbed architecture, shown in Figure 4, records data on all four traffic types used in this experiment. The equipment utilized for conducting the field experiment and gathering data is listed below:

Aruba Switch 2930F (JL259A) is an L3 switch that facilitates network traffic switching with support for the OpenFlow protocol, enabling its utilization in SDN-networks. The experiment employed the Aruba Switch 2930F (JL259A) to handle traffic and accumulate statistical data about SDN-network flows.

The Mikrotik CRS109-8G-1S-2HnD-IN is a versatile network device that combines router and switch functionalities, providing various options for routing and traffic switching on the network. It features Ethernet ports for device connectivity and a built-in wireless module for creating wireless networks. During the study, we used the Mikrotik Cloud Router Switch CRS109-8G-1S-2HnD-IN to establish a wireless network segment and connect devices via Wi-Fi.

TP-Link TL-MR100 is a wireless router that enables internet access through cellular networks. It supports high-speed internet access with 4G LTE standards, making it ideal for locations without a wired connection or requiring mobile communication. The router was used in the experiment to create an isolated network segment for devices to access the internet.

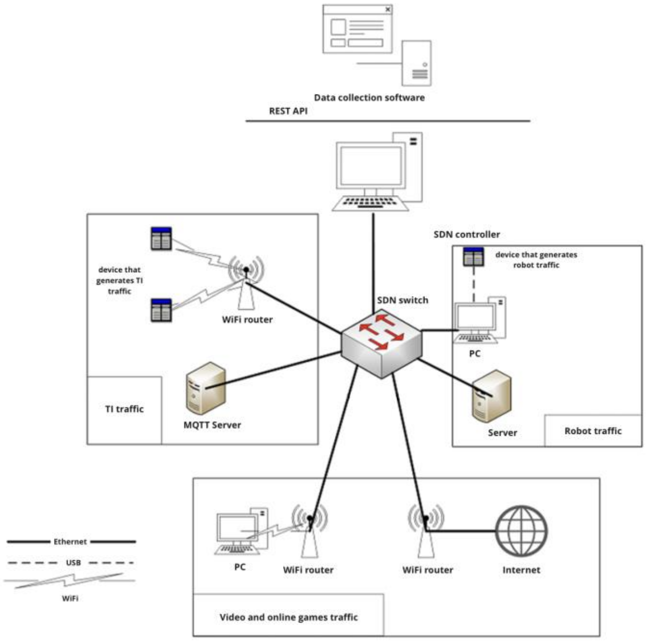


Fig. 4 - General stand architecture

The generators connected to the router through Wi-Fi. Two traffic generators gathered data by sending messages using the Message Queuing Telemetry Transport (MQTT) protocol to an MQTT server running on a personal computer. An SDN switch was vital in the network infrastructure by linking the router and the SDN controller.

VI. Experimental results

As part of our experiment, we carefully chose the size of the training batch (batch size). Increasing the batch size results in a significant improvement in learning speed, but it also leads to a decrease in accuracy. Similarly, increasing the batch size beyond 128 leads to a decrease in accuracy that falls short of 90%. Conversely, decreasing the batch size below 32 results in longer epoch times, taking more than 10 seconds to complete. To maintain a balanced outcome, we recommend a batch size between 32 and 128. Ultimately, we selected a training packet size of 128 since it achieves 95% accuracy and reduces training time by a factor of 2.6.

A neural network model with 256 neurons in the first hidden layer and 128 neurons in the second hidden layer was selected. These parameters resulted in high classification accuracy and sufficient training efficiency. Figure 5 depicts the neural network model for the traffic recognition service in the 6G network.

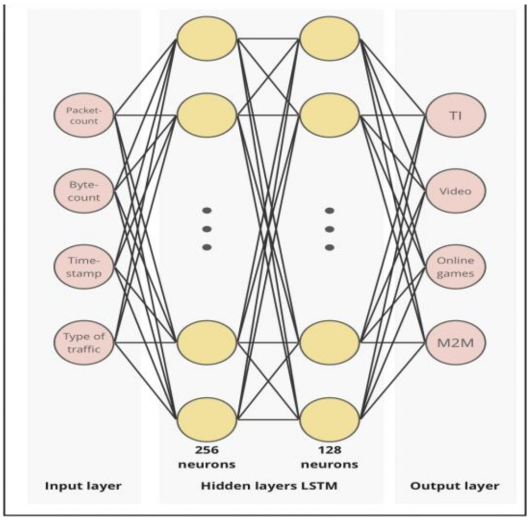


Fig 5 - ANN architecture

After training and testing the artificial neural network, we generated an error matrix, also referred to as a confusion matrix. This table helps us assess the model's performance by listing the number of examples that were correctly and incorrectly classified. Please refer to Figure 6 for the matrix.

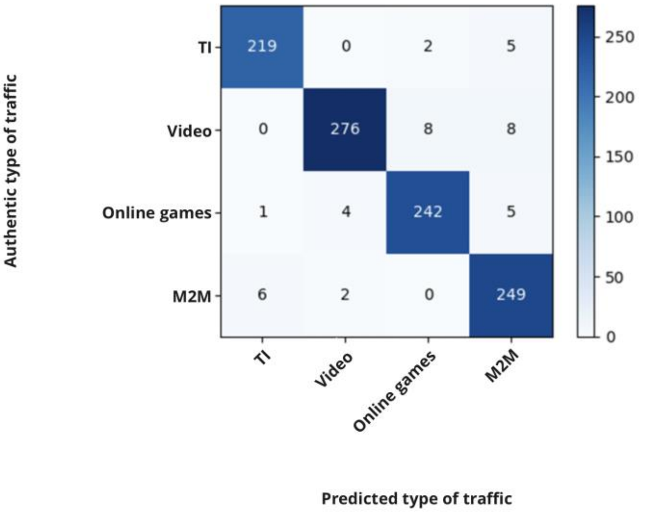


Fig. 6 - Error matrix

By analyzing the error matrix, we can draw conclusions regarding the model's performance:

The model accurately classifies haptic internet traffic and M2M traffic due to the values in the matrix cells being close to the diagonal.

The model struggles with categorizing both video and online game traffic due to non-zero values in the corresponding cells.

The total number of errors is small, indicating good performance of the model.

We examined the loss function, a metric that gauges the difference between the model's forecasted values and the actual values of the target variable. For an ANN, the loss function is its RMS error at every epoch. Figure 7 displays the loss function graph.

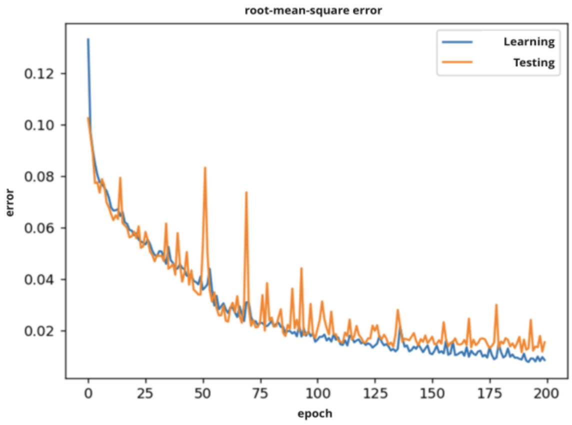


Fig. 7 - Loss graph

Analyzing the loss graph leads to the following conclusion:

The reduction in loss values with increasing epochs demonstrates the model's enhancements in its predictions and ability to typify traffic.

The final loss values are between 0.015 and 0.020, indicating effective model training.

After the ANN was trained, a diagram displaying the accurately identified traffic types and their corresponding probabilities was generated and presented in Figure 8.

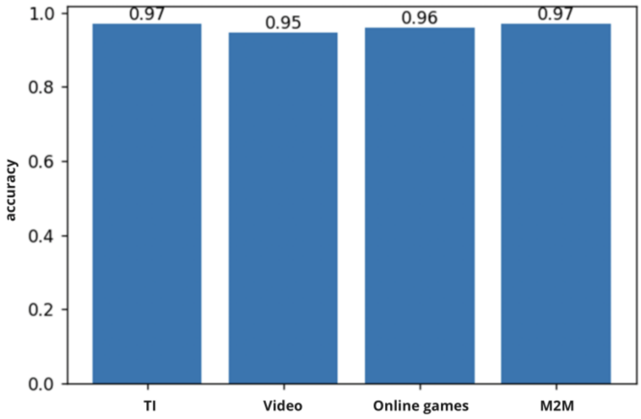


Fig. 8 - Probability of recognizing traffic types

The analysis indicates that the model accurately identifies and categorizes all four traffic types with high precision. Tactile Internet and M2M exhibit greater accuracy and fewer errors, whereas video traffic and online game traffic have slightly lower accuracy yet remain at a reasonably high level. Based on the column chart results and error matrix, it seems that the second and third traffic types share similar characteristics, causing the model to struggle when differentiating between them. This assumption is supported by the fact that the recognition accuracy for these two traffic types (0.96 and 0.92 respectively) are closer to each other compared to the other types. In the error matrix, the off-diagonal feature values (misclassified samples) for the second and third traffic types are more similar to each other than to the other traffic types. This may indicate that there are common features or similar characteristics in the data that cause difficulty in classification.

VII. CONCLUSION

This study is a part of a larger project - researching the extensive application of artificial intelligence for 6G networks, specifically focused on ultra-low latency and ultra-high density networking technologies. The study results show that the developed model attained high accuracy on the validation data, and during the testing process, the oscillation and RMS loss were minimized while accuracy was enhanced.

The model effectively detects haptic Internet and M2M traffic with high probability values. These outcomes indicate that the model achieved high accuracy and successful recognition of promising traffic types. The empirical method of training parameter selection also proved effective, confirming the significance of the model for analyzing and classifying network traffic.

VIII. REFERENCES

[1] Qiao, L.; Li, Y.; Chen, D.; Serikawa, S.; Guizani, M.; Lv, Z. A survey on 5G/6G, AI, and Robotics. Comput. Electr. Eng. 2021,95, 107372.

[2] Dogra, A.; Jha, R.K.; Jain, S. A survey on beyond 5G network with the advent of 6G: Architecture and emerging technologies.IEEE Access 2020, 9, 67512–67547.

[3] Muthanna, A., Ateya, A.A., Al Balushi, M. and Kirichek, R., 2018, May. D2D enabled communication system structure based on software defined networking for 5G network. In 2018 International Symposium on Consumer Technologies (ISCT) (pp. 41-44). IEEE.

[4] Duo, R.; Wu, C.; Yoshinaga, T.; Ji, Y. SDN-based handover approach in IEEE 802.11 p and LTE hybrid vehicular networks.In Proceedings of the 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scal-able Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smart-World/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Guangzhou, China, 8–12 October 2018; IEEE: Manhattan, NY, USA, 2018;pp. 1870–1875.

[5] FG-NET2030 – Focus Group on Technologies for Network 2030. Network 2030 Architecture Framework. – Geneva: ITU-T, 2020.

[6] Artem, V., Ateya, A.A., Koucheryavy, A., 2019. Novel AI-Based Scheme for Traffic Detection and Recognition in 5G Based Networks, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, pp. 243–255. doi:10.1007/978-3-030-30859-9_21

[7] Akyildiz, I.F. Holographic-type communication: A new challenge for the next decade / I.F. Akyildiz, H. Gu // ITU Journal on Future and Evolving Technologies. – 2022.

[8] Long, Q.; Chen, Y.; Zhang, H.; Lei, X. Software defined 5G and 6G networks: A survey. Mob. Netw. Appl. 2019, 1–21.

[9] Navarro-Ortiz, J.; Romero-Diaz, P.; Sendra, S.; Ameigeiras, P.; Ramos-Munoz, J.J.; Lopez-Soler, J.M. A survey on 5G usagescenarios and traffic models. IEEE Commun. Surv. Tutor. 2020, 22, 905–929 (10)

[10] Ateya, A.A., Muthanna, A., Vybornova, A., Algarni, A.D., Abuarqoub, A., Koucheryavy, Y. and Koucheryavy, A., 2019. Chaotic salp swarm

algorithm for SDN multi-controller networks. Engineering Science and Technology, an International Journal, 22(4), pp.1001-1012.

[11] Ge, X.; Li, Z.; Li, S. 5G software defined vehicular networks. IEEE Commun. Mag. 2017, 55, 87–93.

[12] Muthanna, A.; Shamilova, R.; Ateya, A.A.; Paramonov, A.; Hammoudeh, M. A mobile edge computing/software-definednetworking-enabled architecture for vehicular networks. Internet Technol. Lett. 2020, 3, e10 (10)

Microservice-Based Fog Testbed for 6G Applications

Ekaterina Kuzmina
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
kuzmina120601@yandex.ru

Meriem Tefikova
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
tmrvmr@mail.ru

Artem Volkov
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
artemanv.work@gmail.com

Ammar Muthanna
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
muthanna.asa@spbgut.ru

Abdelhamied A. Ateya
Department of Electronics and Communications Engineering, Zagazig University, Zagazig 44519, Egypt;
EIAS Data Science Lab, College of Computer and Information Sciences,
Prince Sultan University,Riyadh 11586, Saudi Arabia
aateya@psu.edu.sa

Andrey Koucheryavy
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
akouch@mail.ru

***Abstract*— This paper provides a real-time fog computing model based on a microservice architecture that enables testing and modeling of eventual implementations of ultra-reliable low-latency communications (uRLLC) services. The work provides fog-based architecture for sixth-generation cellular (6G) applications, including telepresence and uRLLC. A testbed of a robot swarm was developed to prototype the proposed network architecture. Computing tasks are offloaded and handled based on a proposed microservice scheme introduced to meet the 6G requirements. Furthermore, we developed a novel migration scheme for the proposed architecture to support the mobility of end devices. The optimum server for migrating computing tasks is allocated by solving a proposed optimization problem using particle swarm optimization (PSO). All proposed algorithms were implemented in the developed prototype. The proposed work is introduced to provide an architectural foundation for testing fog-based 6G applications and services and to implement and test novel network methods in the future.**

Keywords*— fog computing, microservice architecture, 6G networks, telepresence services, uRRLC, Dobot Magician

I. Introduction

The sixth-generation cellular (6G) system is expected to provide novel services and applications not supported by the previous cellular generations [1]. Such applications include ultra-reliable low-latency (uRLL) applications such as telepresence, holographic, and haptic communications [2]. However, the design of communication networks for 6G applications faces many challenges due to the ultimate requirements of such applications [3]. The 6G systems should achieve ultra-high availability, reliability, and resilience. The end-to-end latency of the communicated data over 6G networks should be sub-millisecond [4]. Developing a testbed for implementing 6G algorithms, methods, and applications is also challenging.

Telepresence services enable users to communicate virtually, even if they may be miles apart, considering using telepresence technologies [5]. Everyone can establish a real-time sensation of presence and interaction using these services. Technologies like virtual reality, augmented reality, and holography are employed to implement this paradigm [6]. Telepresence can be used in various essential areas, such as remote medical diagnosis, consultations, and industrial automation [7].

The introduction of telepresence services significantly improves the quality of life, optimizes the use of human resources, and meets sustainability requirements. However, designing communication networks that provide telepresence services faces many design challenges, including dependability, reliability, availability, and latency [8]. Telepresence services are ultra-reliable low-latency communication (uRLLC) that require ultra-high data rates and extremely-low latency. Such services are one of the main services of the 6G, as per the third-generation partnership project (3GPP) and the international telecommunication union (ITU) [9, 10].

It is possible to achieve high data transmission rates, high reliability, and fault tolerance for the 6G networks by optimizing particular network segments. Monolithic programs can be swapped out for microservice architectures, and popular cloud computing can be replaced with distributed edge computing (e.g., fog computing) [11]. Microservice architecture, the first element of this optimization, enables extensive monolithic systems to be broken into a collection of

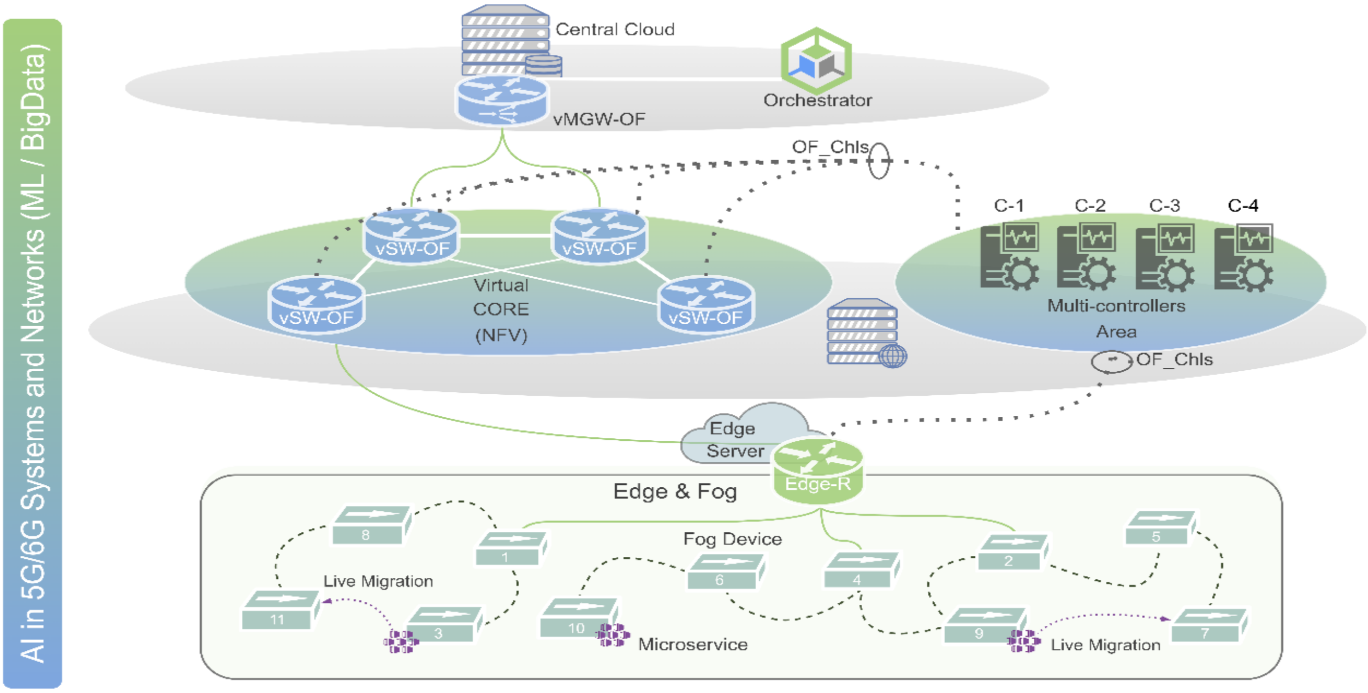


Fig. 1. Fog-based 6G network structure.

interconnected components, known as microservices, which cooperate as a single application [12]. This improves the system's fault tolerance, decreases the deployment complexity, boosts the application's scalability, and speeds up the introduction of new features compared to the conventional approach [13].

While cloud computing always involves data processing only on a remote centralized server, fog computing enables data processing in the end device's domain [14]. This can be achieved by distributing the processing of computing tasks to different fog points through remote micro-servers and user end devices[15]. The data transmission delay can be reduced by processing data closer to the user, hastening the response to a specific user request. Moreover, fog computing enables the creation of "virtual clouds" (also known as computational fog) by creating clouds "out of thin air" [16].

The ultimate objective of this work is to investigate novel strategies for establishing real-time microservice-based fog architecture to implement uRLLC 6G services. The main contributions of this work are summarized as follows.

- Developing a fog-based architecture for 6G applications.
- Developing a testbed of a swarm of robots based on the proposed fog architecture.
- Configuring the developed robots testbed.
- Developing a microservice architecture for the proposed system.
- Implementing the developed microservice scheme on the developed testbed.
- Developing a migration scheme for migrating microservices in the developed system.
- Optimizing the performance of the proposed migration scheme by selecting the optimum server for migration using particle swarm optimization (PSO).

II. Fog-Based Network Structure

Fig. 1 presents the proposed fog-based architecture of 6G networks. The higher layer is the core network layer, which deploys a network orchestrator and a central cloud unit of multiple servers with significant computing resources. The network orchestrator is the primary resource management tool, including processing, data storage, bandwidth, and energy.

The core network also deploys the software-defined networking (SDN) paradigm with multiple SDN controllers. SDN is mainly deployed to improve network flexibility, availability, resilience, and security [17]. Network function virtualization (NFV) is deployed over SDN to offer a software-oriented approach to network administration that allows managing network resources through network interfaces [18].

The lower layer of the architecture (Edge & Fog network) is the distributed edge layer that deploys two levels of edge servers: multiple access edge computing (MEC) servers and fog computing servers. Fog nodes interact with edge routers (Edge-R) to transmit data to the higher level of the proposed architecture.

The Edge & Fog network segment nodes are machines that process incoming data through fog computations. Each fog device offers a microservice or a set of them with specific functionality. These devices collectively provide services organized based on a microservice architecture. Microservice migration is also enabled for load balancing, scalability, and application deployment flexibility in the Edge & Fog segment [19]. It is important to pay close attention to the artificial intelligence (AI) block, which permeates every level and component of the design. With the advancements in chip element design and architecture, it is now possible to integrate AI functions for managing network resources [20].

III. Proposed Testbed

We considered a swarm of robots as a use case to implement the previously proposed architecture. Fig. 2 presents the proposed prototype implementation. The Edge & Fog segment introduced in Fig.1 is implemented in the proposed robot implementation.

The proposed testbed consists of four machines (i.e., robots), each consisting of two main parts: moveable and logical.

Moving part: The robotic manipulator is employed to control the stand's dynamics. The deployed robots are configured with the parameters introduced in Table. I.

TABLE I. Configuration parameters of the moving part of the deployed swarm of robots

Parameter	Value
Working area	340 mm in radius
Number of axles	Four
Accuracy	repeatability up to 0.02 mm
	placement up to 0.2 mm
Power supply	50/60 Hz
	100-240V
Power consumption	200 W

The dynamic segment was built using a flexible 4-axis robotic manipulator deployed in our lab and can carry out a variety of robotics and automation-related tasks. Figure 3 presents the considered robotic arm compatible with many programming languages, including C++, C#, Python, and Java [21].

Logical part (Fog-nodes emulation): This part represents the edge segment implementation. Four Raspberry Pi 4 Model B microcomputers are employed as the stand's logical component to provide edge computing capabilities. The main characteristics of the deployed fog nodes (i.e., Raspberry Pi nodes) are introduced in Table II.

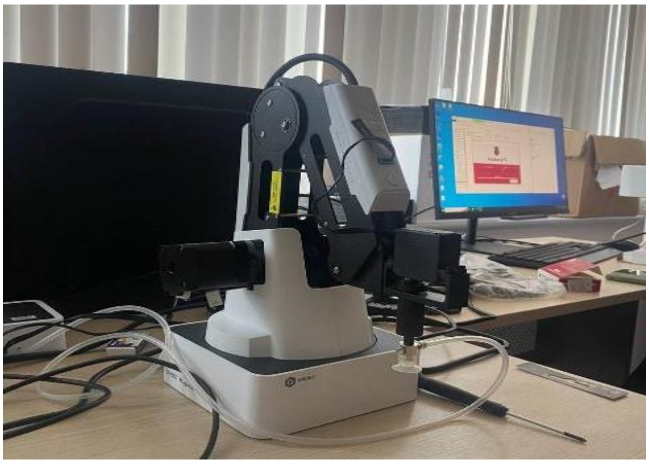


Fig. 3. Robotic arm as an element of the dynamic segment.

TABLE II. Configuration parameters of the robots' embedded fog nodes

Parameter	Value
Processor	Broadcom BCM2711
	Running at 1.5GHz on four ARM Cortex-A72 cores
RAM	LPDDR4-3200 SDRAM (8 GB)
Wireless interfaces	Bluetooth 5.0, BLE, and 802.11ac Wi-Fi
Storage	32 GB MicroSD memory card slot
General-purpose input/output (GPIO)	GPIO header with 40 pins
Power requirements	USB-C or GPIO at DC 5V, 3A

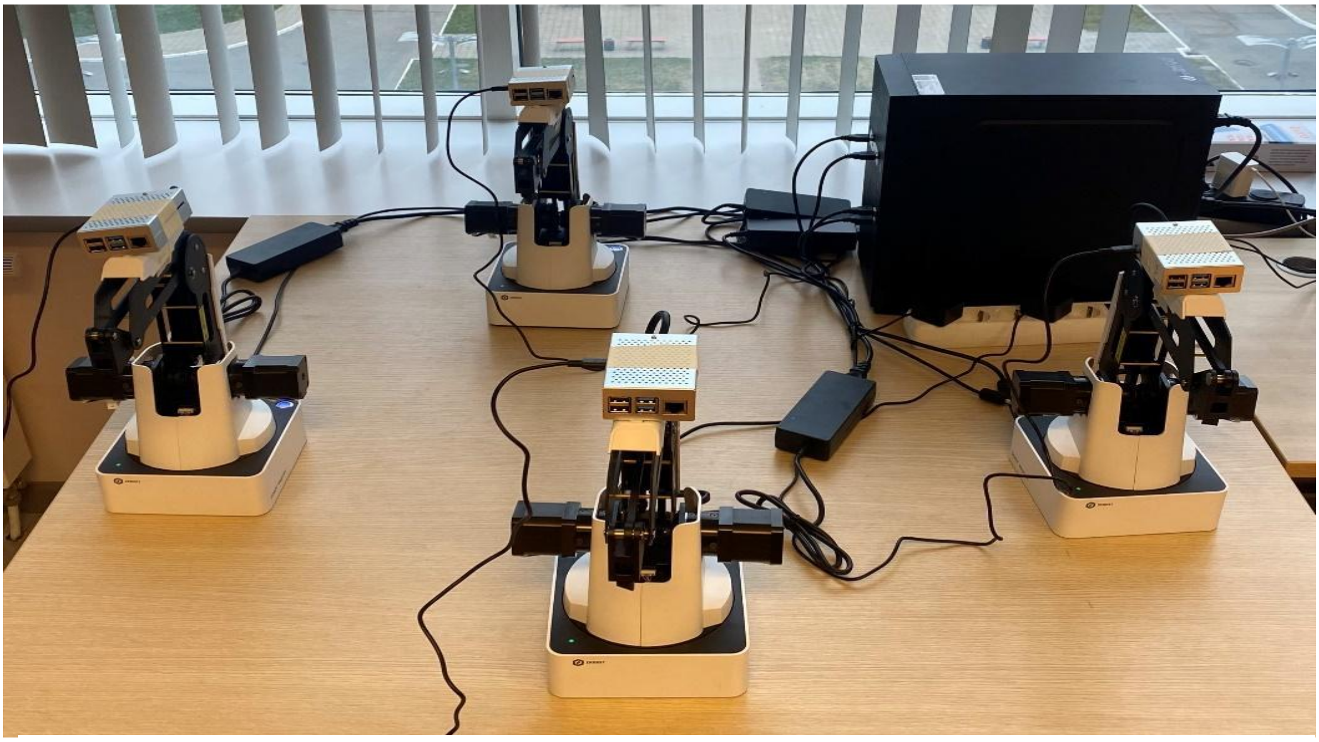


Fig. 2. The developed testbed.

A. *Moving part of the Developed Testbed*

The movement system of the proposed four robots was implemented using a program code written in Python version 3 that moves the robots over the optimum trajectory along a single course. Thus, the system was built to simulate the motion of a swarm of particles. However, in such architectures, veering off course can frequently happen for many reasons. The construction of the moving part of robots considers any deviation of one of the manipulators from the overall trajectory of the swarm because it is important to tackle several difficulties when deviations occur. Thus, three of the manipulators will continue moving synchronously along a typical trajectory after a certain period (Δt). In contrast, the fourth manipulator will deviate and move randomly to the other three.

The system is developed using client-server architecture, where two personal computers (PCs) connected to the same local network act, respectively, as a client and a server, to enable remote and flexible control of the system. Fig. 4 presents the network diagram of the dynamic segment

PC#1 is connected to the switch as a client, and PC#2 is linked to the switch as a management server over an Ethernet connection. The client PC is connected to four Dobot Magician robots via individual USB 3.0 connectors. Sockets, which are programming interfaces for inter-process communication, are used for communication between the server and the client. Logical addresses and ports are used for the communication between devices utilizing the TCP/IP protocol stack. The Internet protocol version 4.0 (IPv4) and version 6.0 (IPv6) are used in the network layer of the proposed structure. For the IPv4 and IPv6 protocols, the address is a structure of 32 bits and 128 bits, respectively, while the port number is an integer value between 0 and 65535 (for the TCP protocol).

The pair of address and port numbers define the socket. The fundamental benefit of using sockets is that data can be sent and received immediately through the socket. As a result, it is possible to transfer data between two or more devices without getting bogged down in the details of the protocols. The UDP protocol was chosen for data transfer between the server and the client to achieve a high data rate.

We built four client listening sockets on four separate official UDP ports, an external client IP address, and two server sockets that deliver commands to the corresponding client sockets to achieve complete remote control of robots. This architecture enables efficient remote control of robots.

B. *Coordinate Generation*

The robotic arm's movement limit system was taken into account to create random coordinates. This plan uses the difference between two hemispheres to calculate space, along with limits on the movement of the manipulator itself. Using the four half-planes introduced in Fig.5, any combination of these half-planes can be used to get the coordinates (X, Y, Z). The limitations of the three coordinates as introduced in Table III.

TABLE III. LIMITATIONS OF THE COORDINATES OF THE ROBOTIC ARM'S MOVEMENT AREA

Coordinate	Limits
X	$X \in (-120, 320)$
Y	$Y \in (-300, 320)$
Z	$Z > 0$

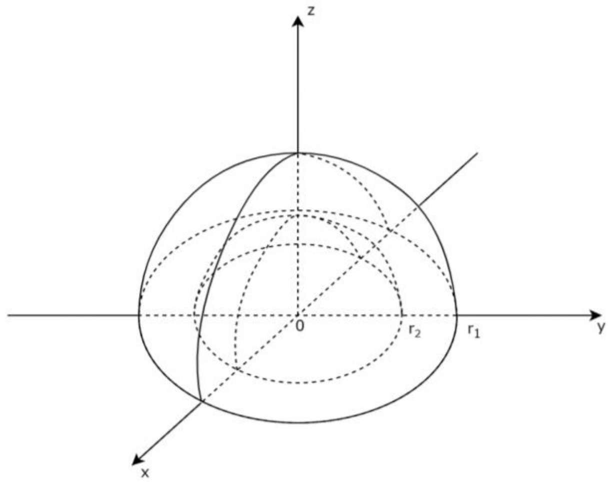


Fig. 5. Dobot magician movement restrictions.

The generated coordinates should meet the requirements introduced in Eq. (1) to be included in the valid list.

$$r_1{}^2 < x^2 + y^2 + z^2 < r_2{}^2 \ , \ r_1 = 300, r_2 = 180 \qquad (1)$$

Where r$_1$ and r$_2$ are the radiuses of the outer and inner spheres, respectively, calculated using the manipulators' known coordinate plane (X, Y, Z). The outer sphere, with the radius r$_1$, is denoted as Ω_1, while the inner sphere, with the radius r$_2$, is denoted as Ω_2.

Additionally, coordinates must be generated to simulate the deviance of one of the four robots from the general swarm trajectory at time Δt. In order to achieve this, we rotate the

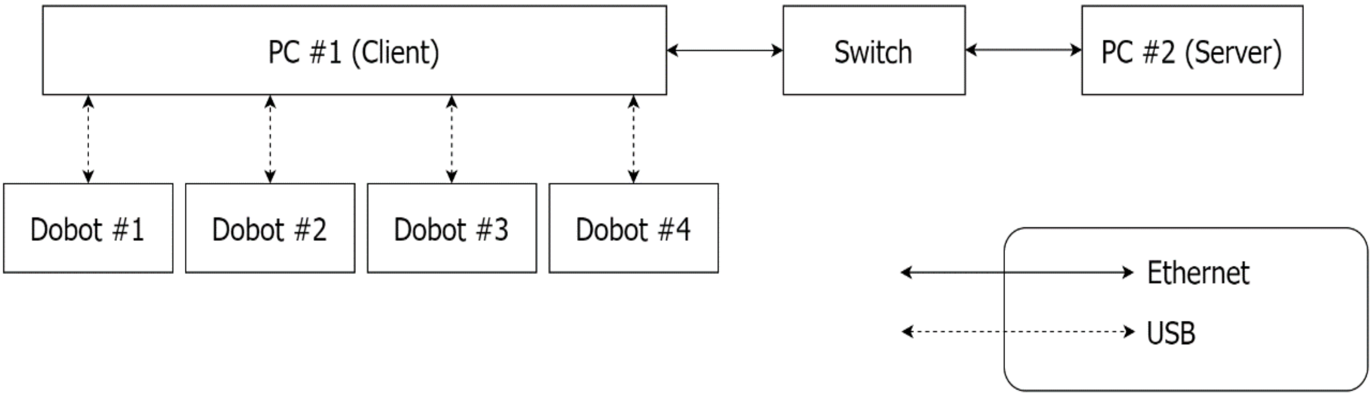
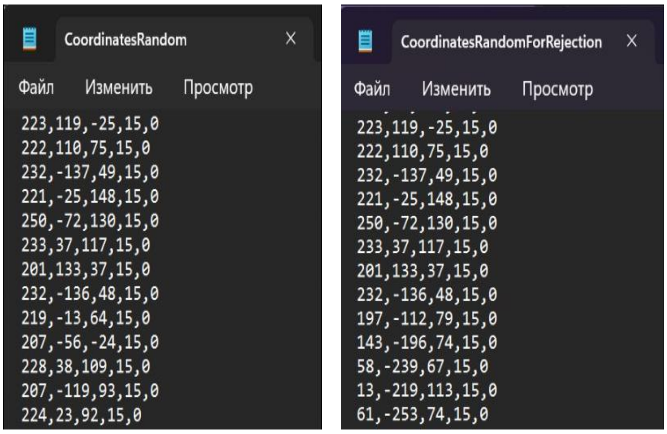


Fig. 4. Dynamic segment structure based on robotic manipulators.

manipulator artificially to a different half-plane by setting the corresponding coordinates directly. Then, after the intended movement to the last of them, produce coordinates for movements only in the new half-plane after the time Δt. We used Dobot Studio to find the turn's coordinates. This fulfills the abovementioned parameters while allowing the robotic arm to travel freely to a new half-plane.

In addition to the previously mentioned requirements, the maximum and minimum values of X, Y, and Z were determined using factory software to construct coordinates. Shifting the manipulator to a different plane changes the X, Y, and Z limits. We receive 2 CSV files with moving coordinates as the program's output. The server then processes these files. Figure 6 displays fragments of the received files with the coordinates.



```
223,119,-25,15,0
222,110,75,15,0
232,-137,49,15,0
221,-25,148,15,0
250,-72,130,15,0
233,37,117,15,0
201,133,37,15,0
232,-136,48,15,0
219,-13,64,15,0
207,-56,-24,15,0
228,38,109,15,0
207,-119,93,15,0
224,23,92,15,0
```

```
223,119,-25,15,0
222,110,75,15,0
232,-137,49,15,0
221,-25,148,15,0
250,-72,130,15,0
233,37,117,15,0
201,133,37,15,0
232,-136,48,15,0
197,-112,79,15,0
143,-196,74,15,0
58,-239,67,15,0
13,-219,113,15,0
61,-253,74,15,0
```

Fig. 6. Fragments of the received coordinates.

The manipulator movement characteristics must be considered when solving the given problem. The robot travels in a straight line between coordinates, occasionally resulting in an overshoot. To avoid this issue, we produce two coordinates (x_1, y_1,z_1) and (x_2, y_2,z_2) for each loop iteration and add a requirement that verifies the plane going through the two points acquired does not encroach upon the sphere Ω_2. This circumstance enables the robotic arm to stay inside the permitted positions in space.

C. *Clint-Server Model*

Two modules with the program code can be found in the server component. The first module examines the coordinates file for the first three robots (CoordinatesRandom.csv), and the second module examines the coordinates file for the fourth robot that deviates from the norm (CoordinatesRandomForRejection.csv). Only the names of the files used for analysis and the ports used to link the server with the client to which the robots are connected differ in these modules' code. Three sets of client socket addresses are defined in the program's main() function.

A socket is created using the socket.socket (*socket.AF_INET*, *socket.SOCK_DGRAM*) command, which is used to send datagrams using the UDP protocol (the first three robots use the official UDP ports 8000, 8880, and 8888, while the fourth deviant robot uses the official UDP port 8400). The code to read the coordinates from the file is run after creating the socket, and the data array it returns is provided to the function to transmit commands to client sockets. Each socket is closed when the loop is completed.

Four client programs make up the client portion of the software package. Each robot listens to a different port in order to maintain command parallelism. Moreover, each robot must be initialized on a different USB port to implement the chosen solution. Every client program produces a distinct socket that is linked to the appropriate port in order to receive orders from the server. As a result, the port for the USB connection between the robot and the client computer is chosen at the start of the program execution. The robot is connected to it for further control inside the *initialDobot()* function utilizing the common set of *DobotDllType* library functions designed for this operation. In the event of a successful connection, the robot's usual operating parameters are set. In the event of an initialization error, the module terminates, and an error message is sent to the console.

The robot startup function is initially called on the port given at the program's beginning in the main() function. A socket is then constructed and defined to the client computer's external IP address and the designated port. The socket is currently listening, which means it watches for messages (i.e., commands). The arriving data is decoded after configuring the buffer size of messages of the socket. An *OK* code is transmitted to the server socket in response to the successful reception of the message. The signaling flow of the proposed code is presented in Fig.7.

The robot moves when the received coordinates are relayed to it. After sending all the coordinates to the robot, the client PC should obtain its current position, then determine how it compares to the one the server needs. The *ERROR* code is provided to the server if the locations do not match; otherwise, the *OK* code is performed. The socket is closed when the loop is finished. Fig.7 shows a description of how the program's job is done. Only the port numbers on which the socket is created and the port numbers on which the robot is defined vary between client programs.

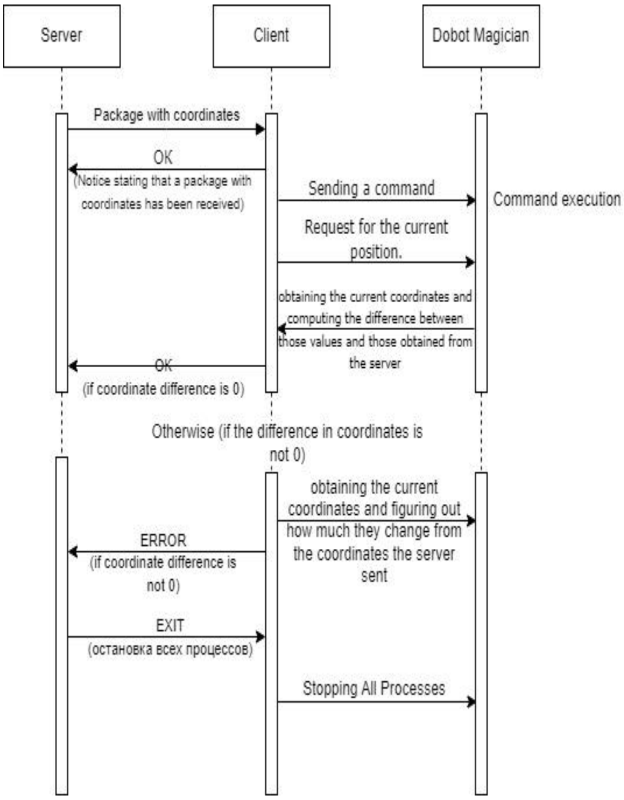

Server
Client
Dobot Magician
Package with coordinates
OK
(Notice stating that a package with coordinates has been received)
Sending a command
Command execution
Request for the current position.
obtaining the current coordinates and computing the difference between those values and those obtained from the server
OK
(if coordinate difference is 0)
Otherwise (if the difference in coordinates is not 0)
obtaining the current coordinates and figuring out how much they change from the coordinates the server sent
ERROR
(if coordinate difference is not 0)
EXIT
(остановка всех процессов)
Stopping All Processes

Fig. 7. Signaling flow of the proposed logical code.

D. Logical part of the testbed

The prototype's logical component was developed using Raspberry Pi 4 Model B microcomputers, which are little machines with an ARM Cortex-A72 processor clocked at up to 1.5 GHz and 8 GB of RAM. The Linux operating system is compatible with this microcomputer, ensuring optimal performance [22]. The Raspberry Pi 4 Model B is an effective and adaptable platform for a wide range of applications that can be used in both domestic and commercial settings. In the large-scale nebula model, the Raspberry Pi is a microprocessor on which one or more microservices with particular capabilities are installed. Microservice migration techniques can also be implemented using Raspberry.

The core network uses a deployed orchestrator to manage and monitor network parts. Kubernetes (K8s) is used to automatically launch, scale, and manage containerized applications in a cluster [23]. Applications can be executed in isolated environments with the help of containers. The management of containers and their deployment on a cluster of devices may be automated with Kubernetes, which streamlines the development process. Kubernetes is the ideal choice for our booth because it also offers scalability and fault tolerance. K8s offers a general API that enables you to run applications on a cluster using different container platforms like Docker. By polling only the master node and gathering network statistics, the offered API will enable the production and monitoring of each cluster node.

Docker and all the necessary Kubernetes software (kubelet, kubeadm, and kubectl) were installed on each Raspberry Pi. The child nodes (hence referred to as workers) and the main node (master-node) are both monitored by the Kubelet service, which begins and operates on each cluster node. Kubeadm is a tool used to manage a Kubernetes cluster as well as to configure nodes. To transmit commands to a Kubernetes cluster, you require the Kubectl program. The cluster worker nodes were set up and connected to the master node after the appropriate components had been installed. Our microcomputers operate as worker nodes, and our server serves as a master node. Fig. 8 presents the cluster's organizational structure.

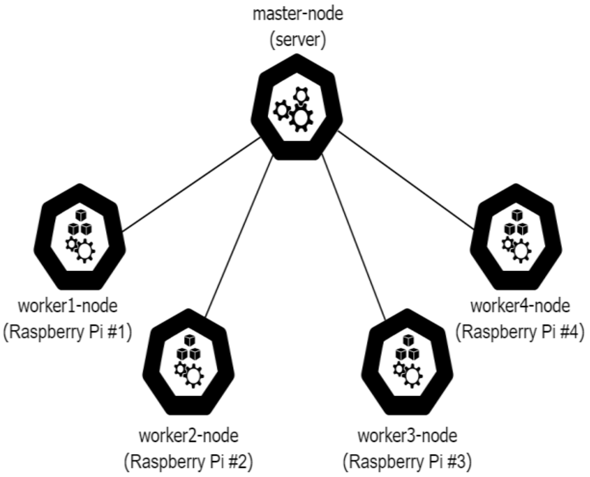


Fig. 8. Clustering scheme.

IV. Microservice architecture & migration scheme

The microservice architecture enables splitting up a monolithic application's power into sections that are part of the major service's functionality. The application developers specify the tasks carried out by a specific microservice scheme. The result of such a system is a unique product that does not differ in functionality from the source but is more reliable and fault-tolerant in operation, with a significant increase in the performance and scalability of the entire system. The group of microservices into which the application is divided functions as a single unit, where each element interacts with all other elements. Fig. 9 presents a comparison between traditional architecture and microservice architecture.

Each microservice can be implemented using kernel-level virtualization (i.e., containerization). Organizing the architecture of various environments in one processing unit without needing a hypervisor is called containerization, a specific type of virtualization. In the conventional virtualization method, separate virtual environments are built within the hypervisor infrastructure. This method works primarily because the hypervisor's primary duty is ensuring that guest operating systems communicate with the host operating system.

Environments are created using the kernel's capabilities while using containerization. This procedure separates the kernel's capacity into parts, each only accessible by a single container, preventing the use of the system's other resources [24]. By excluding the hypervisor's impact on the system, which manifests as a decline in the performance of newly formed services and the trouble of packaging and delivering applications, a new stage in the development of virtualization has made it feasible.

In the proposed scheme, microservices are created as containers, each holding a specific portion of a major service's functionality. Planning for instances where individual services fail, even those beyond their control, is vital to provide better system reliability. An example of such cases is when the logical component of the system is in motion during the work, and one robot unit deviates from the group of others.

In this case, the tasks should be migrated to maintain the system and prevent failure. Thus, we developed a microservices migration mechanism in order to prevent this scenario. The migration scheme is introduced to maintain the service while moving between different coverage areas and to improve system performance and resource usage.

Migrating microservices offers many benefits that can be summarized as follows [25, 26].

- Load balancing: Migration assists in distributing the network load among all the computing power if it is severely out of balance. The service switches from an overloaded host to a more powerful or less loaded one.
- Fault tolerance: Failures must be foreseen and proactively handled to avoid their negative effects on the performance of the system and the running applications. Task migration provides a way for fault tolerance by providing another path for data offloading.
- Power administration: Migration makes it possible to distribute microservices to the fewest number of active servers dynamically. In order to save energy, inactive servers (i.e., those with no offloaded tasks) can be switched to sleep mode.

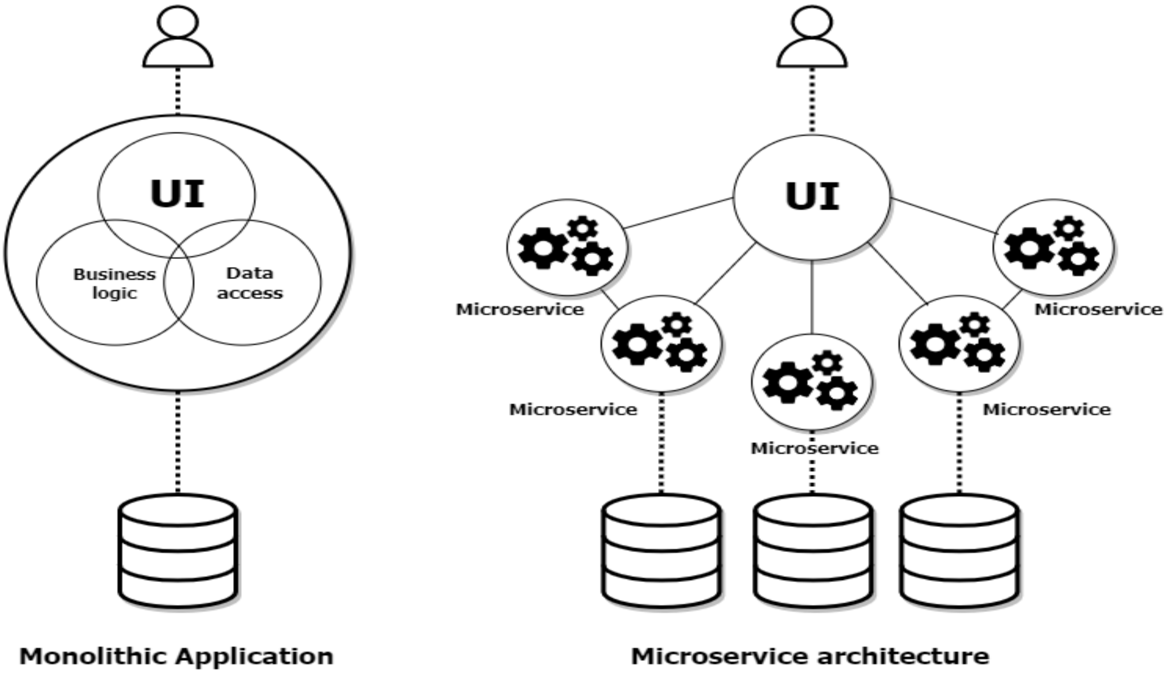


Fig. 9. Traditional and microservice application architecture.

- Prompt maintenance of the system: Suppose a particular server has to be updated or maintained. In that case, all of the microservices hosted there can be moved to a different server in the user's access zone, and the required maintenance can be carried out simultaneously on both servers.

When migrating a microservice live over a wide-area network (WAN), the online services must remain while the processor state, memory pages, and disc storage are moved from a source server to a remote destination server. Pre-copy and post-copy are the two most common migration procedures used concurrently. While the microservice (i.e., container) is still running on the source server and creating memory pages (modified memory pages), pre-copying first copies all pages of memory and disc storage to the destination server. The freshly created pages must be iteratively copied to the destination at a pace quicker than the rate at which changes are made in memory after copying the contents of the disc storage. Finally, the remaining pages and the processor state are transferred to the destination server, the source container is resumed on the new server, and the container is halted on the original server.

On the other hand, post-copy first suspends the container on the source server, moves its processor state to the destination server, resumes the container on the new server, and then migrates memory pages and disc storage. Memory pages and disc storage are actively transferred to the destination server only once since the container does not cause dirty pages to be generated on the source server.

Pre-copying, by its capacity to recover the state of the container at the source, even if the migration fails, clearly triumphs in terms of reliability. Due to the source server's lack of memory page generation, which necessitates an iterative update memory transfer, the post-copy has a quick migration time.

The post-copy technique is used in our model since it must migrate fast without the user recognizing when the moving part leaves the coverage region.

The selection of the destination server is challenging with the considered migration scheme. Selecting the optimum server for migration is an optimization problem that can be solved using practical swarm optimization (PSO) [27].

V. OPTIMIZING MIGRATION SCHEME

A. *Problem Formulation*

The optimization problem seeks the optimum selection of hardware required for the migration processes (i.e., optimum powerful server). The problem is a minimization problem that can be defined as follows.

$$\min_{S} f(x) \qquad (2)$$

Where f(x) is the objective function, and S is the optimum server for the migration process. The proposed objective function of the optimization problem is introduced to achieve the fastest microservice response to a request, maximum reliability and availability, and load balancing. Consequently, the objective function incorporates the server's current CPU load (in%) and RAM usage (in%). It can be defined as follows.

$$f(x) = \sum_{i=1}^{n} \mu_i \cdot \emptyset_i \qquad \forall \quad i \in \mathbb{R} \qquad (3)$$

Where f(x) is the objective function (calculated in %), μ_i is the weight of the i^th^ parameter, φ_i is the ith parameter that describes the server's state, and n is the number of selected parameters. We consider only two parameters for the proposed system: CPU usage and storage requirements. Both parameters are given equal attention (i.e., equally weighted); thus, the objective function is modified as follows.

$$f(x) = 0.5 \cdot CPU + 0.5 \cdot RAM \qquad (4)$$

B. *Particle Swarm Optimization*

The PSO is a metaheuristic algorithm that simulates how a group of birds behaves when seeking the best position in space to approach a food source. The algorithm addresses optimization issues, such as determining a function's lowest or maximum value.

An initial set of "particles" with individual positions and speeds in the multidimensional space is created at the beginning of the procedure. Each particle moves across space, adjusting its speed and position iteratively to arrive at the food position (i.e., optimum solution). The particles update their location in space by exchanging information with other particles in the swarm, using Eq.(2) [27].

$$v_{i,j} = v_{i,j} + c \cdot r_1 \cdot (p_{i,j} - x_{i,j}) + c \cdot r_2 \cdot (p_{g,j} - x_{i,j}) \quad \forall \quad r_1, r_2 \in [0,1] \tag{5}$$

$$x_{i,j} = x_{i,j} + v_{i,j} \tag{6}$$

Where $v_{i,j}$ is the speed of the particle i, $p_{i,j}$ is the best position found, $p_{g,j}$ is the best-found position over all particles, $x_{i,j}$ is the position of the particle i in space, c is a constant with value 2, and r_1, r_2 are random numbers.

The value of the objective function is calculated for each available device using equation (4) to identify the best device for migration. The next step is to select a device from the available devices with the minimum objective function.

VI. CONCLUSION

Along with the advancement of information technology, the demand for society to incorporate novel concepts in this field into everyday life is growing. Unfortunately, the current infrastructure established by conventional communication networks is inadequate for introducing new technologies. This procedure requires gradually modernizing a well-established, global infrastructure that has expanded worldwide for years. Consequently, research on microservice architecture networks and fog computing is growing in significance. Using these and other related solutions, you can create more efficient algorithms and models to satisfy the increasing demands of contemporary technologies. Using the under-consideration architecture to research uRLLC services would accelerate bridging the gap between current data and the Tactile Internet concept's specifications.

ACKNOWLEDGMENT

The studies at St. Petersburg State University of Telecommunications. Prof. M.A. Bonch-Bruevich was supported by the Ministry of Science and High Educationof the Russian Federation by grant number 075-15-2022-1137.

REFERENCES

[1] C. X. Wang et al., "On the road to 6G: Visions, requirements, key technologies and testbeds," *IEEE Commun. Surv. Tutor.*, pp. 1–1, 2023.

[2] M. M. Azari et al., "Evolution of non-terrestrial networks from 5G to 6G: A survey*," IEEE Commun. Surv. Tutor.*, vol. 24, no. 4, pp. 2633–2672, Fourthquarter 2022.

[3] C. Yeh, G. D. Jo, Y.-J. Ko, and H. K. Chung, "Perspectives on 6G wireless communications," *ICT Express*, vol. 9, no. 1, pp. 82–91, 2023.

[4] L.-H. Shen, K.-T. Feng, and L. Hanzo, "Five facets of 6G: Research challenges and opportunities," *ACM Comput. Surv.*, 2022.

[5] D. Calandra, F. G. Pratticò, A. Cannavò, C. Casetti, and F. Lamberti, "Digital twin- and extended reality-based telepresence for collaborative robot programming in the 6G perspective," *Digit. Commun. Netw.*, 2022.

[6] A. A. Ateya, A. Muthanna, A. Koucheryavy, Y. Maleh, and A. A. A. El-Latif, "Energy efficient offloading scheme for MEC-based augmented reality system," *Cluster Comput.*, vol. 26, no. 1, pp. 789–806, 2023.

[7] D. M. Hilty et al., "A review of telepresence, virtual reality, and augmented reality applied to clinical care," *J. Technol. Behav. Sci.*, vol. 5, no. 2, pp. 178–205, 2020.

[8] F. Hernandez, M. Waechter, and A. C. Bullinger, "A first approach for implementing a telepresence robot in an industrial environment," in Advances in Human Factors and System Interactions, Cham: Springer International Publishing, 2021, pp. 141–146.

[9] M. Osama, A. A. Ateya, S. Ahmed Elsaid, and A. Muthanna, "Ultra-reliable low-latency communications: Unmanned aerial vehicles assisted systems*," Information (Basel)*, vol. 13, no. 9, p. 430, 2022.

[10] G. P. Sharma et al., "Towards deterministic communications in 6G networks: State of the art, open challenges and the way forward," arXiv [cs.NI], 2023.

[11] V. Benavente, L. Yantas, I. Moscol, C. Rodriguez, R. Inquilla, and Y. Pomachagua, "Comparative analysis of microservices and monolithic architecture," in 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN), 2022, pp. 177–184.

[12] T. Cerny, A. S. Abdelfattah, V. Bushong, A. Al Maruf, and D. Taibi, "Microservice Architecture Reconstruction and Visualization Techniques: A Review," in 2022 IEEE International Conference on Service-Oriented System Engineering (SOSE), 2022, pp. 39–48.

[13] A. Razzaq and S. A. K. Ghayyur, "A systematic mapping study: The new age of software architecture from monolithic to microservice architecture—awareness and challenges," *Comput. Appl. Eng. Educ.*, 2022.

[14] L. Tawalbeh, F. Muheidat, M. Tawalbeh, M. Quwaider, and A. A. Abd El-Latif, "Edge enabled IoT system model for secure healthcare," *Measurement (Lond.)*, vol. 191, no. 110792, p. 110792, 2022.

[15] F. Al-Doghman, N. Moustafa, I. Khalil, N. Sohrabi, Z. Tari, and A. Y. Zomaya, "AI-enabled secure microservices in edge computing: Opportunities and challenges," *IEEE Trans. Serv. Comput.*, vol. 16, no. 2, pp. 1485–1504, 2023.

[16] B. Costa, J. Bachiega Jr, L. R. de Carvalho, and A. P. F. Araujo, "Orchestration in fog computing: A comprehensive survey," *ACM Comput. Surv.*, vol. 55, no. 2, pp. 1–34, 2023.

[17] A. Volkov, K. Proshutinskiy, A. B. M. Adam, A. A. Ateya, A. Muthanna, and A. Koucheryavy, "SDN load prediction algorithm based on artificial intelligence," in Communications in Computer and Information Science, Cham: Springer International Publishing, 2019, pp. 27–40.

[18] K. Kaur, V. Mangat, and K. Kumar, "A review on Virtualized Infrastructure Managers with management and orchestration features in NFV architecture," *Comput. Netw.*, vol. 217, no. 109281, p. 109281, 2022.

[19] W. Lv et al., "Microservice deployment in edge computing based on deep Q learning," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 11, pp. 1–1, 2022.

[20] A. Bandi, "A review towards AI empowered 6G communication requirements, applications, and technologies in mobile edge computing," in 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), 2022, pp. 12–17.

[21] "DOBOT Magician robotic arm," Automate. [Online]. Available: https://www.automate.org/products/shenzhen-yuejiang-technology-co-ltd-dobot/dobot-magician-robotic-arm. [Accessed: 07-May-2023].

[22] Raspberry Pi Ltd, "Raspberry Pi 4 Model B specifications –," Raspberry Pi. [Online]. Available: https://www.raspberrypi.com/products/raspberry-pi-4-model-b/specifications/. [Accessed: 07-May-2023].

[23] "Production-Grade Container Orchestration," Kubernetes. [Online]. Available: https://kubernetes.io/. [Accessed: 07-May-2023].

[24] S. Muhizi, A. A. Ateya, A. Muthanna, R. Kirichek, and A. Koucheryavy, "A novel slice-oriented network model," in Developments in Language Theory, Cham: Springer International Publishing, 2018, pp. 421–431.

[25] R. Pérez, M. Rivera, Y. Salgueiro, C. R. Baier, and P. Wheeler, "Moving microgrid hierarchical control to an SDN-based Kubernetes cluster: A framework for reliable and flexible energy distribution," *Sensors (Basel)*, vol. 23, no. 7, p. 3395, 2023.

[26] M. Adeppady, P. Giaccone, H. Karl, and C. F. Chiasserini, “Reducing microservices interference and deployment time in resource-constrained cloud systems,” *IEEE Trans. Netw. Serv. Manag.*, pp. 1–1, 2023.

[27] M. Jain, V. Saihjpal, N. Singh, and S. B. Singh, “An overview of variants and advancements of PSO algorithm,” *Appl. Sci. (Basel)*, vol. 12, no. 17, p. 8392, 2022.

Migration routing algorithm for microservice based Fog computing system

Ekaterina Kuzmina
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
kuzmina120601@yandex.ru

Meriem Tefikova
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
tmrvmr@mail.ru

Artem Volkov
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
artemanv.work@gmail.com

Ammar Muthanna
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
muthanna.asa@spbgut.ru

Andrey Koucheryavy
Department of Telecommunication Networks and Data Transmission,
Saint-Petersburg State University of Telecommunications
Saint Petersburg, Russia
akouch@mail.ru

***Abstract*— The new generation of communication networks has provided the basis for the realization of new classes of services, one of which are URLLC services. These services include concepts such as the Tactile Internet. The transmission medium for URLLC requires high bandwidth and reliability of communication channels as well as ultra-low latency in the network. To cover these requirements, various technologies such as fog computing and microservice architecture are being introduced into the network architecture. The latter is more often recommended together with support for migration of microservices between devices. This paper considers one of the most important migration tasks, namely determining an efficient path to the destination server. The parameters that need to be considered for efficient routing in microservice migration were identified, as well as the network architecture layer that should control the migration. The results show that the proposed solution is preferred for application over the AODV protocol.**

Keywords*— *Fog computing; 6G; Protocol; service migration

I. Introduction

Previously, communication networks could not meet all the requirements strictly set for new approaches of machine-to-machine communication, but now there is a qualitative leap in the capabilities of modern networks. The most recently released mobile communication standard is IMT-2020 (5G), which has greatly surpassed its predecessors in all directions. Data transmission speed, the number of devices connected to the network, network reliability - the fifth generation of communication networks provided growth of all these indicators. It is also important that the network latency has been significantly reduced [1]. All this allowed us to approach the realization of such concepts as the Tactile Internet. However, the requirements of the Tactile Internet to the transmission medium are very specific, for example, the network latency should not exceed 1 ms [2]. Moreover, in addition to the changing composition of the technological background of human life, its scale is also changing. The number of devices connected to the network is growing relentlessly, and at the same time the amount of traffic transmitted over the network is also increasing [3]. Classical approaches to the construction of communication networks will simply cease to cope with the service of such a large number of requests, which entails the impossibility of providing the necessary QoS. Nevertheless, researchers have already proposed a huge number of innovative solutions to avoid the worst-case scenario of communication networks development. These include the concept of Fog computing and microservice architecture. However, the implementation of new solutions in the upper layers of the network infrastructure requires partial or complete changes in the technologies used at the lower layers. Thus, for quality operation of a network containing nodes supporting microservice architecture, it is necessary to implement the migration of microservices between network devices, and this process necessitates the development of new routing principles, when considered in relation to different network topologies and architectures.

II. The AODV routing protocol

AODV (Ad hoc On-Demand Distance Vector) is an open-source reactive routing protocol which is used in wireless self-organizing networks [4]. It is designed for mobile ad-hoc MANETs, Sensor networks and other wireless networks. AODV creates and maintains dynamic routes between nodes.

The working principle of the protocol can be described as follows [5]:

1. When there is a need to transmit data, the sending node starts searching for a route to the destination node. It sends a RREQ (Route Request) message to all its neighboring nodes.

2. Neighboring nodes forward the RREQ message to their own neighbors, and this continues until the message

reaches a node that knows the route to the destination node, or has a record of a shorter route to it.

3. When a RREQ message reaches either the destination node or a node that knows the route to it, the corresponding node sends a RREP (Route Reply) message back to the source node, the initiator of the route search. The RREP message follows the path specified in the RREQ message. In the process of returning RREP message, each node encountered on the path remembers the route to the sending node. After receiving the RREP message, the source node knows the path to the destination node.

One of the main advantages of the AODV protocol is its efficiency in the use of network resources: since routes are established only, when necessary, a constant stream of service messages is not transmitted over the network, as in the work of proactive routing protocols. This allows you to reduce the load on the network and reduce the requirements for the resources of nodes. Moreover, the results of the protocol determine the shortest route to the destination node, which means that data transmission will require fewer transitions through transit nodes. AODV also has the property of self-organization, which allows network nodes to dynamically adapt to changes in the network.

AODV protocol is preferred and applicable to solve the problem of microservice migration route determination in Fog computing network, but it does not allow to consider all the necessary circumstances of path selection to the destination node, so there is a need to create a new solution.

III. Defining the parameters

The AODV protocol involves choosing the shortest route to the destination node, and the shortest route is determined by the number of hops required to reach the destination. The mentioned indicator is more often referred to as Hop-count. One hop implies passing through one network node, in other words, one transit section on the packet's path to the destination.

However, in the context of the task of route selection for microservice migration in the considered dynamic Fog computing network, it is required to take into account more network state parameters than the AODV protocol algorithm allows. This is due to the fact that the nodes in the network are represented by Fog devices performing different computational tasks, respectively spending their computational resources. It is necessary to balance the distribution of resources in the network, so it is necessary to choose a route whose transit nodes do not perform any resource-intensive tasks at the time of migration. For this purpose, the main server only needs to have information about the CPU and RAM utilization of each node in the network.

In addition, the link state indicators between the network nodes are also important for solving the described problem. Migration must be performed quickly and reliably enough, which requires from the communication channel a sufficiently low delay, and packet losses must also be minimal, otherwise the time spent on migration will grow.

Thus, the most important network state parameters for determining the microservice migration route in the current task are:

- CPU load on network nodes;
- RAM load on network nodes;
- Delay on communication channels between nodes;
- Share of packet loss on communication channels between nodes.

By tracking these parameters, it is possible to find a migration path that will be the most efficient, with the lowest latency, the lowest proportion of lost packets, and the most reliable due to the fact that the transit nodes, which covers the route, will have enough resources to pass the microservice through them, or their load will not approach the limit in contrast to the rest of the nodes in the network, in the process of migration to the destination.

The described characteristics should be collected in the form of real-time and continuous statistics for Fog zones. There should be a temporary or permanent data repository where the collected statistics will be recorded so that, if necessary, it will be possible to analyze the load in the zone over a certain period of time. This is also necessary for predicting the loads that will occur in Fog zones in the future.

IV. Proposed algorithm to route selecting

Turning to graph theory, it can be seen that a route has an important generalizing characteristic - the path cost, also referred to as Cost. The definition of a migration route should be based on the path cost, which will be calculated on the basis of the described considered parameters.

The routing tables of the network nodes can be represented by a graph, and then path finding algorithms in the graph can be used to find a path. One of the classical algorithms for solving such a problem is Dijkstra's algorithm. The algorithm assumes that each edge of the graph has its own weight, and the progression to the final node in the graph should follow the edges with the lowest weights among the possible neighbors of the origin node and transit nodes.

In the microservice migration pathfinding problem, there is no need to search for paths from the initial node of the graph to all other nodes, since the final destination is known in advance, it is determined by existing algorithms. Therefore, it is necessary to find only one path, which is the "shortest" path to the destination node. The "shortest" path is the route whose cost is minimal among the other available ones.

However, the costs of edges are not known in advance. They should be calculated and then the algorithm should be started. In general, the considered Fog computing network can be represented by the graph shown in Figure 1:

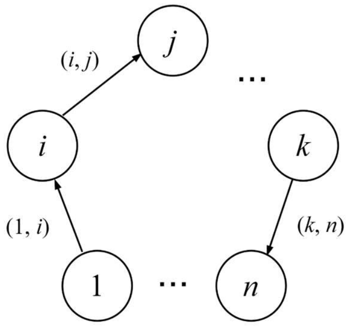


Figure 1. Generalized graph of the topology of the considered network

Figure 1 shows the generalized topology of the considered network. Let node 1 be the network node from which the microservice should be migrated, and node n be the network node to which the microservice should migrate. The graph is directed, since the microservice must exactly pass each node only once during the migration process, in other words, no loops in the routes should be formed.

Then, to determine the shortest route from node 1 to node n, we need to find all the weights of the edges encountered on the path to the destination node. The cost of an edge in this case should take into account not only the characteristics of the edge itself, but also the characteristics of the node, as was clarified earlier. Then the cost of an edge or link (i,j) will be determined as follows:

f_(i,j)= 〖edgeOptions〗_(i,j)+ 〖topOptions〗_j, (1)

Where 〖edgeOptions〗_(i,j) – parameters describing the state of the edge, and 〖topOptions〗_j – parameters describing the state of the vertex to which the edge leads. The mentioned parameters were previously defined as CPU load of node j, RAM load of node j, delays in communication channels (i,j), PacketLoss in communication channels (i,j). All these characteristics have different units of measurement and different priorities, so they need to be brought to an equivalent form later.

Based on expression 1, the cost of the entire route can be expressed as:

F_x (1,n)=∑▒f_(i,j) , (2)

and the required migration path is defined as:

min { F_x (1,n) }. (3)

Returning to the parameters describing the state of the network, as already mentioned, it is necessary to take into account that each of the selected parameters has its own magnitude of influence on the result and its own units of measurement. It is necessary to determine the importance of each parameter in relation to the others, since not all of them equally influence the choice of one or another path to the destination device. The importance of a parameter in this context is called its weight.

Parameter weights should be determined using methods of comparing a set of objects with each other based on the conditions of the task, such approaches include the Method of Pairwise Comparisons.

As a result of the calculations, the CPU, Delay, and PacketLoss parameters have weights of 0.275 and RAM has a weight of 0.175.

Thus, the target function from expression 1 of finding the cost of one edge (i,j) of the graph, with the found weights and certain parameters of the network state will be of the form:

f_(i,j)= 〖CPU〗_j·w_CPU+ 〖RAM〗_j·w_RAM+D_(i,j)·w_D+ 〖PL〗_(i,j)·w_PL, (4)

где w_CPU=w_D=w_PL=0,275, w_RAM=0,175.

Figure 2 shows the proposed microservice migration algorithm for route selection.

It is worth noting that the layer of the network architecture that controls this process (Figure 2) should be the Fog Computing Orchestrator. This task falls under the list of reasons for its use and implementation, because it is the orchestrator that deals with load distribution between Fog nodes, forecasting, prioritization of various tasks performed by Fog nodes, etc.

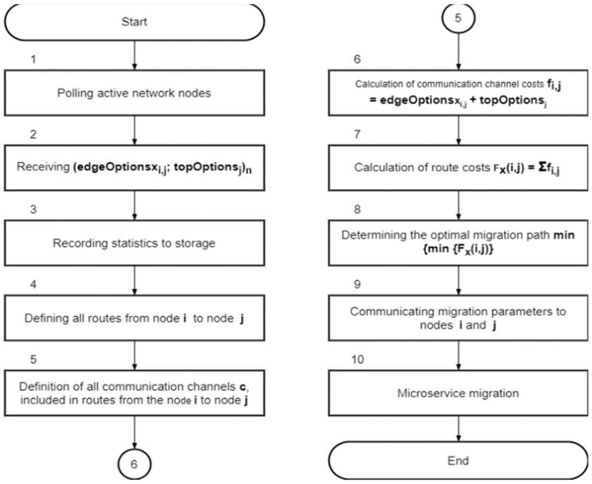

Start
1
Polling active network nodes
2
Receiving (edgeOptionsx$_{i,j}$; topOptions$_j$)$_n$
3
Recording statistics to storage
4
Defining all routes from node i to node j
5
Definition of all communication channels c, included in routes from the node i to node j
6
5
6
Calculation of communication channel costs f$_{i,j}$ = edgeOptionsx$_{i,j}$ + topOptions$_j$
7
Calculation of route costs F$_X$(i,j) = Σf$_{i,j}$
8
Determining the optimal migration path min {min {F$_X$(i,j)}
9
Communicating migration parameters to nodes i and j
10
Microservice migration
End

Figure 2. Proposed migration routing algorithm for microservice architecture

V. RESULTS

To test the performance of the proposed migration route determination algorithm and to evaluate its performance, 5 experiments were conducted on a model-based Fog computing network.

To conduct the experiments, the nodes of the model network were started and put into operation. Within the framework of these experiments, the nodes were subjected to artificial increase and decrease of load in order to collect objective statistics on the performance of two routing algorithms: the classical route selection algorithm of the AODV protocol, which is based on the Hop-count metric, i.e., evaluates paths according to the choice of the shortest route by the number of transitions through transit nodes to the destination server, as well as the proposed algorithm for determining the migration path, which is based on the search for a route on less loaded links connecting and less loaded links.

Summary graphs on the indicators of network nodes CPU, RAM, Delay in data transmission (Delay), packet loss (PacketLoss) when using the mentioned algorithms for 5 experiments are presented in Figure 3.

The results show that the proposed algorithm selects a route less prone to loss of transmitted packets, the superiority over the AODV protocol algorithm can be estimated at 7% in this indicator. Also, the proposed algorithm considers the load at each node of the route, which is confirmed by 77% and 14% reduction in CPU and RAM resources consumed at the transit nodes of the network, respectively. At the same time, the delay performance increased by 4%, which is explained by the fact that previously the CPU and Delay parameters were determined to be equivalent in the considered problem, and the network nodes were subjected to high loads of 90% CPU utilization, with the link delay not taking such high values in the measurements, respectively, the differences in the parameters were large, in which case CPU caused a strong increase in the cost of the whole route. However, the algorithm allows us to adjust the weights of the parameters and the list of parameters themselves as needed for a particular task.

The experimentally proven efficiency of the algorithm gives an understanding that when performing microservice migration in a dynamic Fog computing network, it is the proposed algorithm that should be used to select the microservice migration route. In other words, it is preferred for solving this problem because it shows less utilization of network resources when migrating a microservice along the route that the algorithm determines, which means that the proposed solution takes into account the peculiarities of the considered network and improves its performance.

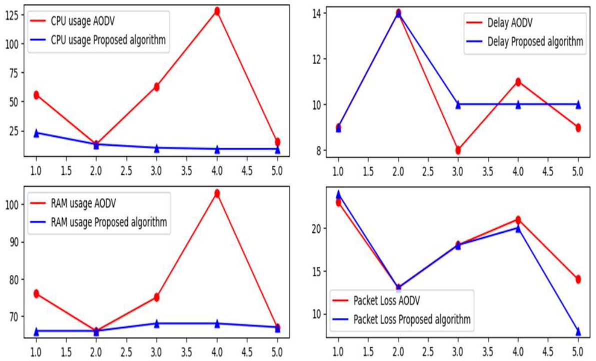


Figure 3. Load indicators when using the two algorithms

VI. CONCLUSION

In this paper, an algorithm for selecting the migration route of a microservice in a Fog computing network has been proposed. The architecture layer that controls the migration process has been defined. The developed solution has been tested together with the already classical approach used by the AODV protocol, and from the results obtained it is clear that the proposed algorithm is more efficient than the AODV protocol algorithm for solving the microservice migration path finding problem. By improving the network performance that is observed in the microservice migration process using the proposed algorithm, it is possible to achieve minimization of network failures and provide the required QoS, which is an undeniably important issue in the deployment of fifth generation communication networks.

ACKNOWLEDGMENT

The studies at St. Petersburg State University of Telecommunications. Prof. M.A. Bonch-Bruevich was supported by the Ministry of Science and High Educationof the Russian Federation by grant number 075-15-2022-1137.

REFERENCES

[1] Recommendation ITU-R M.2083-0: IMT Vision, “Framework and overall objectives of the future development of IMT for 2020 and beyond,” Sep. 2015.

[2] D. Calandra, F. G. Pratticò, A. Cannavò, C. Casetti, and F. Lamberti, “Digital twin- and extended reality-based telepresence for collaborative robot programming in the 6G perspective,” *Digit. Commun. Netw.*, 2022.

[3] A. A. Ateya, A. Muthanna, A. Koucheryavy, Y. Maleh, and A. A. A. El-Latif, “Energy efficient offloading scheme for MEC-based augmented reality system,” *Cluster Comput.*, vol. 26, no. 1, pp. 789–806, 2023.

[4] D. M. Hilty et al., “A review of telepresence, virtual reality, and augmented reality applied to clinical care,” *J. Technol. Behav. Sci.*, vol. 5, no. 2, pp. 178–205, 2020.

[5] M. Osama, A. A. Ateya, S. Ahmed Elsaid, and A. Muthanna, “Ultra-reliable low-latency communications: Unmanned aerial vehicles assisted systems*,” Information (Basel)*, vol. 13, no. 9, p. 430, 2022.

[6] V. Benavente, L. Yantas, I. Moscol, C. Rodriguez, R. Inquilla, and Y. Pomachagua, “Comparative analysis of microservices and monolithic architecture,” in 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN), 2022, pp. 177–184.

[7] A. Razzaq and S. A. K. Ghayyur, “A systematic mapping study: The new age of software architecture from monolithic to microservice architecture—awareness and challenges,” *Comput. Appl. Eng. Educ.*, 2022.

[8] L. Tawalbeh, F. Muheidat, M. Tawalbeh, M. Quwaider, and A. A. Abd El-Latif, “Edge enabled IoT system model for secure healthcare,” *Measurement (Lond.)*, vol. 191, no. 110792, p. 110792, 2022.

[9] B. Costa, J. Bachiega Jr, L. R. de Carvalho, and A. P. F. Araujo, “Orchestration in fog computing: A comprehensive survey,” *ACM Comput. Surv.*, vol. 55, no. 2, pp. 1–34, 2023.

[10] A. Volkov, K. Proshutinskiy, A. B. M. Adam, A. A. Ateya, A. Muthanna, and A. Koucheryavy, “SDN load prediction algorithm based on artificial intelligence,” in Communications in Computer and Information Science, Cham: Springer International Publishing, 2019, pp. 27–40.

[11] K. Kaur, V. Mangat, and K. Kumar, “A review on Virtualized Infrastructure Managers with management and orchestration features in NFV architecture,” *Comput. Netw.*, vol. 217, no. 109281, p. 109281, 2022.

[12] W. Lv et al., “Microservice deployment in edge computing based on deep Q learning,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 11, pp. 1–1, 2022.

Dual-RIS Assisted 3D Positioning and Beamforming Design in ISAC System

Dejie Ma[1], Zhiquan Bai[1*], Jinqiu Zhao[1], Hao Xu[1], Zeyu Liu[2], Di Zhou[1], Mingyan Jiang[1], and KyungSup Kwak[3]
[1]Shandong Provincial Key Lab. of Wireless Communication Technologies,
School of Information Science and Engineering, Shandong University, Qingdao 266237, China
[2]Department of Engineering Construction, China Mobile Inner Mongolia Co., Ltd. Baotou Branch, Baotou 014000, China
[3]Department of Information and Communication Engineering, INHA University, Incheon 22212, Korea
madj0212@163.com, zqbai@sdu.edu.cn*, 202020373@mail.sdu.edu.cn, xhxhn999@163.com, 13500621551@139.com,
emailofzhoudi@163.com, jiangmingyan@sdu.edu.cn, kskwak@inha.ac.kr

Abstract— **Integrated sensing and communication (ISAC) technology as a research focus in 6G communications commonly works in high frequency band, which may suffer severe fading caused by obstacle. Reconfigurable intelligent surface (RIS) can overcome the above issue and improve the performance of ISAC system through phase adjustment. In this paper, dual-RIS assisted 3D positioning and beamforming design in ISAC system are studied. Firstly, the localization in the ISAC system is transformed into a compressed sensing (CS) problem, and a stepwise matching pursuit (SMP) algorithm is proposed for better positioning ability and lower complexity, compared with the typical matching pursuit (MP) algorithm. Then, the positioning information is utilized for the beamforming design of the RISs to maximize the system achievable rate through the alternating optimization algorithm based on the triangle inequality (TI-AO). Simulation results show that the system achievable rate of the optimization design is close to the optimal one and verifies the effectiveness of the proposed framework.**

Keywords— **Integrated sensing and communication (ISAC), reconfigurable intelligent surface (RIS), stepwise matching pursuit (SMP), beamforming design**

I. Introduction

Integrated sensing and communication (ISAC) has become one of the main technologies in 6G communications [1]. By realizing the communication and sensing in the same spectrum, ISAC can avoid the interference and improve the spectrum efficiency. However, the increasing demands for high communication spectrum leads to the increase of costs and energy consumption, which also influences the signal propagation range. The emerging reconfigurable intelligent surface (RIS) technology can properly adjust the transmission environment and control the wireless channel to a certain extent. It is feasible to introduce RIS into the ISAC system and improve the communication quality and coverage [2].

In the case of non-line-of-sight (NLoS) transmission, various obstacles may affect the accurate positioning of the ISAC system. However, the RIS can work as a good solution to this problem by establishing a virtual line of sight (VLoS) path between the transmitter and the receiver, and further achieve reliable communication by adjusting the signal phases [3][4]. At present, there have been many studies on the design of RIS assisted ISAC systems. R. Liu *et. al.* proposed a location method based on the time difference of arrival (TDOA), which realized user positioning in RIS assisted wireless network [5]. The authors in [6] employed RISs to increase the received signal intensity at the adjacent locations, so that multiple users can be located based on the received signal strength (RSS). An iterative positioning algorithm with centimeter-level positioning accuracy was presented in [7] based on the angle of arrival (AOA) and angle of departure (AOD) estimated by the maximum likelihood (ML) algorithm. [8] deduced the analytic position error bound (PEB) and rotation error bound (REB) based on the RIS phase shift functions. Then, to optimize all RISs' phase shifts, the lowest attainable PEB and REB were obtained by the particle swarm optimization (PSO) algorithm. In [9]-[11], it was shown that the passive RIS and semi-passive RIS can be jointly taken into ISAC system to realize the user positioning and share the position information for the beamforming design. The authors in [12] proposed an ISAC system based on a hybrid RIS with active and passive components, and employed the alternating optimization (AO) for better worst-case target illumination power performance.

Currently, AO algorithm is commonly utilized to optimize the transmit beamforming and RIS phase shifts. A hybrid RIS-assisted ISAC system similar to [12] was studied in [13], which can sense the positions of nearby targets by the traditional multiple signal classification (MUSIC) algorithm and complete the optimization by maximizing the total power of average echo signals of the RIS sensors. The AO algorithm based beamforming in the single RIS assisted communication system was presented in [14] to maximize the total received signal power of the users. Meanwhile, [15] further investigated a multi-user communication system assisted by two RISs with collaborative passive beamforming design, where the base station (BS) in the uplink can obtain a larger receive signal-to-noise ratio (SNR) by the AO algorithm. In this paper, we study the 3D positioning and beamforming design in dual-RIS assisted ISAC system, then we transform the localization in the ISAC system into a compressed sensing (CS) problem and propose a stepwise matching pursuit (SMP) algorithm. Based on the positioning information, we complete the beamforming design of the BS and RISs by an alternating optimization algorithm based on the triangle inequality (TI-AO).

The work was supported in part by the Shandong Provincial Natural Science Foundation under Grant ZR2021LZH003, in part by the Regional Innovation and Development Joint Foundation by National Natural Science Foundation of China under Grant U23A20277, in part by the National Key Research and Development Plan under Grant 2022YFF0604900, in part by the National Nature Science Foundation of China under Grant 61771291, and in part by the National Research Foundation of Korea under Grant MSIT-NRF-2020R1A2B5B02002478. *Corresponding author.

II. System Model and Problem Formulation

The dual-RIS assisted ISAC system is presented in Fig. 1, where two RISs, RIS$_1$ and RIS$_2$, each with M passive reflection units, are deployed to assist the uplink transmission from a single antenna user to a BS with N receive antennas.

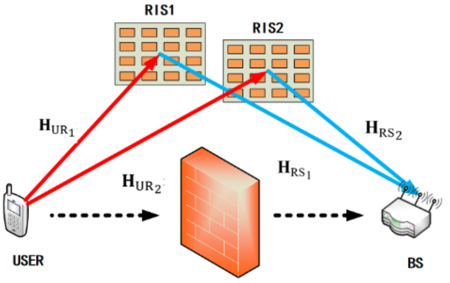

RIS1
RIS2
$\mathbf{H}_{\mathrm{UR}_1}$
$\mathbf{H}_{\mathrm{UR}_2}$
$\mathbf{H}_{\mathrm{RS}_1}$
$\mathbf{H}_{\mathrm{RS}_2}$
USER
BS

Figure 1. Dual-RIS assisted ISAC system

It is assumed that the positions of the BS and RISs are known. $\mathbf{H}_{RS_i}$ denotes the channel matrix between the ith RIS and the BS, and it is written as

$$\mathbf{H}_{RS_i}=\delta_{RS_i}\mathbf{a}(\varphi_{iDR},\theta_{iDR})\otimes\mathbf{b}(\varphi_{iAS})^{\mathrm{T}}, \tag{1}$$

where δ_{RS_i} is the free space path loss as [16], $\mathbf{a}(\varphi_{iDR},\theta_{iDR})$ represents the response of the ith RIS transmit array, and $\mathbf{b}(\varphi_{iAS})$ is the response of receive antennas at the BS. The uniform linear array (ULA) characteristics of the BS antennas and the uniform planar array (UPA) characteristics of the RISs are denoted respectively as

$$\mathbf{b}(\varphi_{iAS})=\mathbf{b}_{isy}(\varphi_{iAS}), \tag{2}$$

and

$$\mathbf{a}(\varphi_{iDR},\theta_{iDR})=\mathbf{a}_{irz}(\theta_{iDR})\otimes\mathbf{a}_{iry}(\varphi_{iDR},\theta_{iDR}), \tag{3}$$

where we have

$$\mathbf{b}_{isy}(\varphi_{iAS})=\left[1,\cdots,e^{j(N-1)\frac{2\pi d\cos\varphi_{iAS}}{\lambda}}\right], \tag{4}$$

$$\mathbf{a}_{iry}(\varphi_{iDR},\theta_{iDR})=\left[1,\cdots,e^{-j\left(\sqrt{M}-1\right)\frac{2\pi d\sin\varphi_{iDR}\cos\theta_{iDR}}{\lambda}}\right], \tag{5}$$

$$\mathbf{a}_{irz}(\theta_{iDR})=\left[1,\cdots,e^{-j\left(\sqrt{M}-1\right)\frac{2\pi d\sin\theta_{iDR}}{\lambda}}\right], \tag{6}$$

$\otimes$ is the Kronecker product, φ_{iAS} represents the signal AOA at the BS, φ_{iDR} and θ_{iDR} are the transmit azimuth angle and the transmit elevation angle at the ith RIS, respectively, d is the distance between the neighboring RIS reflection units, and λ is the signal wavelength.

Similarly, $\mathbf{H}_{UR_i}$ means the channel matrix from the user to the ith RIS and is expressed as

$$\mathbf{H}_{UR_i}=\delta_{UR_i}\mathbf{a}(\varphi_{iAR},\theta_{iAR}), \tag{7}$$

where $\mathbf{a}(\varphi_{iAR},\theta_{iAR})$ is the receive array response of the ith RIS.

According to the above setting, the channel matrix of the user-RIS-BS link can be modelled as

$$\mathbf{G}_i=\mathbf{H}_{RS_i}\mathbf{\Phi}_i(\mathbf{H}_{UR_i})^{\mathrm{H}},\ i=1,2, \tag{8}$$

where $\mathbf{\Phi}_i=diag\left[e^{j\phi_i^1},e^{j\phi_i^2},\cdots,e^{j\phi_i^M}\right]$ is the reflection coefficient matrix of the ith RIS and ϕ_i^m is the corresponding phase of the mth reflection unit.

The signal sent by the user at time l is $x(l)$ with power $p=\mathbb{E}\left[x(l)x^{\mathrm{H}}(l)\right]$. The received signal at the BS from the two RISs is

$$y(l)=\sqrt{p}\boldsymbol{\omega}\left(\mathbf{H}_{RS_1}\mathbf{\Phi}_1\left(\mathbf{H}_{UR_1}\right)^{\mathrm{H}}+\mathbf{H}_{RS_2}\mathbf{\Phi}_2\left(\mathbf{H}_{UR_2}\right)^{\mathrm{H}}\right)x(l)$$
$$+\boldsymbol{\omega}(n_1(l)+n_2(l)), \tag{9}$$

where $\boldsymbol{\omega}\in\mathbb{C}^{1\times N}$ is the receive beamforming vector at the BS, and $n_i(l)\sim\mathcal{CN}(0,\sigma^2)$ represents the complex additive white Gaussian noise (AWGN) with mean 0 and variance σ^2. $(\varphi_{iDR},\theta_{iDR})$ in $\mathbf{H}_{UR_i}$ can be obtained by the received signal, and the user position can be estimated according to the geometric information of the user and RISs. Considering the signal sparsity in the channel estimation, the angle estimation problem can be transformed into a traditional CS problem.

Firstly, the angles of azimuth and elevation are uniformly sampled into discrete values with the size of W and Q, respectively, as

$$\boldsymbol{\varphi}=\left[\varphi_0,\varphi_1,\cdots,\varphi_{W-1}\right]^{\mathrm{T}},\varphi_w\in\left(0,\frac{\pi}{2}\right),w\in\left(0,W-1\right), \tag{10}$$

$$\boldsymbol{\theta}=[\theta_0,\theta_1,\cdots,\theta_{Q-1}]^{\mathrm{T}},\theta_q\in\left(0,\frac{\pi}{2}\right),q\in\left(0,Q-1\right). \tag{11}$$

Based on (10) and (11), the received signal at the BS from the ith RIS can be reformulated as

$$y_i(l)=\sum_{w=0}^{W-1}\sum_{q=0}^{Q-1}\sqrt{p}\boldsymbol{\omega}\mathbf{H}_{RS_i}\mathbf{\Phi}_i\left(\mathbf{H}_{UR_i}\left(\varphi_w,\theta_q\right)\right)^{\mathrm{H}}x(l)\xi_i^{w,q}+\boldsymbol{\omega}n_i(l),$$
$$=\mathbf{\Gamma}_i\boldsymbol{\xi}_i+\boldsymbol{\omega}n_i(l), \tag{12}$$

with

$$\mathbf{\Gamma}_i=\sum_{w=0}^{W-1}\sum_{q=0}^{Q-1}\sqrt{p}\boldsymbol{\omega}\mathbf{H}_{RS_i}\mathbf{\Phi}_i\left(\mathbf{H}_{\mathrm{UR}_i}\left(\varphi_w,\theta_q\right)\right)^{\mathrm{H}}x(l)\in\mathbb{C}^{1\times WQ}, \tag{13}$$

$\boldsymbol{\xi}_i=\left[\xi_i^{0,0},\xi_i^{0,1},\cdots,\xi_i^{W-1,Q-1}\right]^{\mathrm{T}}$, and $\xi_i^{w,q}\in\{0,1\}$. Due to the sparsity of $\boldsymbol{\xi}_i$, the location recovery problem can be transformed into a classical CS problem. Thus, $y_i(l)$ can be viewed as the observed value in the CS problem, and $\mathbf{\Gamma}_i$ is the sensing matrix.

III. Position Awareness and Beamforming Design

A. Construct Perception Matrix

According to the above CS problem, only the user's angle corresponding to each RIS is unknown in the formula. We assume the uniform sampling for $\boldsymbol{\varphi}=\left[\varphi_0,\varphi_1,\cdots,\varphi_{W-1}\right]^{\mathrm{T}}$ as $\varphi_0=0$ and $\varphi_w=\varphi_0+\pi w/2W$. Similarly, we have $\theta_0=0$ and $\theta_q=0+\pi q/2Q$. Thus, $\mathbf{H}_{UR_i}(\varphi_w,\theta_q)$ can be rewritten as

$$\mathbf{H}_{UR_i}(\varphi_w,\theta_q)=\delta_{UR_i}\mathbf{a}_{irz}\left(\theta_q\right)\otimes\mathbf{a}_{iry}\left(\varphi_w,\theta_q\right). \tag{14}$$

Since $\mathbf{H}_{RS_i}$ and $\mathbf{\Phi}_i$ are assumed to be known, we can construct the sensing matrix $\mathbf{\Gamma}_i$.

B. Stepwise Matching Pursuit

Due to the existence of two unknown variables φ and θ in the sensing matrix $\mathbf{\Gamma}_i$, the computational complexity for the traditional CS algorithm is high, and the number of matching calculations is $W\times Q$. Meanwhile, different azimuth and elevation angles may result in the same results and affect the positioning performance. Therefore, in this subsection, we take the SMP algorithm to realize the positioning and assume that the azimuth angle φ_{iDR} is $\pi/2$.

To reduce the effect of channel noise on the positioning perception, we first set $\mathbf{\Phi}_i = diag(\boldsymbol{\omega}\mathbf{a}(\varphi_{iDR}, \theta_{iDR})^{\mathbf{H}})$ and let the user send the identical signals in two adjacent time slots. After the completion of the first signal transmission, we have

$$\mathbf{\Phi}_i = -\mathbf{\Phi}_i, \tag{15}$$

$$\mathbf{\Phi}_i(k,k) = -1 \times \mathbf{\Phi}_i(k,k),\ k = 1:\sqrt{M}:M. \tag{16}$$

Then, the two received signals are superimposed, and the same operation is performed on the sensing matrix. Because of the change of $\mathbf{\Phi}_i$, the received signals from the *i*th RIS can be regarded as a column vector. Thus, $\mathbf{H}_{UR_i}(\pi/2, \theta_q)$ is equivalent to

$$\mathbf{H}_{UR_i}(\pi/2, \theta_q) = \delta_{UR_i}\left[1, \cdots, 0, \cdots, \mathbf{e}^{-j(\sqrt{M}-1)\pi\sin\theta_q}, \cdots, 0\right]. \tag{17}$$

It can be seen that $\mathbf{H}_{UR_i}(\pi/2, \theta_q)$ only depends on $\sin\theta_q$. We can take MP algorithm to obtain more accurate θ_{iAR} .

According to the above analysis, the available elevation angle θ_{iAR} can be substituted into $\mathbf{H}_{UR_i}\left(\varphi_w, \theta_q\right)$, and the azimuth angle can be further obtained by the MP algorithm. The specific procedure of the positioning algorithm is shown in Table I.

TABLE I. SMP ALGORITHM

```
Algorithm 1: SMP
Input: $y_i, \mathbf{\Gamma}_i$
Result: $\xi_i$
Initialization: $r_{\mathbf{0}} = y_i, \Lambda_{i0} = \varnothing, \Gamma_{i0} = \varnothing, t = 1$;
  Step 1: $\mathbf{\Phi}_i = diag(\boldsymbol{\omega}\mathbf{a}(\varphi_{iDR}, \theta_{iDR})^{\mathbf{H}})$ , calculate $y_i$ and $\mathbf{\Gamma}_i$ according to (13) and (14);
  Step 2: Transform $\mathbf{\Phi}_i$ as shown in (16) and (17), then calculate $y_i^{'}$ and $\mathbf{\Gamma}_i^{'}$ according to (13) and (14), so that $r_0 = y_i + y_i^{'}$ and $\mathbf{\Gamma}_i = \mathbf{\Gamma}_i + \mathbf{\Gamma}_i^{'}$ ;
  Step 3: Execute matching pursuit algorithm
    for $t<C\&\&r_0>0$ do
        a) $\mathbf{J} = \arg\min_{j \notin \mathbf{\Lambda}_{it-1}} |\zeta_i^j - r_{t-1}|$ ( for all $\boldsymbol{\zeta}_\mathbf{i}^\mathbf{j} \in \mathbf{\Gamma} i$ ) ;
        b) $\mathbf{\Lambda}_{it} = \mathbf{\Lambda}_{it-1} \cup \mathbf{J}, \mathbf{\Gamma}_{it} = \mathbf{\Gamma}_{it-1} \cup \zeta_i^j$ ( for all $\boldsymbol{\zeta}_\mathbf{i}^\mathbf{j} \in \mathbf{\Gamma} i$ ) ;
        c) $\xi_{it} = \arg\min_{\xi_{it}} \| y_i - \mathbf{\Gamma}_{it}\xi_{it} \| = \left(\mathbf{\Gamma}_{it}^{\mathrm{T}}\mathbf{\Gamma}_{it}\right)^{-1}\mathbf{\Gamma}_{it}^{\mathrm{T}} y_i$ ;
        d) $r_t = y_i - \mathbf{\Gamma}_{it}\xi_{it} = y_i - \mathbf{\Gamma}_{it}\left(\mathbf{\Gamma}_{it}^{\mathrm{T}}\mathbf{\Gamma}_{it}\right)^{-1}\mathbf{\Gamma}_{it}^{\mathrm{T}} y_i$ ;
      $t = t+1$
    end
  Step 4: Set the maximum $K$ items of $\xi_{it}$ obtained in Step 3 to be 1, recover the elevation $\theta_{iAR}$ by $\xi_{it}$ and substitute it into $\mathbf{H}_{UR_i}(\varphi, \theta)$ ;
  Step 5: Reset $r_{\mathbf{0}} = y_i, \Lambda_{i0} = \varnothing, \Gamma_{i0} = \varnothing, t = 1$;
  Step6: Set $\mathbf{\Phi}_i = diag(\boldsymbol{\omega}\mathbf{a}(\varphi_{iDR}, \theta_{iDR})^{\mathbf{H}})$ , and calculate $y_i$ and $\mathbf{\Gamma}_i$ according to (13) and (14);
  Step 7: Repeat the matching pursuit algorithm;
  Step 8: Set the maximum $K$ terms of $\xi_{it}$ obtained in Step 7 to be 1, and recover the azimuth angle $\varphi_{iAR}$ .
```

The angle relationship between the user and the RISs can be obtained by the SMP algorithm, and then the user's location can be reconstructed according to the geometric relationship. After having the above angles and positions, we can get the estimated RIS array response, and thus obtain the phase offset information of the channel. Therefore, the received signal can be rewritten as

$$y_i(l) = \sqrt{p}\boldsymbol{\omega}\mathbf{H}_{RS_i}\mathbf{\Phi}_i(\mathbf{H}_{UR_i}(\hat{\varphi}_{iAR}, \hat{\theta}_{iAR}))^{\mathrm{H}} x(l) + \boldsymbol{\omega} n_i(l). \tag{18}$$

C. Beamforming Design

In this work, with the known channel state information (CSI), we aim to maximize the system achievable rate at the BS through the design of the receive beamforming of the BS antennas and the phase shift matrix of the RISs. The above optimization problem can be decomposed into two sub-problems, one is the phase shift design of the two RISs and the other is the receive beamforming design of the BS. We can first formulate the independent optimization problems by fixing two of these variables. Then, the optimal solution of each sub-problem can be obtained by decoupling. According to (18), the system achievable rate at the BS is computed as

$$R = \log_2(1+\gamma), \tag{19}$$

with

$$\gamma = \frac{\left\| \sqrt{p}\boldsymbol{\omega}(\mathbf{H}_{RS_1}\mathbf{\Phi}_1(\mathbf{H}_{UR_1}(\varphi_{1AR}, \theta_{1AR}))^{\mathbf{H}} + \mathbf{H}_{RS_2}\mathbf{\Phi}_2(\mathbf{H}_{UR_2}(\varphi_{2AR}, \theta_{2AR}))^{\mathbf{H}}) \right\|^2}{2\sigma^2}. \tag{20}$$

where $\|\cdot\|$ is the two-norm.

The optimization problem of the RIS beamforming design can be described as

$$\begin{aligned} \text{(P1)} \quad & \max \quad R \\ & \text{s.t.} \quad \left|\phi_i^n\right| = 1, \forall i \in \{1,2\}, n \in \{1, 2, \cdots, N\}. \end{aligned} \tag{21}$$

The above problem is not convex due to the constant modulus constraint of the phase shift matrix $\left|\phi_i^n\right| = 1$ and the coupling of $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$. Considering that $\mathbf{\Phi}_1$, $\mathbf{\Phi}_2$, and $\boldsymbol{\omega}$ in (P1) should be optimized simultaneously, it is impossible to find the optimal solution directly. Therefore, as an approximate optimal solution, AO algorithm is considered. Firstly, we can transform (P1) into a sub-problem about $\boldsymbol{\omega}$ and find the optimal solution. Then, the sub-problems of $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ can be solved in turn, which can ensure the quality of the received signal of each time slot at the BS.

At the BS, in order to maximize the receive SNR, we can take the maximal ratio combining (MRC) method. With fixed $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$, the optimal $\boldsymbol{\omega}$ can be obtained as

$$\boldsymbol{\omega}^* = \frac{\sqrt{p}(\mathbf{H}_{RS_1}\mathbf{\Phi}_1(\mathbf{H}_{UR_1}(\hat{\varphi}_{1AR}, \hat{\theta}_{1AR}))^{\mathrm{H}} + \mathbf{H}_{RS_2}\mathbf{\Phi}_2(\mathbf{H}_{UR_2}(\hat{\varphi}_{2AR}, \hat{\theta}_{2AR}))^{\mathrm{H}})}{\left\|\sqrt{p}(\mathbf{H}_{RS_1}\mathbf{\Phi}_1(\mathbf{H}_{UR_1}(\hat{\varphi}_{1AR}, \hat{\theta}_{1AR}))^{\mathrm{H}} + \mathbf{H}_{RS_2}\mathbf{\Phi}_2(\mathbf{H}_{UR_2}(\hat{\varphi}_{2AR}, \hat{\theta}_{2AR}))^{\mathrm{H}})\right\|}. \tag{22}$$

Since it is difficult to obtain the optimal solution of the non-convex optimization problem (P1), we can fix the phase shift matrix of one RIS, while optimizing the phase shift matrix of the other RIS. Specifically, for given $\boldsymbol{\omega}$ and $\mathbf{\Phi}_2$, we have

$$\begin{aligned} &\left|\sqrt{p}\boldsymbol{\omega}(\mathbf{H}_{RS_1}\mathbf{\Phi}_1(\mathbf{H}_{UR_1}(\varphi_{1AR}, \theta_{1AR}))^{\mathrm{H}} + \mathbf{H}_{RS_2}\mathbf{\Phi}_2(\mathbf{H}_{UR_2}(\varphi_{2AR}, \theta_{2AR}))^{\mathrm{H}})\right| \\ &\quad \le \left|\sqrt{p}\boldsymbol{\omega}\mathbf{H}_{RS_1}\mathbf{\Phi}_1(\mathbf{H}_{UR_1}(\hat{\varphi}_{1AR}, \hat{\theta}_{1AR}))^{\mathrm{H}}\right| + A = \left|\mathbf{B}\mathbf{\Psi}_1\right| + A, \end{aligned} \tag{23}$$

with

$$A = \sqrt{p}\boldsymbol{\omega}\mathbf{H}_{RS_2}\mathbf{\Phi}_2(\mathbf{H}_{UR_2}(\varphi_{2AR}, \theta_{2AR}))^{\mathrm{H}}, \tag{24}$$

and

$$\mathbf{B} = \sqrt{p}\boldsymbol{\omega}\mathbf{H}_{RS_1} diag(\mathbf{H}_{UR_1}(\hat{\varphi}_{1AR}, \hat{\theta}_{1AR}))^{\mathrm{H}}. \tag{25}$$

If and only if $\angle\mathbf{B}\boldsymbol{\Psi}_1=\angle A$ in (23) is valid, the optimal solution is obtained as

$$\boldsymbol{\Psi}_1^*=e^{j(\angle A-\angle\mathbf{B})}. \tag{26}$$

Finally, the optimal solution $\boldsymbol{\Psi}_1^*$ is converted into a diagonal matrix $\boldsymbol{\Phi}_1^*=diag(\boldsymbol{\Psi}_1^*)$ for the convenience of subsequent iterations, and we can get $\boldsymbol{\Phi}_1$ correspondingly.

Similar to the solution of $\boldsymbol{\Phi}_1$, for given $\boldsymbol{\omega}$ and $\boldsymbol{\Phi}_1$, we can set $D=\sqrt{p}\boldsymbol{\omega}\mathbf{H}_{RS_1}\boldsymbol{\Phi}_1(\mathbf{H}_{UR_1}(\varphi_{1AR},\theta_{1AR}))^{\mathrm{H}}$ and $\mathbf{E}=\sqrt{p}\boldsymbol{\omega}\mathbf{H}_{RS_2}diag(\mathbf{H}_{UR_2}(\varphi_{2AR},\theta_{2AR}))^{\mathrm{H}}$, and get the optimal solution of $\boldsymbol{\Psi}_2$ as

$$\boldsymbol{\Psi}_2^*=e^{j(\angle D-\angle\mathbf{E})}. \tag{27}$$

Then, $\boldsymbol{\Psi}_2^*$ can be further transferred into a diagonal matrix for the convenience of subsequent iterations.

TABLE II. AO ALGORITHM BASED ON TRIANGLE INEQUALITY

```
Algorithm 2: TI-AO
Input: $\mathbf{H}_{RS_1},\mathbf{H}_{UR_1}(\hat{\varphi}_{1AR},\hat{\theta}_{1AR}),\mathbf{H}_{RS_2},\mathbf{H}_{UR_2}(\hat{\varphi}_{2AR},\hat{\theta}_{2AR})$
Result: $\boldsymbol{\omega}^*,\boldsymbol{\Phi}_1^*,\boldsymbol{\Phi}_2^*$
Initialization: $\boldsymbol{\Phi}_1^{(0)},\boldsymbol{\Phi}_2^{(0)}, p=100, k=0, \gamma(0)=\infty$
while $|\gamma(k+1)-\gamma(k)| \&\& k<100$ do
    Step 1: Calculate $\boldsymbol{\omega}^*$ by (22);
    Step 2: Calculate $\boldsymbol{\Psi}_1^*$ by (26);
    Step 3: Calculate $\boldsymbol{\Psi}_2^*$ by (27);
    Step 4: Calculate $\gamma(k+1)$ from $\boldsymbol{\omega}^*$, $\boldsymbol{\Psi}_1^*$, and $\boldsymbol{\Psi}_2^*$;
    $k=k+1$.
end
```

IV. NUMERICAL RESULTS

In this section, we provide the simulation results of the proposed dual-RIS assisted ISAC system. The system settings in the simulation are as follows. The user's 3D location is $\mathbf{u}$=(7, 6, 1)m. Considering that the positioning accuracy is low when the heights of the two RISs are close [5], the coordinates of the two RISs are set as $\mathbf{r}_1$=(0, 2, 3)m and $\mathbf{r}_2$= (0, 3, 1)m, and the coordinate of the BS is set to be $\mathbf{s}$=(1, 1, 1)m. We first set the number of RIS reflection units M=36, the number of BS antennas N=8, and the path loss coefficient g=2.2 with SNR=50dB. In the positioning stage, the sampling numbers W and Q for the azimuth and elevation angles are both set to be 1000.

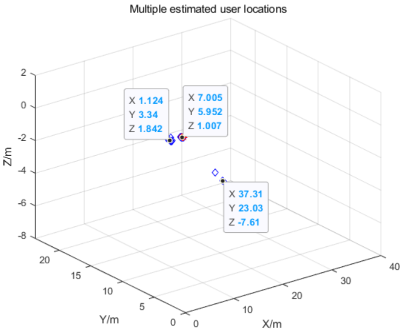


Figure 2. Location performance of SMP and MP Algorithms

As shown in Fig. 2, the red star-shaped dot set is the positioning result of the SMP algorithm, and the blue diamond-shaped one is the result of the traditional MP algorithm. It is seen that the SMP algorithm achieves better positioning performance compared with the MP algorithm. The reason is that different combination values of $(\varphi_{iAR},\theta_{iAR})$ may get relatively similar results for the MP algorithm due to the two-dimensional CS and the positioning error may increase for the influence of noise. The SMP algorithm proposed in this paper can deal with this issue well and it can greatly reduce the computational complexity. The computational complexity of the MP algorithm is $\mathcal{O}(W\times Q)$, while the SMP algorithm has lower complexity as $\mathcal{O}(W+Q)$.

We further study the influence of the number of BS receive antennas on the mean square error (MSE) of positioning with different SNRs. The length of time is L=100. We set SNR=50dB, 60dB, and 70dB, respectively, and the number of BS antennas N=8 and 16, respectively. As seen from Fig. 3, the increasing SNR can significantly improve the positioning performance. That is, the positioning MSE decreases significantly with the increase of SNR since the influence of noise on the positioning error becomes marginal with the increase of SNR. If the SNR increases continuously, the MSE will become very small but not zero even with increasing reflection units of the RISs, which is due to the limitation of the uniform sampling number of angles. When the user is not at the grid points decided by the perception matrix, certain errors will happen and increasing the sampling number can further solve this problem. It is also shown that, when the number of BS receive antennas and the number of RIS reflection units increase, the MSE gradually decreases and then becomes marginal, indicating that the MSE can be affected by both of them, but with limited influence at large number. When the number of RIS units reaches 64, the subsequent MSE changes very little and becomes stable.

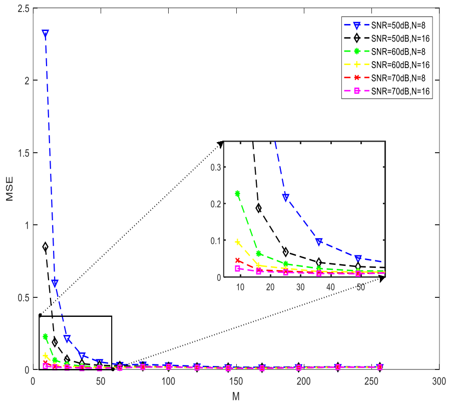


Figure 3. Effect of the receive antenna number N on MSE with different SNRs

The effect of angle division on the MSE of positioning is depicted in Fig. 4. The accuracy of CS depends largely on the meshing of channels in the MP algorithm. When the uniform sampling number is small and the user is not on the grid point, a large positioning error may appear. As illustrated in Figure 4, when the number of uniform samples increases, the MSE gradually decreases and the volatility of MSE also decreases. When the number of uniform samples increases indefinitely, the MSE will tend to be 0. If the user is located on the grid, the MSE also becomes 0. The MSE tends to be stable when the number of RIS reflection units exceeds 64.

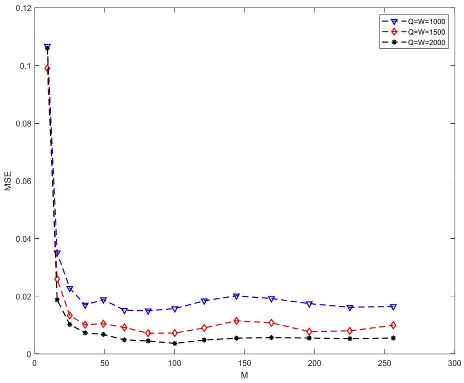


Figure 4. Effect of sampling number *W* and *Q* on MSE

In the following, we study the TI-AO algorithm based beamforming design of the BS antennas and the RIS units according to the obtained positioning information. The system achievable rate with the proposed beamforming design is shown in Fig. 5. We set the uniform sampling number as 2000, the number of BS antennas as 8 and 16, and the number of RIS units as 36 and 64, respectively. The random phase shifts of RIS are also considered for comparison. It is shown that the increase of the SNR, the number of BS antennas, and the number of RIS units, will improve the system achievable rate. At the same time, it is found that, when the number of RIS units is small, the system achievable rate does not increase linearly with the increased SNR due to the large positioning MSE at the first stage shown in Fig. 3. However, when the number of RIS units is 64, the achievable system rate increases linearly with the SNR.

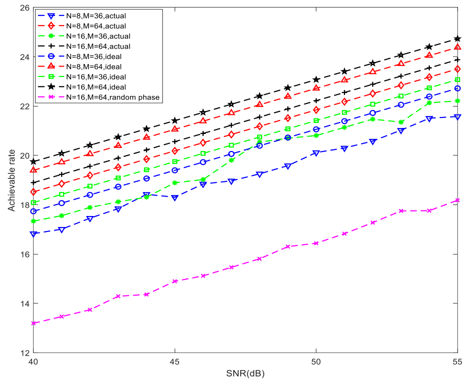


Figure 5. Achievable rate of the proposed beamforming design

V. CONCLUTION

In this paper, a dual-RIS assisted ISAC system is proposed to achieve high-precision accurate positioning of users in the uplink transmission and the optimal beamforming design of the BS and RISs is presented based on the positioning information. The proposed ISAC system realizes the sensing and communication under the same frequency and time resources. It is verified through simulation that the presented SMP algorithm has excellent positioning performance and the near-optimal system achievable rate can be achieved with the optimal beamforming design of the BS and RISs.

REFERENCES

[1] F. Liu et al., "Integrated Sensing and Communications: Toward Dual-Functional Wireless Networks for 6G and Beyond," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 1728-1767, June. 2022.

[2] Q. N. Le, V. -D. Nguyen, O. A. Dobre and H. Shin, "RIS-assisted Full-Duplex Integrated Sensing and Communication," *IEEE Wireless Commun. Lett.*, July. 2023.

[3] Q. Wu and R. Zhang, "Towards Smart and Reconfigurable Environment: Intelligent Reflecting Surface Aided Wireless Network," *IEEE Commun. Mag.* vol. 58, no. 1, pp. 106-112, January. 2020.

[4] X. Ma, S. Guo, H. Zhang, Y. Fang and D. Yuan, "Joint Beamforming and Reflecting Design in Reconfigurable Intelligent Surface-Aided Multi-User Communication Systems," *IEEE Trans. Wirel.Commun.*, vol. 20, no. 5, pp. 3269-3283, May. 2021.

[5] R. Liu, M. Jian and W. Zhang, "A TDoA based Positioning Method for Wireless Networks assisted by Passive RIS," in *Proc. IEEE Globecom*, 2022, pp. 1531-1536.

[6] H. Zhang, H. Zhang, B. Di, K. Bian, Z. Han and L. Song, "MetaLocalization: Reconfigurable Intelligent Surface Aided Multi-User Wireless Indoor Localization," *IEEE Trans. Wirel.Commun.*, vol. 20, no. 12, pp. 7743-7757, Dec. 2021.

[7] W. Wang and W. Zhang, "Joint Beam Training and Positioning for Intelligent Reflecting Surfaces Assisted Millimeter Wave Communications," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 10, pp. 6282-6297, Oct. 2021.

[8] S. Hong, M. Li, X. Li and P. Xu, "The PEB and OEB Analysis of mmWave Positioning System Aided by Multiple RISs," in *Proc. IEEE ICCC*, 2022, pp. 464-468.

[9] C. Liu, X. Hu, M. Peng and C. Zhong, "Sensing for Beamforming: An IRS-Enabled Integrated Sensing and Communication Framework," in *Proc. IEEE ICC*, 2022, pp. 5567-5572.

[10] X. Hu, C. Liu, M. Peng and C. Zhong, "IRS-Based Integrated Location Sensing and Communication for mmWave SIMO Systems," *IEEE Trans. Wirel.Commun.*, vol. 22, no. 6, pp. 4132-4145, June. 2023.

[11] Z. Yu, X. Hu, C. Liu, M. Peng and C. Zhong, "Location Sensing and Beamforming Design for IRS-Enabled Multi-User ISAC Systems," *IEEE Trans. Signal Process.*, vol. 70, pp. 5178-5193, 2022.

[12] R. S. P. Sankar and S. P. Chepuri, "Beamforming in Hybrid RIS assisted Integrated Sensing and Communication Systems," in *Proc. IEEE ESPC*, 2022, pp. 1082-1086.

[13] X. Shao, C. You, W. Ma, X. Chen and R. Zhang, "Target Sensing With Intelligent Reflecting Surface: Architecture and Performance," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 7, pp. 2070-2084, July 2022.

[14] Q. Wu and R. Zhang. Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming[J]. *IEEE Trans. Wirel.Commun.*, vol. 18, no. 11, pp. 5394-5409, Nov. 2019.

[15] Zheng B ,You C ,Zhang R . Double-IRS Assisted Multi-User MIMO: Cooperative Passive Beamforming Design[J]. *IEEE Trans. Wirel.Commun.*, 2021, PP(99):1-1.

[16] J. He, H. Wymeersch, T. Sanguanpuak, O. Silven and M. Juntti, "Adaptive Beamforming Design for mmWave RIS-Aided Joint Localization and Communication," in *Proc. IEEE WCNC*, 2020, pp. 1-6.

Dejie Ma is currently pursuing the M.S. degree in Electronic Information at the School of Information Science and Engineering, Shandong University, Qingdao, China. His research interests include reconfigurable intelligent surface, integrated sensing and communication and signal processing.

Zhiquan Bai received the M.Eng. degree in communication and information system from Shandong University, Jinan, China, in 2003, and the Ph.D. degree (Hons.) from INHA University, Incheon, South Korea, in 2007, under the Grant of Korean Government IT Scholarship. He held a postdoctoral position with INHA University, and was a Visiting Professor with The University of British Columbia, Canada. He is currently a Professor with the School of Information Science and Engineering, Shandong University. His research interests include cooperative technology and spatial modulation, orthogonal time frequency space modulation, MIMO technology, resource allocation and optimization, and deep-learning based 5G wireless communications. He is a member of the editorial board of Journal of Systems Engineering and

Electronics and also an associate editor of the International Journal of Communication Systems.

Jinqiu Zhao received B.E. degree from Shandong Normal University, Jinan, China, in 2020. She is currently pursuing her Ph.D. degree in the School of Information Science and Engineering, Shandong University, Qingdao, China. Her main research interests include reconfigurable intelligent surface and machine learning.

Hao Xu was born in Heze, Shandong Province, China in Dec 2001. He studied at Shandong Agricultural University from 2018 to 2022 and obtained a bachelor's degree in communication engineering. Now he is studying for a master's degree in electronic information engineering at Shandong University. His specific research fields include optimal design on orthogonal time frequency space modulation and signal detection based on nonlinear equalization.

Zeyu Liu received B.E. degree from Inner Mongolia University , Huhehaote, China, in 2000. He is currently an Engineer in China Mobile , Baotou, China. His main research interests include Mobile Communication and Transmission Network Technology.

Di Zhou is pursuing her Ph.D degree in Electronic information from the School of Information Science and Engineering, Shandong University, Qingdao, China. She graduated with a M.S. degree in Telecommunications Engineering from the University of Sydney, Sydney, Australia in 2022. Her research interests include wireless network, Intelligent reflective surfaces, image processing and deep learning techique.

Mingyan Jiang received a B.E. degree in Radio Electronics from the Department of Radio Electronics of Shandong University in 1987, a Master of Science degree in Intelligent Measurement and Control from the Department of Electronic Engineering of Shandong University in 1992, a Doctor of Science degree in Communication and Information Systems Engineering from the School of Information Science and Engineering of Shandong University in 2005, Spain in 2007 (CTTC) Communication signals and systems outbound postdoc.

Kyung Sup Kwak received his BS degree from the Inha University, Inchon, Korea,in 1977 and his MS degree from the University of Southern California in 1981and his PhD degree from the University of California at San Diego in 1988, under the Inha University Fellowship and the Korea Electric Association Abroad Scholarship Grants, respectively.From 1988 to 1989, he was with Hughes Network Systems, San Diego, California. From 1989 to 1990, he was with the IBM Network Analysis Center, North Carolina. Since then, he has been with the School of Information and Communication Engineering, Inha University, Korea, as a professor. He is the director of UWB Wireless Communications Research Center (UWB-ITRC).Since 1994, he served as a member of the board of directors and the vice president and the president of Korean Institute of Communication Sciences (KICS) in 2006 and the president of Korea Institute of Intelligent Transport Systems (KITS) in 2009. He received many research awards, such as the award of research achievements in UWB radio from the Ministry of Information and Communication and Prime Ministry of Korea in 2005 and 2006, respectively. In 2008, he is elected as Inha Fellow Professor (IFP). In 2010, he received the Korean President official commendation for his contribution to ICT innovation and industrial promotion.He published more than 100 SCI journal papers, 300 conference/domestic papers, obtained 20 registered patents and 35 pending patents, and proposed 21 technical proposals on IEEE 802.15 (WPAN) PHY/MAC. He is one of the members of the IEEE, IEICE, KICS, and KIEE. His research interests include multiple access communication systems, cognitive radio, UWB radio systems and WBAN, WPAN, and sensor networks.

Deep Reinforcement Learning Based Beamforming in RIS-assisted MIMO System Under Hardware Loss

Yuan Sun[1], Zhiquan Bai[1*], Jinqiu Zhao[1], Dejie Ma[1], Zhaoxia Xian[1], and KyungSup Kwak[2]
[1]Shandong Provincial Key Lab. of Wireless Communication Technologies,
School of Information Science and Engineering, Shandong University, Qingdao, Shandong, China
[2]Department of Information and Communication Engineering, INHA University, Incheon 22212, Korea
202212666@mail.sdu.edu.cn, zqbai@sdu.edu.cn*, 202020373@mail.sdu.edu.cn,
madj0212@163.com, xianzhaoxia2000@163.com, kskwak@inha.ac.kr

***Abstract*—Reconfigurable intelligent surface (RIS) is considered as one of the key enabling technologies for future 6G wireless communication by realizing an intelligent radio environment. RIS is used as reflective array to change the transmission and coverage of radio frequency (RF) signals. In this paper, we propose a deep reinforcement learning (DRL) based RIS beamforming design in practical scenarios where RIS may have hardware loss, and the soft actor-critic (SAC)-exploration algorithm is presented to solve the beamforming design. The algorithm reduces the prediction error by introducing a perturbation signal to influence the action prediction. Simulation results show that our proposed SAC-exploration algorithm has significant improvement over the typical SAC algorithm, which verifies the effectiveness of the proposed algorithm.**

***Index Terms*—Reconfigurable intelligent surfaces (RIS), radio frequency, multiple input multiple output (MIMO), soft actor-critic (SAC), time division duplex (TDD).**

I. Introduction

Reconfigurable intelligent surface (RIS) is an emerging technology that is widely recognized as a foundational technology for the forthcoming generation of wireless communications [1]. It comprises multiple reflectors with adjustable impedances to achieve the desired phase shifts of the reflected wave. Consequently, RIS improves radio frequency (RF) signal transmission performance, reduces signal interference, extends signal coverage, and thus improves communication quality.

Recently, the research of RIS has focused on how to perform beamforming design of RIS to achieve the optimal system performence. In [2], passive beamforming is designed for the RIS to maximize the sum rate of the RIS-assisted cellular network considering the frequency selectivity characteristics of the RIS. The signal-to-noise ratios (SNRs) of the target detections are weighted and maximized in [3] by jointly optimizing the transmit beamforming and the RIS reflection coefficients while satisfying the quality of service (QoS) requirements of communication.

Considering the increasingly intricate radio environment, traditional optimization algorithms have become impractical for their complexities. However, with the continuous advancements in hardware design, computers now possess significantly enhanced computing capability. Consequently, a new alternative algorithm called reinforcement learning (RL) has emerged as a main method for machine learning (ML) based RIS-assisted communication. It gets applications in various areas, such as computing offloading of edge devices and fog computing [4], communication and trajectory planning within the internet of vehicles domain [5], system capacity enhancement in unmanned aerial vehicle (UAV)-assisted nonorthogonal multiple access (NOMA) network [6], and the optimization of beamforming and the phase design of RIS for efficient transmission [7] [8]. The phase design of RIS in previous scenarios has been idealized, and there is a lack of consideration for the hardware loss of the RIS.

In this paper, a deep reinforcement learning (DRL) approach is used for the RIS beamforming design in the RIS-assisted multiple input multiple output (MIMO) communication to maximize the system sum rate in the downlink transmission. The main contributions of this paper are summarized as follows. (1) We consider the hardware loss of RIS and propose a phase shift design for RIS-assisted MIMO communication. (2) We design a DRL based RIS beamforming algorithm for the above scheme, by introducing interference signals and interfering with action selection to prevent from falling into local optimal solutions, and the simulation results show that the proposed algorithm has a significant improvement over the typical algorithm.

II. System Model and Problem Formulation

A. System Model

For the RIS-assisted wireless communication system, the system diagram is shown in Figure 1, with k single antenna users and one RIS with L elements, the channel model

The work was supported in part by the Shandong Provincial Natural Science Foundation under Grant ZR2021LZH003, in part by the Regional Innovation and Development Joint Foundation by National Natural Science Foundation of China under Grant U23A20277, in part by the National Defense Preliminary Research Project under Grant 90903030202, in part by the National Key Research and Development Plan under Grant 2022YFF0604900, in part by the National Nature Science Foundation of China under Grant 61771291, and in part by the National Research Foundation of Korea under Grant MSIT-NRF-2020R1A2B5B02002478. *Corresponding author.

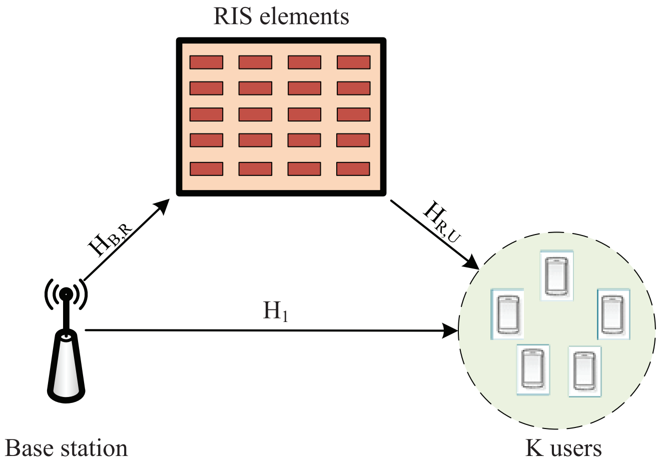


Fig. 1. RIS-assisted MIMO System

between the base station (BS) and the user k can be expressed as

$$\mathbf{h}_k = \mathbf{h}_{bs,k} + \sum_{l=1}^{L} \mathbf{h}_{bs,l}\phi_l h_{l,k}, \tag{1}$$

where $\mathbf{h}_{bs,k} \in \mathbb{C}^{N_B \times 1}$, $\mathbf{h}_{bs,l} \in \mathbb{C}^{N_B \times 1}$, and $h_{l,k}$ respectively denote the channel vector from the BS to the user k, from the BS to the RIS, and from the RIS to the user k. Considering the conditions of hardware loss, the phase-independent amplitude model of RIS [9] can be obtained as

$$\Phi \triangleq diag(\phi_1, ..., \phi_L) \in \mathbb{C}^{L \times L}, \tag{2}$$

where $\phi_l = \zeta\left(\varphi_l\right) e^{j\varphi_l}$ is phase-independent amplitude model for each RIS, the phase-shift is $\varphi_l \in [0, 2\pi)$, and the amplitude satisfies

$$\zeta\left(\varphi_l\right) = \left(1 - \zeta_{\min}\right)\left(\frac{\sin\left(\varphi_l - \mu\right) + 1}{2}\right)^{\kappa} + \zeta_{\min}, \tag{3}$$

where $\zeta_{min} \in [0, 1]$, $\mu \geq 0$, and $\kappa \geq 0$ are constants depend on the hardware configuration of RIS.

The single-user channel is aggregated to obtain the multi-user channel, which is denoted as

$$\mathbf{H} = \mathbf{H}_1 + \mathbf{H}_{B,R}\Phi\mathbf{H}_{R,U}, \tag{4}$$

with

$$\mathbf{H}_1 = \left[\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K\right],$$

$$\mathbf{H}_{B,U} = \left[\mathbf{h}_{bs,1}, \mathbf{h}_{bs,2}, \ldots, \mathbf{h}_{bs,K}\right],$$

and

$$\mathbf{H}_{R,U} = \begin{bmatrix} h_{1,1}, h_{1,2}, \ldots, h_{1,K} \\ h_{2,1}, h_{2,2}, \ldots, h_{2,K} \\ \vdots \quad \cdots \quad \cdots \quad \vdots \\ h_{L,1}, h_{L,2}, \ldots, h_{L,K} \end{bmatrix}.$$

B. Problem Formulation

The time division duplex (TDD) technique is utilized to implement the wireless transmissions in order to validate the advantages of the proposed algorithm and ensure its applicability and generality [10].

In the uplink stage, the BS simultaneously receives the pilot signals transmitted by the k users. The pilot signal vector received by the BS is expressed as

$$\mathbf{Y}_K = \mathbf{HP} + \mathbf{N}, \tag{5}$$

where $\mathbf{P} \in \mathbb{C}^{K \times K}$ is the pilot pattern.

The BS uses minimum mean square error (MMSE) to estimate the channel matrix of the received pilot signals, that is

$$\hat{\mathbf{H}} = \mathbf{Y}_U\mathbf{P}^H\left(\mathbf{PP}^H + \sigma_U^2\mathbf{I}\right)^{-1}. \tag{6}$$

When $\mathbf{P}$ is a unitary matrix, the above estimated channel matrix can be transformed.

$$\hat{\mathbf{H}} = \frac{\mathbf{Y}_U\mathbf{P}^H}{1 + \sigma_U^2}. \tag{7}$$

In the downlink transmission, data transmission with zero-forcing (ZF) precoding is performed with the following expression

$$\mathbf{C} = \left[\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\right]^H = \mathbf{P}_{norm}\left(\hat{\mathbf{H}}^H\hat{\mathbf{H}}\right)^{-1}\hat{\mathbf{H}}^H, \tag{8}$$

where $\mathbf{P}_{norm}$ is the normalized power matrix and written as

$$\mathbf{P}_{norm} = \mathrm{diag}\left(\left[\frac{1}{\|\mathbf{c}_1\|_2}, \frac{1}{\|\mathbf{c}_2\|_2}, \ldots, \frac{1}{\|\mathbf{c}_K\|_2}\right]\right). \tag{9}$$

The signal received by the k-th user is

$$y_{P,k} = \mathbf{c}_k^H\mathbf{h}_k x_k + \sum_{z \neq k}^{K} \mathbf{c}_z^H\mathbf{h}_k x_z + n_k, \tag{10}$$

where x_k is the signal sent to the user k, n_k is additive gaussian white noise. For the RIS-assisted wireless communication MIMO system, we will use the sum downlink rate as an indicator to measure the performance of the system. Thus, the problem can be expressed as

$$\begin{aligned} \max_{\boldsymbol{\phi}} \quad & P_m \\ \text{s.t.} \quad & \varphi \in [0, 2\pi) \\ & \mu \geq 0 \\ & \kappa \geq 0 \end{aligned}, \tag{11}$$

where $P_m = \sum_{k=1}^{K} r_k = \sum_{k=1}^{K} log_2(1 + SINR_k)$, the signal to interference plus noise ratio (SINR) is given as

$$SINR_k = \frac{|\mathbf{c}_k^H\mathbf{h}_k|}{\sum_{z \neq k}^{K} |\mathbf{c}_z^H\mathbf{h}_k|^2 + \sigma_k^2}. \tag{12}$$

III. Deep Reinforcement Learning Algorithms

In this section, we propose the deep reinforcement learning framework. We will introduce the principle of soft actor-critic (SAC) algorithm first, and then we present the improved SAC algorithm with the environment parameters.

A. Soft Actor-Critic Algorithm

In contrast to typical DRL algorithms, we use the advanced SAC algorithm [11]. The training of the algorithm involves a trade-off between the maximized expected returns and the maximized entropy. Unlike *on-policy* algorithms, SAC is an off*-policy* algorithm that requires an experience replay buffer for caching experiences to facilitate the subsequent model training. By introducing entropy regularization, the SAC algorithm enables the agent to explore more possible actions while preventing it from getting trapped in local optima.

The SAC agent consists of three networks: two Q networks and a random policy network (or actor network). Each of them is a multi-layer perceptron (MLP). The reason for taking two Q networks is to reduce the overestimation of Q values. The Q networks take the states provided by the environment and actions produced by the actor network as inputs and produce Q-value estimates, which are scalar values. Given a network of actors π_ψ parameterized by ψ, the Q network is trained jointly in the SAC algorithm as

$$\hat{\mathbf{y}} \triangleq \tilde{\mathbf{r}} + \min_{i=1,2} Q_{\theta_i'}\left(\mathbf{s}', \mathbf{a}'\right)|_{\mathbf{a}' \backsim \pi_\psi(\cdot|\mathbf{s}')} - \alpha \log\left(\mathbf{a}' \mid \mathbf{s}'\right), \quad (13)$$

$$\mathbf{J}\left(\theta_i\right) = \frac{1}{N}\left\|\hat{\mathbf{y}} - Q_{\theta_i}\left(\mathbf{s}, \mathbf{a}\right)\right\|_2^2, \quad (14)$$

$$\theta_i \leftarrow \theta_i - \eta \nabla_{\theta_i} J(\theta_i), \quad (15)$$

where $(\mathbf{s}, \mathbf{a}, \mathbf{r}, \mathbf{s}')_{i=1}^N$ is a small batch of experience tuples randomly drawn from the experience replay buffer, N denotes the size of the experience tuple. θ_i is the network parameter of the ith Q-network. α is the entropy regularization term, and $||\cdot||_2$ is the L_2 norm.

Similarly, the policy network takes state vectors from the environment as inputs and produces numerical action vectors. The loss function for the policy network in the SAC algorithm is calculated as

$$J(\psi) \triangleq \frac{1}{N}\sum_i^N \alpha \log \pi_\psi\left(\hat{\mathbf{a}}_i \mid \mathbf{s}_i\right) - \min_{j=1,2} Q_{\theta_j}\left(\mathbf{s}_i, \hat{\mathbf{a}}_{\hat{\beta},i}\right)\Big|_{\hat{\mathbf{a}} \sim \pi_\psi(\cdot|\mathbf{s})} \quad (16)$$

For the update of the policy network, we use the stochastic gradient algorithm for the policy gradient $\nabla_\psi J(\psi)$ and the gradient ascent algorithm to update the parameter of the policy network as

$$\psi \leftarrow \psi + \eta \nabla_\psi J(\psi). \quad (17)$$

Finally, the regularization factor of entropy, α is a hyperparameter, and a larger value of α means that the agent will have a larger exploration space, which enables agent to explore more possibilities instead of being limited to the current action.

B. SAC-exploration

Environments for RL are categorized as episodic and non-episodic. When the termination condition is satisfied, the agent stops interacting with the environment, whereas the non-episodic tasks, that donot have a defined termination condition will lead to endless interaction between the agent and the environment. Therefore, we need to set a systematic threshold as a condition for termination.

- **Action**: For the optimization problem we have constructed, the action of the agent is the beamforming design of the RIS. Since the neural networks cannot handle complex numbers, we need to separate the real and imaginary parts of the RIS phases, and the action space contains $2L$ elements.
- **State**: To take full advantage of the information obtained by the agent interacting with the environment, we define the state space as $\mathbf{H}$ and $\mathbf{\Phi}$, the equivalent wireless channel matrix and RIS beamforming. In order to reduce the influence of the correlation of the channel state matrix, we perform a whitening operation on the channel matrix after each environmental interaction.
- **Reward**: For each episode, we will use the sum rate of the downlink transmission to determine the reward in order to choose the right action to get a higher reward for the agent.

For the RL, we cannot just focus on the reward that we get instantaneously, future gains are necessary to consider. We can increase or decrease the size of the future influence on the present reward by adjusting γ, and we set initially $\gamma = 1$.

For the non-episode task, we can use average reward instead of the instantaneous reward to allow the agent to learn from the environment as

$$\tilde{r} \triangleq r - \bar{r}, \quad (18)$$

where r is the immediate reward computed by the environment in the current state and r is the average of the rewards collected before the current state. For the typical SAC algorithm, it is not a good solution of our proposed problem. Therefore, we propose an SAC-exploration algorithm that adds the explorer network to the typical SAC algorithm to interfere with the actions chosen by the policy, and constantly maximize the prediction error of the Q network. As a result, this causes the agents to enter the state-action space where Q prediction is difficult to perform, thus allowing them to correct the prediction error of the unknown or less-selected actions as

$$a_{\hat{\zeta}} \triangleq a \odot \lambda \cdot \xi_\omega(s) \Longrightarrow \hat{\phi}_{\hat{\zeta}} \triangleq \left[\hat{\zeta}_1 e^{j\varphi_1} \ldots \hat{\zeta}_L e^{j\varphi_L}\right]^\top, \quad (19)$$

where $\odot$ is the Hadamard product and the $\lambda \in (0, 1]$ is used to restrict the explorer network from influencing the policy network to choose the wrong action. At each update step, perturbed actions are sampled from the experience replay buffer instead of the original actions generated by the policy network, and then the loss functions for the Q network and the actor network are as follows

$$\hat{\mathbf{y}}_{\hat{\zeta}} \triangleq \tilde{\mathbf{r}} + \min_{i=1,2} Q_{\theta_i'}\left(\mathbf{s}', \mathbf{a}' \odot \lambda \cdot \xi_{\omega'}(\mathbf{s})\right) - \alpha \log\left(\mathbf{a}' \mid \mathbf{s}'\right), \quad (20)$$

Algorithm 1 Proposed SAC- exploration Algorithm

```
Initialize Q-networks $Q_{\theta_i}$ with parameters $\psi$, $\theta_1$, $\theta_2$, actor network $\pi_\psi$ and explorer network $\xi_\omega$ with parameters $\omega$
Initialize target networks: $\hat{\theta}_1 \leftarrow \theta$, $\theta'_2 \leftarrow \theta_2$, $\omega' \leftarrow \omega$ and experience replay buffer $M$
for each environment step do
    Observe state $s$ and select an action $a \backsim \pi_\psi(\cdot|s)$
    Perturb the selected action: $a_{\hat{\zeta}} = a \odot \lambda \cdot \xi_\omega(s)$
    Import$a_{\hat{\zeta}}$, and observe reward $r$ and next state $s'$
    Store$(s, a, a_{\hat{\zeta}}, r, s')$ into the experience replay buffer
end for
for each update step do
    Randomly select an empirical tuple $(\mathbf{s}, \mathbf{a}, \mathbf{a}_{\hat{\beta}}\mathbf{r}, \mathbf{s}')_{i=1}^{N}$
    Update Q-networks by(20)and(16);
    $\theta_i \leftarrow \theta_i - \eta\nabla_{\theta_i} J(\theta_i)$
    Update the actor network by(19);
    $\psi \leftarrow \psi + \eta\nabla_\psi J(\psi)$
    Update the explorer network by(23)(24) and (25)
    Update the target networks:
    $\theta'_{i=1,2} \leftarrow (1-\tau)\theta'_{i=1,2} + \tau\theta_{i=1,2}$
    $\omega' \leftarrow (1-\tau)\omega' + \tau\omega$
end for
```

$$J_{\hat{\zeta}}\left(\theta_i\right) \triangleq \frac{1}{N}\left\|\hat{\mathbf{y}}_{\hat{\beta}} - Q_{\theta_i}\left(\mathbf{s}, \mathbf{a}_{\hat{\beta}}\right)\right\|_2^2, \tag{21}$$

and

$$J_{\hat{\zeta}}(\psi) \triangleq \frac{1}{N}\sum_{i}^{N} \alpha \log \pi_\psi\left(\hat{\mathbf{a}}_i \mid \mathbf{s}_i\right) - \min_{j=1,2} Q_{\theta_j}\left(\mathbf{s}_i, \hat{\mathbf{a}}_{\hat{\zeta},i}\right)\Bigg|_{\hat{\mathbf{a}}\sim\pi_\psi(\cdot|\mathbf{s})}, \tag{22}$$

which $\hat{\mathbf{a}}_{\beta,i} = \hat{\mathbf{a}} \odot \lambda \cdot \xi_\omega$. We define the loss function of the explorer network by maximizing the sum of the absolute temporal-difference (TD) errors corresponding to the two Q networks as

$$\tilde{\delta}_{\hat{\zeta}_i}(\mathbf{s}, \mathbf{a}_{\hat{\zeta}}) \triangleq \frac{1}{N}||\hat{\mathbf{y}}_{\hat{\zeta}} - Q_{\theta_i}(\mathbf{s}, \mathbf{a}_{\hat{\zeta}})||_2^2, \tag{23}$$

$$J_\omega = \tilde{\delta}_{\hat{\zeta}_1}(\mathbf{s}, \mathbf{a}_{\hat{\zeta}}) + \tilde{\delta}_{\hat{\zeta}_2}(\mathbf{s}, \mathbf{a}_{\hat{\zeta}}). \tag{24}$$

The deterministic exploration policy gradient is then computed by the deterministic policy gradient algorithm [12] and used to perform the gradient ascent over the TD errors defined by (23),

$$\nabla_\omega J(\omega) = \sum_{i=1}^{2} \mathbb{E}[\nabla_{\lambda\cdot\xi_\omega(\mathbf{s})}\tilde{\delta}_{\hat{\zeta}_i}(\mathbf{s}, \mathbf{a}_{\hat{\zeta}})\nabla_\omega\xi_\omega(\mathbf{s})], \tag{25}$$

$$\omega \leftarrow \omega + \eta\nabla_\omega J(\omega). \tag{26}$$

Overall, we refer to the resulting algorithm as SAC-exploration and provide the pseudocode in Algorithm 1.

IV. Simulation Results

The BS is equipped with a uniform linear array (ULA) that is placed at [1, 0, 0] (i.e., x-axis), and RIS is a ULA, and located at [0, 1, 0] (i.e., y-axis), and the user is equipped with single antenna and is randomly generated in a circular range. We set the number of the BS antennas as 4, the number of users as 4, and the number of RIS elements as 32. One channel episode includes 20 channel blocks. The parameters of the Q network and the parameters of the RIS hardware loss model are given in Table I.

TABLE I

Hyper-Parameter	Value
hidden layers	2
units in each hidden layer	256
Hidden layers activation	ReLU
actor final layer activation	tanh
explorer final layer activation	tanh
Total time steps per training	20000
Mini-batch size	16
Discount term γ	1
Learning rate for target networks τ	10^{-3}
Initial α	0.2
μ	0
κ	1.5

To verify the effectiveness of the proposed algorithm, we compare the three algorithms, SAC, SAC-exploration and random, under the same configuration. Since the DRL is highly randomized, we conduct 10 experiments and average them in order to facilitate the observation, which makes the results more convincing. In Fig. 2, we analyze the influence of different algorithms on the system sum rate. Compared to the random algorithm, the RIS beamforming design can significantly improve the system sum rate, and the proposed algorithm achieves significant improvement over the SAC algorithm. This is because, unlike the typical DRL algorithm, the error in the value function for the proposed algorithm can be corrected by introducing a perturbation signal to interfere with the action selection. Thus, the agent may select the optimal action more accurately, and the effectiveness of the

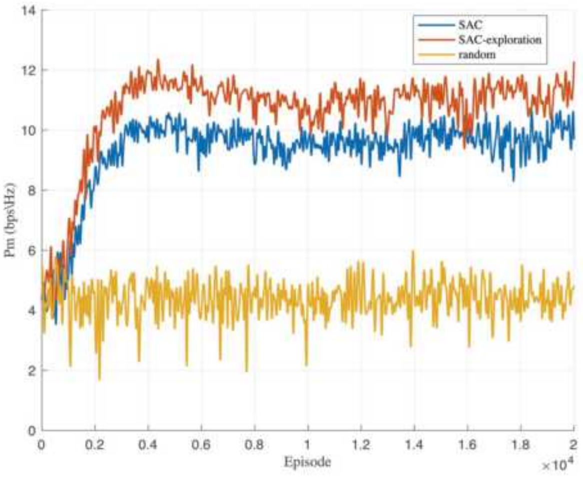


Fig. 2. Effect of different algorithms on P_m at $\zeta = 0.6$ and $L = 32$.

proposed algorithm can be verified by the fact that the system sum rate P_m is significantly improved.

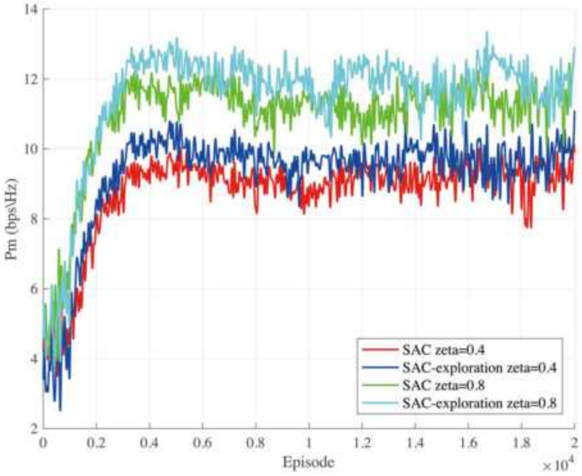


Fig. 3. Effect of different ζ values on P_m at $L = 32$.

The performance of the proposed algorithm for different ζ is investigated in Fig. 3. For the convenience of observation, we set the typical SAC algorithm as a benchmark. Comparing with Fig. 2, it can be seen that the proposed DRL algorithm with $\zeta = 0.6$ has a significant improvement in the system sum rate over the SAC algorithm, and it is also observed that the system sum rate P_m of the proposed algorithm increases as the value of ζ increases, which is due to the fact that the larger the value of ζ, the closer it is to the ideal condition, thus validating the effectiveness of the algorithm.

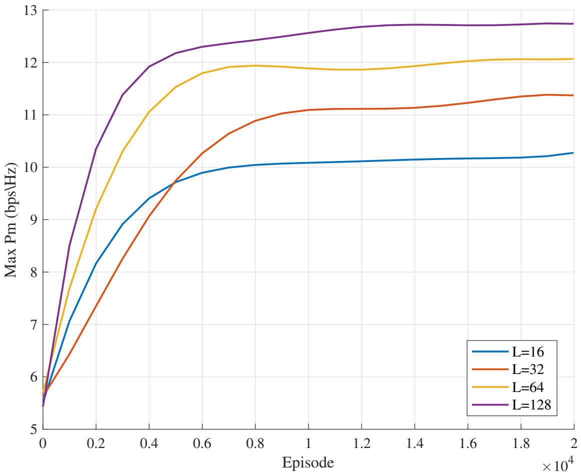


Fig. 4. Effect of different number of RIS elements on the maximum system performance at $\zeta = 0.6$.

Fig. 4 depicts the effect of different number of RIS elements on the system sum rate performance for the proposed algorithm. In order to better observe the experimental results, we take the max P_m within 1000 episodes for the simulation analysis. We can observe that, as the number of RIS elements L increases, the maximum system sum rate P_m increases, and the increase in the P_m is achieved at the expense of the complexity. The simulation results show that the proposed algorithm has significant advantages in future practical applications as well.

V. Conclusion

In this paper, we study the RIS beamforming design in the RIS-assisted MIMO communication and propose a modified SAC-exploration algorithm to solve the RIS beamforming design considering the RIS hardware loss. The simulation results show that the proposed algorithm is able to realize better RIS beamforming and has a significant improvement in the downlink system sum rate compared with the typical SAC algorithm. Considering the case of multiple RISs, we will consider the multi-agent algorithm to solve the beamforming design problem in the future.

References

[1] C. Cai, X. Yuan and Y. -J. A. Zhang, “RIS Partitioning Based Scalable Beamforming Design for Large-Scale MIMO: Asymptotic Analysis and Optimization,” *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6061-6077, Sept. 2023.

[2] S. Liu, R. Liu, M. Li, Y. Liu and Q. Liu, “Joint BS-RIS-User Association and Beamforming Design for RIS-Assisted Cellular Networks,” *IEEE Trans. Veh. Technol.*, vol. 72, no. 5, pp. 6113-6128, May 2023.

[3] H. Luo, R. Liu, M. Li and Q. Liu, “RIS-Aided Integrated Sensing and Communication: Joint Beamforming and Reflection Design,” *IEEE Trans. Veh. Technol.*, vol. 72, no. 7, pp. 9626-9630, July 2023.

[4] M. Goudarzi, M. Palaniswami and R. Buyya, “A Distributed Deep Reinforcement Learning Technique for Application Placement in Edge and Fog Computing Environments,” *IIEEE Trans Mob Comput*, vol. 22, no. 5, pp. 2491-2505, 1 May 2023.

[5] Q. Liu, Y. Zhu, M. Li, R. Liu, Y. Liu and Z. Lu, “DRL-based Secrecy Rate Optimization for RIS-Assisted Secure ISAC Systems,” *IEEE Trans. Veh. Technol.*, pp.1-5, 2023.

[6] H. Zhang, M. Huang, H. Zhou, X. Wang, N. Wang and K. Long, “Capacity Maximization in RIS-UAV Networks: A DDQN-Based Trajectory and Phase Shift Optimization Approach,” *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2583-2591, April 2023.

[7] Y. Zhu, M. Li, Y. Liu, Q. Liu, Z. Chang and Y. Hu, “DRL-based Joint Beamforming and BS-RIS-UE Association Design for RIS-Assisted mmWave Networks,” *IEEE Wireless Commun. Netw. Conf. (WCNC),* 2022, pp. 345-350.

[8] C. Huang et al., “Multi-Hop RIS-Empowered Terahertz Communications: A DRL-Based Hybrid Beamforming Design,” *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1663-1677, June 2021.

[9] S. Abeywickrama, R. Zhang, Q. Wu and C. Yuen, “Intelligent Reflecting Surface: Practical Phase Shift Model and Beamforming Optimization,” *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5849-5863, Sept. 2020.

[10] W. Wang and W. Zhang, “Intelligent Reflecting Surface Configurations for Smart Radio Using Deep Reinforcement Learning,” *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2335-2346, Aug. 2022.

[11] C. Banerjee, Z. Chen and N. Noman, “Improved Soft Actor-Critic: Mixing Prioritized Off-Policy Samples With On-Policy Experiences,” *IEEE Trans Neural Netw Learn Syst*, pp. 1-9, 2022.

[12] X. Yang, H. Zhang and Z. Wang, “Data-Based Optimal Consensus Control for Multiagent Systems With Policy Gradient Reinforcement Learning,” *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 8, pp. 3872-3883, Aug. 2022.

Yuan Sun is currently pursuing the M.S. degree in Electronic Information at the School of Information Science and Engineering, Shandong University, Qingdao, China. His research interests include reconfigurable intelligent surface, deep reinforcement learning, beamforming and signal processing.

Zhiquan Bai received the M.Eng. degree in communication and information system from Shandong University, Jinan, China, in 2003, and the Ph.D. degree (Hons.) from INHA University, Incheon, South Korea, in 2007, under the Grant of Korean Government IT Scholarship. He held a postdoctoral position with INHA University, and was a Visiting Professor with The University of British Columbia, Canada. He is currently a Professor with the School of Information Science and Engineering, Shandong University. His research interests include cooperative technology and spatial modulation, orthogonal time frequency space modulation, MIMO technology, resource allocation and optimization, and deep-learning based 5G wireless communications. He is a member of the editorial board of Journal of Systems Engineering and Electronics and also an associate editor of the International Journal of Communication Systems.

Jinqiu Zhao received B.E. degree from Shandong Normal University, Jinan, China, in 2020. She is currently pursuing her Ph.D. degree in the School of Information Science and Engineering, Shandong University, Qingdao, China. Her main research interests include reconfigurable intelligent surface and machine learning.

Dejie Ma is currently pursuing the M.S. degree in Electronic Information at the School of Information Science and Engineering, Shandong University, Qingdao, China. His research interests include reconfigurable intelligent surface, integrated sensing and communication and signal processing.

Zhaoxia Xian received the B.S. degree from the School of Physical and Electronic Sciences at Shandong Normal University in Jinan, China. She is currently pursuing the M.S. degree in Electronic Information at the School of Information Science and Engineering, Shandong University, Qingdao, China. Her research interests include signal processing, signal recognition and deep learning.

Kyung Sup Kwak received his BS degree from the Inha University, Inchon, Korea, in 1977 and his MS degree from the University of Southern California in 1981 and his PhD degree from the University of California at San Diego in 1988, under the Inha University Fellowship and the Korea Electric Association Abroad Scholarship Grants, respectively.From 1988 to 1989, he was with Hughes Network Systems, San Diego, California. From 1989 to 1990, he was with the IBM Network Analysis Center, North Carolina. Since then, he has been with the School of Information and Communication Engineering, Inha University, Korea, as a professor. He is the director of UWB Wireless Communications Research Center (UWB-ITRC).Since 1994, he served as a member of the board of directors and the vice president and the president of Korean Institute of Communication Sciences (KICS) in 2006 and the president of Korea Institute of Intelligent Transport Systems (KITS) in 2009. He received many research awards, such as the award of research achievements in UWB radio from the Ministry of Information and Communication and Prime Ministry of Korea in 2005 and 2006, respectively. In 2008, he is elected as Inha Fellow Professor (IFP). In 2010, he received the Korean President official commendation for his contribution to ICT innovation and industrial promotion.He published more than 100 SCI journal papers, 300 conference/domestic papers, obtained 20 registered patents and 35 pending patents, and proposed 21 technical proposals on IEEE 802.15 (WPAN) PHY/MAC. He is one of the members of the IEEE, IEICE, KICS, and KIEE. His research interests include multiple access communication systems, cognitive radio, UWB radio systems and WBAN, WPAN, and sensor networks.

Session 3B: Artificial Intelligence 3

Chair: Dr. Jung Joo Yoo, Electronics Telecommunications Research Institute (ETRI), Korea

1 Paper ID: 20240447, 199~205

Transforming Education Policy: Evaluating UAQTE Program Implementation through LDA, BoW and TF-IDF Techniques

Mr. Christian Sy, Dr. Lany Maceda, Dr. Thelma Palaoag, Dr. Mideth Abisado,

Bicol University. Philippines

2 Paper ID: 20240444, 206~210

Leveraging Machine Learning to Uncover Key Factors Influencing Satisfaction Among Free Tertiary Education Recipients in the Philippines

Prof. John Raymund Barajas, Dr. Lea Austero, Dr. Jennifer Llovido, Dr. Lany Maceda, Dr. Mideth Abisado,

Bicol University. Philippines

4 Paper ID: 20240402, 211~215

Classifying Gastric Cancer carcinoma stages with deep semantic features and GLCM Texture Features

Mr. Sikandar Ali, Ms. Samman Fatima, Mr. Ali Hussain, Mr. Maisam Ali , Mr. Muhammad Yaseen, Mr. Tagne Poupi Theodore Armand, Prof. Hee Cheol Kim,

Inje University. Korea(South)

5 Paper ID: 20240453, 216~220

Enhanced Experiences: Benefits of AI-powered recommendation systems

Ms. Kouayep Sonia Carole, Mr. Tagne Poupi Theodore Armand, Prof. Hee-Cheol Kim,

inje university. Korea(South)

Transforming Education Policy: Evaluating UAQTE Program Implementation through LDA, BoW and TF-IDF Techniques

Christian Y. Sy*, Lany L. Maceda*, Thelma D. Palaoag**, Mideth B. Abisado***

*Bicol University, Legazpi City, Philippines

** University of the Cordilleras, Baguio City, Philippines

*** National University, Manila, Philippines.

cysy@bicol-u.edu.ph, llmaceda@bicol-u.edu.ph, tdpalaoag@uc-bcf.edu.ph, mbabisado@national-u.edu.ph

Abstract— **This research aimed to deepen our understanding of the issues and challenges faced by beneficiaries of the Philippines' Universal Access to Quality Tertiary Education (UAQTE) program by utilizing Latent Dirichlet Allocation (LDA) complemented by Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) techniques. The study used the "Boses Ko" participatory toolkit, a digital portal to collect student responses from the various Higher Education Institutions (HEIs). The analysis showcased the LDA model's proficiency in identifying distinct clusters, underscoring its ability to generate coherent and meaningful topics that provide interpretable and insightful dataset representations. Evaluation of the models included Silhouette and Coherence Scores, supplemented by manual assessments by domain experts. This comprehensive approach led to the identification of key themes, including "Academic Difficulties," "Financial Difficulties," "Pandemic-Related Challenges," "Grant Disbursement," and "Program Implementation." The identified themes present actionable policy recommendations geared towards strengthening academic support, financial assistance, flexible grant disbursement, addressing pandemic-related challenges, and establishing a structured feedback mechanism. These measures collectively aim to enhance the overall implementation of the UAQTE program.**

Keywords**— Topic Modeling, Latent Dirichlet Allocation (LDA), Bag-of-Words, TF-IDF, UAQTE program**

I. Introduction

Philippines, where economic and social inequalities persist, tertiary education emerges as a transformative force. It becomes a powerful channel for social mobility, granting individuals from less privileged economic backgrounds equal access to higher education opportunities traditionally reserved for the more affluent. The emphasis on tertiary education aligns with the global agenda, resonating with the United Nations' fourth Sustainable Development Goal (SDG), which emphasizes "Quality Education" as a catalyst for sustainable development [1]–[3].

In the quest for equitable education, the Universal Access to Quality Tertiary Education (UAQTE) program, also known as Republic Act No. 10931was passed on August 13, 2017, requiring all public higher education institutions (HEIs) and government-run technical-vocational institutions (TVIs) to provide free quality tertiary education to eligible Filipino students. This initiative is instrumental in helping individuals break free from the cycle of poverty and catalyze national progress, thereby playing a vital role in the country's socioeconomic development [4], [5]. The program's commitment to providing free tertiary education is fundamental in equipping students with essential knowledge, skills, and critical thinking abilities that are indispensable tools for overcoming the intricate challenges of today's world [6], [7]. This educational empowerment broadens students' perspectives, unveiling various career possibilities and empowering graduates to contribute substantially to their communities and the country [8], [9].

As the UAQTE program expands, it becomes increasingly important to assess its implementation holistically by considering the perspectives of student beneficiaries. This approach allows for a more thorough evaluation of its effectiveness, identifies potential areas for improvement, and gathers valuable insights. To achieve this, our research aims to leverage topic modeling techniques implementing Latent Dirichlet Allocation (LDA), Bag of Words (BoW), and Term Frequency-Inverse Document Frequency (TF-IDF) to uncover patterns, and correlations within the dataset, providing a comprehensive understanding of the UAQTE program's implementation and its impact on stakeholders.

Topic modeling is a statistical technique within natural language processing domain that is used to uncover underlying themes or topics in a textual data collection. It facilitates the organization and categorization of extensive unstructured text, enabling the extraction of meaningful insights [10], [11]. Through topic modeling, we can effectively reveal the underlying structure in the textual data by grouping words that tend to co-occur frequently within the same context [12]–[15].

Central to this study is the "Boses Ko" participatory toolkit, a digital platform developed through a collaborative effort between Bicol University and National University, funded by the Commission on Higher Education - Leading the Advancement of Knowledge in Agriculture and Sciences (CHED-LAKAS). This tool is crucial in collecting the viewpoints of student beneficiaries of the UAQTE program, serving as a pivotal avenue for students to articulate the encountered issues and challenges.

Building upon the pre-processed dataset, this study integrates a multifaceted qualitative modeling approach, employing Latent Dirichlet Allocation (LDA), a well-established methodology, complemented by the use of Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) techniques, both of which play a crucial role in the topic modeling process. The Bag of Words (BoW) model lays the groundwork for representing textual data in this study, analyzing documents by the frequency of words while setting aside their order or contextual significance [16]–[18]. This simplification of the dataset aids Latent Dirichlet Allocation (LDA) in efficiently identifying and classifying topics based on the occurrence of words.

Consecutively, the study employs the Term Frequency-Inverse Document Frequency (TF-IDF) technique to elevate the analysis. TF-IDF measures the importance of each word within a document relative to a collection of documents, assigning a weight to the frequency of each word. This method effectively diminishes the impact of commonly occurring words across the dataset, instead highlighting words that are more distinctive to individual documents [19]–[21]. Such weighted differentiation enhances LDA's ability to uncover and distinguish more relevant and unique topics in the dataset.

The generated models are evaluated combining automated metrics using the Silhouette and Coherence Score with manual topic assessment guided by domain experts' insights. The Silhouette Score is instrumental in determining the clustering quality of the model by measuring how similar each data point is to its own cluster compared to neighboring clusters [22], [23]. This metric is crucial for assessing the effectiveness of the clustering approach taken by the LDA model. The Coherence Score, on the other hand, is used to evaluate the interpretability and quality of the topics identified by the model. It quantitatively measures the semantic relationships among the most significant words within each topic, ensuring they form a coherent and meaningful theme [24], [25]. These automated metrics offer objectivity and efficiency in the initial stages of model assessment, allowing for rapid refinement and baseline evaluations of the model's performance.

The domain experts' insights capture the depth and nuances of the model's output. They bring a vital perspective to the evaluation process, considering aspects such as the relevance, interpretability, and real-world applicability of the topics generated by the model[26]–[28]. Their specialized knowledge and experience in the domain allow them to provide a more distinct analysis, aligning the model's findings with the intricacies and specific requirements of the research objectives.

Ultimately, this research aims to provide evidence-based insights to enhance and optimize the UAQTE program implementation, benefiting students and the wider educational sector. It highlights the importance of collaboration and stakeholder involvement, enabling education sectors to actively shape the UAQTE program and influence future tertiary education policies. This strategy demonstrates a commitment to ongoing improvement and adaptive policymaking, ensuring the program stays dynamic, aligns with the evolving educational landscape, and meets the diverse needs of its beneficiaries.

II. Methodology

This section details the methodology used. Figure 1 illustrates the information processing phases, which include data collection and dataset pre-processing, feature extraction, topic modeling, topic interpretation and labeling, and model evaluation.

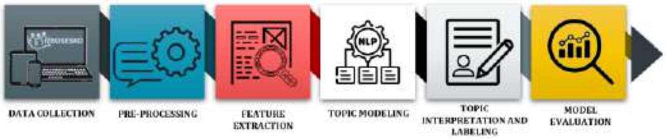


Figure 1. Information Processing Phases

A. *Data Collection*

The "Boses Ko" toolkit has played a pivotal role in implementing a bottom-up approach to data collection from student beneficiaries, emphasizing the importance of insights from those directly involved in the UAQTE program. Tailored to capture the unique perspectives of these students, the central qualitative question of this study is: "What are the specific issues and challenges faced by beneficiaries of the UAQTE program?"

The study included a sample of 2,800 student beneficiaries from various State Universities and Colleges (SUCs) across the Bicol region. This broad selection ensured a thorough evaluation of the UAQTE program's effectiveness and impact in different institutional settings.

B. *Data Pre-processing*

The preparation of the collected responses for feature extraction began with critical data pre-processing steps. This included removing non-English, duplicate, non-grantee, and empty responses, followed by text standardization. Standardization involves converting text to lowercase and removing special characters, punctuation, and digits, aiming to refine the textual representation for clearer analysis and reduced noise. Further processing utilized the Natural Language Toolkit (NLTK) for tokenization and the removal of common stopwords (e.g., "the," "is," "and"). This step segmented the text into individual words or tokens, and removed frequently occurring but semantically limited words, enhancing the clarity and interpretability of the topics. Additionally, domain-specific terms like "UAQTE," "issues," "challenges," and "beneficiaries" were treated as additional stopwords to focus the text analysis and produce more meaningful results by highlighting key terms.

The study intentionally avoided using stemming and lemmatization techniques to retain the nuanced meanings of words. Applying these methods could risk oversimplifying the language and result in a loss of textual clarity, as they reduce words to their base forms or substitute them with similar-meaning terms. For instance, transforming "synchronous" into "synchron" or "better" into "good" might change the original intent or context, potentially leading to misinterpretation or loss of critical information.

C. Feature Extraction

In the feature extraction of Latent Dirichlet Allocation (LDA), two predominant techniques are utilized: the Bag of Words (BoW) model and the Term Frequency-Inverse Document Frequency (TF-IDF) transformation. These methods are instrumental in uncovering underlying structures and enhancing insights from text data [29], [30]. BoW models each document as a vector, quantifying the frequency of words within it [31], [32]. This enables LDA to translate textual data into numerical features, facilitating the identification of underlying topics and their distribution across documents.

TF-IDF augments feature representation quality, building upon the BoW model [33]. It assigns weights to words based on their relevance in individual documents and the entire corpus, balancing word significance. This step, following BoW, produces a more informative and distinct feature matrix, aiding LDA in assigning importance to words unique to specific documents or topics. By doing so, it diminishes the weight of common words.

Integrating TF-IDF after the BoW method is crucial for refining feature representation, reducing noise, and improving topic distinction within textual data [34]. This methodology ensures that LDA more accurately captures the distinctive elements of the dataset, resulting in more precise and meaningful categorization of topics.

$$tf - idf(t) = tf(t, d)\; x\; idf(t) \qquad (1)$$

Equation 1 assesses the importance of the term 't' within a specific document 'd' throughout a collection of documents. This evaluation incorporates two primary factors: the frequency of the term within the document ('tf') and its scarcity or distinctiveness across the entire collection of documents ('idf').

D. Topic Modeling/Extraction

Building on the initial feature extraction phase using the Bag of Words (BoW), where documents are represented as frequency vectors of words, LDA fundamentally uncovers hidden topics within the text corpus. Operating on the document-term matrix derived from BoW, LDA interprets each document as a composition of various topics. Each word occurrence within a document is associated with one of these topics. The algorithm dynamically refines its understanding, adjusting topic associations for each word by analyzing word co-occurrence patterns throughout the corpus.

The integration of the TF-IDF methodology refines the BoW representation. This approach assigns specific weights to terms, considering their frequency within individual documents and their overall relevance in the entire corpus. This weighted TF-IDF matrix then serves as the input for LDA, enabling it to delve into the significance and distribution of terms to pinpoint underlying themes. By evaluating the prominence of particular terms across different documents, the algorithm seeks to grasp the semantic interconnections among words. This approach enriches the topic modeling process, ensuring a more sophisticated and accurate interpretation of the textual content.

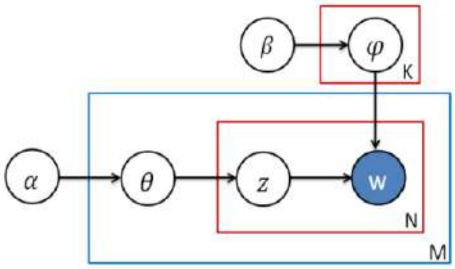


Figure 2. Topic Extraction using LDA

Figure 3 presents the Latent Dirichlet Allocation (LDA) model's generative mechanism, offering a systematic document formation perspective. This includes the process of topic selection, word distribution within these topics, and the role of hyperparameters α and β in shaping the distributions of topics and words.

- α (alpha): This parameter influences the Dirichlet prior for topic distributions within each document (θ). It determines how topics are varied across individual documents.
- β (beta): This parameter governs the Dirichlet prior for word distributions within each topic (φ). It specifies how words are distributed across topics.
- θ(theta)M: This represents the distribution of topics in document M, depicting the range of topics that make up the document.
- φ(phi)K: This denotes the distribution of words within topic K, showing the probability of each word appearing in topic K.
- Zmn: Indicates the topic allocation for specific words, particularly the topic assigned to the nth word in document m.
- Wmn: Represents the actual words in the corpus. Each Wmn is a unique word in the document, and combined with Zmn, it shows the particular word associated with a specific topic in that document.

E. Hyper-Parameters

The following are the hyper-parameters used:

- Number of Topics (K). determines the level of detail and variety of themes extracted from the text corpus.
- Number of Words. Indicates the total number of distinct words accounted for in the analysis, defining the breadth of the vocabulary considered.
- Passes. Dictates how many times the algorithm reviews the entire corpus during its training phase. Each pass involves an update in the distributions of topics for documents and the distributions of words for topics.
- Iterations. Denotes the frequency of iterations within each training pass. This controls how often the model reassesses the entire dataset to refine its topic assignments.
- Alpha (α). Affects how topics are distributed across documents. A higher alpha value results in documents featuring a wider range of topics, whereas a lower alpha

leads to more concentrated topic distributions in each document.

- Eta (η). Impacts the density of word distributions within each topic. Higher eta values foster a more expansive distribution of words across topics, while lower values encourage a tighter, more focused word distribution for each topic.

TABLE 1. HYPER-PARAMETERS USED

Number of topics	Number of words	Passes	Iterations	alpha	eta
5 to 20	5 to 10	10 to 100	100 to 5000	0.01	0.01

Table 1 illustrates the range of hyperparameters examined to ascertain the optimal configurations for the LDA model experiments. These configurations are pivotal in producing coherent, interpretable, and contextually relevant topics derived from the dataset. Fine-tuning these hyperparameters is critical to align the topic modeling process with the dataset's unique attributes, significantly impacting the clarity and relevance of the extracted topics.

F. Identification of Appropriate Themes

The main objective in labeling the topic model was assigning pertinent labels to each topic, guided by the observed word similarities. This approach aimed to amplify topic clarity and strengthen effective communication, substantially enhancing their applicability within the UAQTE framework. Working in close partnership with domain experts from various disciplines, including CHED administrators, social scientists, data scientists, and recipients of the UAQTE program, greatly aided in choosing topic labels that were clearly understandable and closely aligned with the research objectives. Table 2 displays the labels established through this collaborative process, reflecting a well-rounded and research-focused understanding of the topics.

TABLE 2. DOMAIN EXPERTS' IDENTIFIED LABELS

Categories	Description
Financial Difficulties	Refers to the challenges faced due to limited financial resources, including difficulties in budgeting, paying for miscellaneous fees, boarding, daily expenses, and additional educational costs not covered by the program or scholarship.
Grant Disbursement	Refers to issues of scholarships or financial aid being delayed in disbursement, causing financial strain and difficulties meeting educational expenses.
Academic Difficulties	Refers to the challenges in fulfilling various academic obligations, such as preparing and submitting requirements, managing coursework, and striving to achieve satisfactory academic performance. Other factors like lack or poor facilities and infrastructures are likewise part of this category.
Pandemic-Related Challenges	Refers to the obstacles and difficulties arising from the COVID-19 pandemic, including the transition to online learning, lack of access to reliable internet connectivity and necessary devices, disruptions in academic schedules, and the impact on mental health, stress, anxiety, and feelings of isolation.
Program Implementation	Refers to the diverse viewpoints regarding the program's implementation, incorporating a range of positive and negative perspectives.

G. Model Evaluation

The Silhouette score is calculated as follows:

$$s = \frac{b-a}{max(a,b)} \tag{2}$$

The silhouette score (s) of a data point is calculated by comparing its average distance (a) to points in the same cluster and (b) to points in the nearest neighboring cluster. A score close to +1 indicates strong cluster alignment and distinctiveness from neighboring clusters, suggesting clear and well-separated groupings. Scores near zero suggest the data point is on the boundary between clusters, indicating ambiguity in its assignment. Scores approaching -1 imply possible misplacement of the object in the wrong cluster, hinting at issues like cluster overlap or indistinct boundaries. Overall, silhouette scores provide insights into cluster separation and clarity, helping evaluate cluster quality and potential overlaps in the dataset.

The Coherence score is calculated as follows:

$$C\_V(T) = 2 / (|T|(|T| - 1)) * \sum\_\{i = 1\}\hat{}\{|T|\} \sum\_\{j \neq i\} sim(Word\_i, Word\_j) \tag{3}$$

The coherence score (CV(T)) for topic T is determined by considering the number of words in the topic, the representation of word pairs (Wordi and Wordj) in the topic, and their similarity measure. This score assesses the connectedness and meaningfulness of words within a topic. A score near 0 indicates poor word connections, leading to difficult interpretation. Scores around 0.2 to 0.4 suggest limited coherence and interpretability. Scores between 0.4 to 0.6 denote reasonably coherent topics, improving interpretability. Scores from 0.6 to 0.8 reflect well-defined, highly coherent topics that are easy to interpret. Scores approaching 0.8 to 1 indicate exceptionally coherent topics with closely related words, offering high interpretability.

Domain experts are crucial in the evaluation phase, bringing essential contextual insights and specialized knowledge. Their input is vital given the multifaceted requirements of topic models, which include statistical integrity, semantic coherence, and relevance to the specific field. As key validators, they ensure that the generated models align with the research objectives.

Combining expert insights with silhouette and coherence scores establishes a robust evaluation method. Good silhouette and coherence scores, alongside expert validation of topic interpretability, collectively indicate the high quality of the generated topics. Integrating quantitative metrics with the judgment of domain experts creates a well-rounded assessment framework. It guarantees a nuanced and dependable evaluation of the topic model's overall quality and effectiveness, ensuring that the models are not just statistically sound but rich in context and relevance to the UAQTE domain.

III. Results and Discussion

This section presents key results and findings obtained through topic modeling of the UAQTE dataset implementing LDA. Table 3 highlights the configurations that achieved acceptable silhouette and coherence scores from the numerous experiments. Both scores aim to provide quantitative measures of the quality or usefulness of the output produced by unsupervised learning algorithms. Higher scores in both cases generally indicate better results, with higher silhouette scores implying well-defined clusters and higher coherence scores indicating more interpretable topics.

LDA model analysis investigated various configurations, including topics, words, passes, iterations, alpha, and eta, leading to variations in silhouette and coherence scores under different parameter settings, as indicated in Table 3. In particular, configurations with 8 topics, 50 passes, and 3000 iterations showcased higher silhouette scores, implying the model's effectiveness in creating more distinct data clusters. These settings reflect the LDA's strength in delineating clear groupings within the dataset.

TABLE 3. LDA Hyperparameters and Evaluation Scores

Number of topics	Number of words	Passes	Iterations	alpha	eta	Silhouette Score	Coherence Score
8	10	20	1000	0.01	0.01	0.771	0.618
8	10	50	1000	0.01	0.01	0.761	0.594
8	10	10	3000	0.01	0.01	0.769	0.617
8	10	50	3000	0.01	0.01	0.789	0.623
8	10	30	5000	0.01	0.01	0.761	0.591
8	10	100	5000	0.01	0.01	0.770	0.627
10	10	20	1000	0.01	0.01	0.728	0.603
10	10	50	1000	0.01	0.01	0.746	0.596
10	10	10	3000	0.01	0.01	0.740	0.605
10	10	50	3000	0.01	0.01	0.739	0.604
10	10	10	5000	0.01	0.01	0.735	0.645
10	10	50	5000	0.01	0.01	0.771	0.634

Although the coherence scores ranged between 0.591 and 0.645, indicating some challenges in capturing complex semantic relationships due to the probabilistic nature of LDA, the model still showed a considerable degree of effectiveness in associating topics and words. This was instrumental in generating relevant labels. Despite the relatively modest coherence scores, LDA demonstrated its capability to formulate coherent and meaningful topics. This underscores its value in delivering interpretable and insightful dataset representations, highlighting its practical importance in data analysis.

Table 4 features the LDA model labeled by domain experts, with the highest Silhouette and Coherence Scores among all the LDA experiments. This model is characterized by its hyperparameters: it includes 8 topics, the top 10 words, 50 passes, and 3000 iterations, along with an alpha value set at 0.01 and an eta value of 0.01. The model attained a Silhouette Score of 0.789 and a Coherence Score of 0.623 in this configuration.

TABLE 4. LDA Labeled Model

Topic	Words	Label
0	hard, expectations, scholarship, support, pressure, help, high, study, passing, excel	Academic Difficulties
1	lack, facilities, equipment, information, encounter, poor, resources, school, infrastructures, classrooms	Academic Difficulties
2	school, expenses, financial, problems, pay, need, requirements, family, enough, fee	Financial Difficulties
3	education, quality, access, lack, sustain, limited, free, learning, financial, expenses	Financial Difficulties
4	pressure, passing, academic, expenses, perform, requirements, work, studying, school	Academic Difficulties
5	grades, pandemic, online, good, maintaining, experience, pressure, academic, learning, classes	Pandemic-Related Challenges
6	school, free, money, tuition, allowance, delayed, expensive, release, bills, patience	Grant Disbursement
7	encounter, education, program, free, study, things, tuition, regarding, facilities, process	Program Implementation

The extensive LDA experiments highlighted key themes such as "Academic Difficulties," "Financial Difficulties," "Pandemic-Related Challenges," "Grant Disbursement," and "Program Implementation." The analysis underscores the significance of these core topics, illustrating the model's ability to identify and prioritize crucial themes relevant to the dataset.

Integrating Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) methods with LDA remarkably enhances the model's interpretability. BoW contributed to improved topic identification and model stability, ensuring clear and consistent topic representation. On the other hand, TF-IDF refined topic descriptors and balanced word representation, offering nuanced and contextually relevant topics by focusing on unique words.

IV. Conclusions

Investigating various parameters and evaluation metrics in LDA topic modeling yielded valuable insights. Key performance metrics such as silhouette and coherence scores highlighted the importance of parameter optimization in LDA, particularly its ability to define distinct topic clusters. This underlines the necessity of tailoring modeling approaches to align with specific research objectives and the unique characteristics of datasets.

Critically, the involvement of domain experts in assigning labels to the generated LDA models has been instrumental in uncovering key insights for enhancing the UAQTE program. Their domain-specific knowledge guided the interpretation of themes such as "Academic Difficulties," "Financial Difficulties," "Grant Disbursement," "Pandemic-Related Challenges," and "Program Implementation." This has led to robust recommendations, including strengthening academic support, enhancing financial assistance, flexible grant disbursement policies, addressing pandemic-related challenges, and establishing a structured feedback mechanism for program improvement.

These recommendations provide actionable directions for policy reforms within the UAQTE program, aiming to improve student support and the overall efficiency of the educational sector. It is crucial to note that these policy recommendations are part of an iterative process, requiring ongoing evaluation and flexibility to adapt to evolving circumstances, ensuring the long-term impact and relevance of the UAQTE program.

Acknowledgment

The researchers extend their heartfelt thanks to the Philippine Commission on Higher Education (CHED) and the Leading the Advancement of Knowledge in Agriculture and Science (LAKAS) Project No. 2021-007, particularly for the eParticipation 2.1 initiative: "Harnessing Natural Language Processing (NLP) for Community Participation." We deeply appreciate the financial support from this project, which was crucial in facilitating our research.

References

[1] E. G. García, E. C. Magaña, and A. C. Ariza, “Quality education as a sustainable development goal in the context of 2030 agenda: Bibliometric approach,” Sustainability (Switzerland), vol. 12, no. 15, Aug. 2020, doi: 10.3390/SU12155884.

[2] E. Boeren, “Understanding Sustainable Development Goal (SDG) 4 on ‘quality education’ from micro, meso and macro perspectives,” International Review of Education, vol. 65, no. 2, pp. 277–294, Apr. 2019, doi: 10.1007/s11159-019-09772-7.

[3] R. Nazar, I. S. Chaudhry, S. Ali, and M. Faheem, “Role of Quality Education for Sustainable Development Goals (SDGS),” PEOPLE: International Journal of Social Sciences, vol. 4, no. 2, pp. 486–501, 2018, doi: 10.20319/pijss.2018.42.486501.

[4] O. Legusov, R. L. Raby, L. Mou, F. Gómez-Gajardo, and Y. Zhou, “How community colleges and other TVET institutions contribute to the united nations sustainable development goals,” J Furth High Educ, vol. 46, no. 1, pp. 89–106, 2022, doi: 10.1080/0309877X.2021.1887463.

[5] E. Lalith, A. P. S. K., S. Sampath, and R. Lakshmi, “Creating a psychological paradigm shift in students choice for tertiary education in Sri Lanka: The influence of socioeconomic factors,” International Journal of Educational Administration and Policy Studies, vol. 14, no. 1, pp. 1–16, Jan. 2022, doi: 10.5897/ijeaps2021.0722.

[6] C. Hursen, “The Effect of Problem-Based Learning Method Supported by Web 2.0 Tools on Academic Achievement and Critical Thinking Skills in Teacher Education,” Technology, Knowledge and Learning, vol. 26, no. 3, pp. 515–533, Sep. 2021, doi: 10.1007/s10758-020-09458-2.

[7] T. Kromydas, “Rethinking higher education and its relationship with social inequalities: Past knowledge, present state and future potential,” Palgrave Commun, vol. 3, no. 1, Dec. 2017, doi: 10.1057/s41599-017-0001-8.

[8] M. P. S. Mousavi et al., “Stress and Mental Health in Graduate School: How Student Empowerment Creates Lasting Change,” J Chem Educ, vol. 95, no. 11, pp. 1939–1946, Nov. 2018, doi: 10.1021/acs.jchemed.8b00188.

[9] B. A. Burt, B. D. Stone, R. Motshubi, and L. D. Baber, “STEM validation among underrepresented students: Leveraging insights from a STEM diversity program to broaden participation.,” J Divers High Educ, 2020, doi: 10.1037/dhe0000300.

[10] Abdou Khadre Diop, Serban Meza, Mihaela Gordan, and Aurel Vlaicu, “LDA based classification of video surveillance sequences using motion information,” 2018 20th International Conference on Advanced Communication Technology (ICACT), 2018.

[11] M. Hasan, A. Rahman, M. Razaul, M. Saikat Islam Khan, M. Razaul Karim, and M. Jahidul Islam, “Normalized Approach to Find Optimal Number of Topics in Latent Dirichlet Allocation (LDA),” 2021. [Online]. Available: https://www.researchgate.net/publication/344803887

[12] J. A. Lossio-Ventura, S. Gonzales, J. Morzan, H. Alatrista-Salas, T. Hernandez-Boussard, and J. Bian, “Evaluation of clustering and topic modeling methods over health-related tweets and emails,” Artif Intell Med, vol. 117, Jul. 2021, doi: 10.1016/j.artmed.2021.102096.

[13] A. Masoud, W. Cheruiyot, K. Ogada, C. Author, and A. Masoud, “Topic of Interest Discovery on Social Media Using Knowledge Base and Term Frequency-Inverse Document Frequency Techniques,” 2018, 2018, doi: 10.9790/1813-0710030120.

[14] A. Meddeb and L. Ben Romdhane, “Using Topic Modeling and Word Embedding for Topic Extraction in Twitter,” in Procedia Computer Science, Elsevier B.V., 2022, pp. 790–799. doi: 10.1016/j.procs.2022.09.134.

[15] S. M. Ozdemirci and M. Turan, “Case Study on well-known Topic Modeling Methods for Document Classification,” in Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021, Institute of Electrical and Electronics Engineers Inc., Jan. 2021, pp. 1304–1309. doi: 10.1109/ICICT50816.2021.9358473.

[16] M. Yousef and D. Voskergian, “TextNetTopics: Text Classification Based Word Grouping as Topics and Topics’ Scoring,” Front Genet, vol. 13, Jun. 2022, doi: 10.3389/fgene.2022.893378.

[17] J. Wang and Y. Dong, “Measurement of text similarity: A survey,” Information (Switzerland), vol. 11, no. 9. MDPI AG, pp. 1–17, Sep. 01, 2020. doi: 10.3390/info11090421.

[18] W. H. Park, I. F. Siddiqui, C. Chakraborty, N. M. F. Qureshi, and D. R. Shin, “Scarcity-aware spam detection technique for big data ecosystem,” 2023.

[19] X. Li, S. Wang, R. Malekian, S. Hao, and Z. Li, “Numerical Simulation of Rock Breakage Modes under Confining Pressures in Deep Mining: An Experimental Investigation,” IEEE Access, vol. 4, pp. 5710–5720, 2016, doi: 10.1109/ACCESS.2016.2608384.

[20] Institute of Electrical and Electronics Engineers, Institute of Electrical and Electronics Engineers. Morocco Section, and Jāmiʻat al-Ḥasan al-Thānī. Ecole Normale Supérieure de l’Enseignement Technique de Mohammedia, ICOA 2018 Optimization : proceedings of the 2018 International Conference on Optimization and Applications (ICOA) : April 26-27, 2018, Mohammedia, Morocco. 2018.

[21] S. Qaiser and R. Ali, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents,” Int J Comput Appl, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.

[22] R. Vangara et al., “Finding the Number of Latent Topics with Semantic Non-negative Matrix Factorization,” IEEE Access, 2021, doi: 10.1109/ACCESS.2021.3106879.

[23] M. V. Mantyla, M. Claes, and U. Farooq, “Measuring LDA topic stability from clusters of replicated runs,” in International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society, Oct. 2018. doi: 10.1145/3239235.3267435.

[24] H. M. Alash and G. A. Al-Sultany, “Improve topic modeling algorithms based on Twitter hashtags,” in Journal of Physics: Conference Series, IOP Publishing Ltd, Nov. 2020. doi: 10.1088/1742-6596/1660/1/012100.

[25] P. Tijare and P. J. Rani, “Exploring popular topic models,” in Journal of Physics: Conference Series, IOP Publishing Ltd, Dec. 2020. doi: 10.1088/1742-6596/1706/1/012171.

[26] B. Gencoglu, M. Helms-Lorenz, R. Maulana, E. P. W. A. Jansen, and O. Gencoglu, “Machine and expert judgments of student perceptions of teaching behavior in secondary education: Added value of topic modeling with big data,” Comput Educ, vol. 193, Feb. 2023, doi: 10.1016/j.compedu.2022.104682.

[27] E. M. Rinke, T. Dobbrick, C. Löb, C. Zirn, and H. Wessler, “Expert-Informed Topic Models for Document Set Discovery,” Commun Methods Meas, vol. 16, no. 1, pp. 39–58, 2022, doi: 10.1080/19312458.2021.1920008.

[28] Q. Fu, Y. Zhuang, J. Gu, Y. Zhu, and X. Guo, “Agreeing to Disagree: Choosing Among Eight Topic-Modeling Methods,” Big Data Research, vol. 23, Feb. 2021, doi: 10.1016/j.bdr.2020.100173.

[29] A. P. Pimpalkar and R. J. Retna Raj, “Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features,” ADCAIJ: Advances in Distributed Computing and

Artificial Intelligence Journal, vol. 9, no. 2, pp. 49–68, Jun. 2020, doi: 10.14201/adcaij2020924968.

[30] P. Celard, A. S. Vieira, E. L. Iglesias, and L. Borrajo, "LDA filter: A Latent Dirichlet Allocation preprocess method for Weka," PLoS One, vol. 15, no. 11 November, Nov. 2020, doi: 10.1371/journal.pone.0241701.

[31] B. Ozyurt and M. A. Akcayol, "A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA," Expert Syst Appl, vol. 168, Apr. 2021, doi: 10.1016/j.eswa.2020.114231.

[32] C. Yang, J. Barth, D. Katumullage, and J. Cao, "Wine Review Descriptors as Quality Predictors: Evidence from Language Processing Techniques," Journal of Wine Economics, vol. 17, no. 1, pp. 64–80, Feb. 2022, doi: 10.1017/jwe.2022.3.

[33] G. Kontonatsios, S. Spencer, P. Matthew, and I. Korkontzelos, "Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews," 2020, doi: 10.1016/j.eswax.2020.10.

[34] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, S. M. Al-Ghuribi, and F. A. Ghanem, "Enhancing Big Social Media Data Quality for Use in Short-Text Topic Modeling," IEEE Access, vol. 10, pp. 105328–105351, 2022, doi: 10.1109/ACCESS.2022.3211396.

Christian Y. Sy, is working on his DIT degree at the University of the Cordilleras, received his MIT degree in the same institution. An associate professor IV of Bicol University, his interests lie in the dynamic blend of machine learning, data analytics, and natural language processing (NLP). Published several machine learning and natural language processing papers in various scientific journals.

Dr. Lany L. Maceda, received her DIT and MIT degrees at the University of the Cordilleras. Her fields of research include Machine Learning, Data Analytics, and Soft Computing. Her professional affiliations include Computing Society of the Philippines-Special Interest Group on Natural Language Processing (NLP), Special Interest Group in Women in Computing (WIC), and Commission on Higher Education (CHED) Regional Quality Assessment Team (RQAT).

Dr. Thelma D. Palaoag, Director of the Innovation and Technology Transfer Office of the University of the Cordilleras, she is a visionary leader, accomplished researcher, and trailblazer in the field of Information Technology (IT). With an unwavering commitment to advancing technological frontiers, she has dedicated her career to pushing the boundaries of IT research and fostering innovation within academic institutions.

Dr. Mideth B. Abisado, received her DIT degree at the Technological Institute of the Philippines, MCS at the Mapua Institute of the Philippines. Her fields of research include Artificial Intelligence, Adaptive Learning, E-Learning, Emphatic Computing, and Machine Learning. Her professional affiliations include Computing Society of the Philippines, National Research Council of the Philippines, and Association for Computing Machinery.

Leveraging Machine Learning to Uncover Key Factors Influencing Satisfaction Among Free Tertiary Education Recipients in the Philippines

John Raymund B. Baragas*, Lea D. Austero*, Jennifer L. Llovido*, Lany L. Maceda*, Mideth B. Abisado**

*Computer Science and Information Technology, College of Science, Bicol University, Albay, Philippines

** College of Computing and Information Technologies, National University, Manila, Philippines

jrbbarajas@bicol-u.edu.ph, ldaustero@bicol-u.edu.ph, jllovido@bicol-u.edu.ph llmaceda@bicol-u.edu.ph, mbabisado@national-u.edu-ph

Abstract— **In spite of the broad implementation of the Universal Access to Quality Tertiary Education Act (UAQTE) – a groundbreaking legislation benefitting over 2 million students since its enactment in 2017 – a comprehensive evaluation of its outcomes has been notably absent. To bridge this gap, an extensive survey was undertaken among graduating tertiary students in selected regions of the Philippines. This strategically designed survey aimed to pinpoint overlooked aspects of UAQTE and capture firsthand insights from its recipients. The methodology employed to create this survey included focus-group discussions with various stakeholders (i.e., students, parents, faculty) and a pilot test reflecting the target demographic. To facilitate analysis of the results of 1462 responses, five regression machine learning algorithms were then employed to analyze questionnaire data. The decision tree regressor with a root-mean-squared-error of 0.6881 was found to be the best performing model describing the collected questionnaire data. Shapley explanations of the best performing model highlighted the desire of the recipient to pursue international employment as the top predictor of satisfaction in UAQTE among its recipients. Furthermore, insights from employed topic modeling among the open-ended questions in the deployed survey suggested potential inadequacy of UAQTE subsidies, specifically to recipients whose pursued degrees are in the science, technology, engineering, and mathematics courses. This substantial finding promises valuable insights into the effectiveness of the legislation and may inform future policy adjustments to better address the diverse needs of tertiary education in the Philippines. Overall, this research provides a robust framework for assessing the impact of UAQTE and showcases a methodologically sound approach in integrating machine learning and qualitative analysis.**

Keywords— **Machine Learning, Tertiary Education, Satisfaction, Free Education, Philippines**

I. INTRODUCTION

In the dynamic tapestry of Philippine legislation, the Universal Access to Quality Tertiary Education Act (UAQTE), signed into law in 2017, stands as an enduring monument to the commitment of the Philippine government to foster an inclusive and empowered society. This groundbreaking legislation has not only redefined the educational landscape in the country but it also emerged as a beacon of hope which extended the promise of higher learning to at least 2 million eager minds. With this achievement, the UAQTE then represented a pivotal chapter in the ongoing narrative of Philippine education and this signaled a profound shift towards democratizing access to tertiary education especially in a third-world country like the Philippines.

With the goal of enshrining the principle that education is a fundamental right, this newly implemented policy indeed dismantled barriers that once hindered the pursuit of knowledge for many aspiring students. By championing universal access to tertiary education, the legislation has become a catalyst for social mobility that has already started a future where education is not a privilege but essentially a birthright.

However, amidst its widespread implementation, a comprehensive evaluation of the outcomes and impacts of UAQTE has remained conspicuously absent. This void in understanding prompted the initiation of an extensive survey among but not limited to graduating tertiary students in selected regions of the Philippines, with the overarching goal of shedding light on the nuanced facets of UAQTE positioned to capture firsthand insights from its beneficiaries.

In an attempt to achieve this primary goal, this work employed a methodology that was both strategic and inclusive. Through focus-group discussions involving various stakeholders, including students, parents, and faculty, the deployed survey of this study was meticulously designed to identify potential oversights in the UAQTE framework. A subsequent pilot test, mirroring the demographic profile of the target respondents, ensured the relevance and efficacy of the survey instrument. To rigorously analyze the wealth of data gathered from the 1462 collected survey responses, five regression machine learning algorithms were employed. Notably, the decision tree regressor emerged as the best-performing model and achieved the lowest root-mean-squared-error across all tested models. Shapley explanations derived from this model spotlighted as well that the baccalaureate degree was the paramount predictor of satisfaction among UAQTE recipients.

Beyond quantitative analysis, this work also delved into qualitative insights through topic modeling of open-ended survey questions. This exploration unveiled potential inadequacies in UAQTE subsidies, particularly for students pursuing degrees in science, technology, engineering, and mathematics (STEM) fields. It could also be inferred from the results of this work that the significance of these findings extended beyond the academic realm. As a result, this findings from this work also offered invaluable insights into the effectiveness of UAQTE as a policy, potentially guiding future policy adjustments to better align with the diverse needs of tertiary education in the Philippines. In essence, this study not only fill a critical void in the evaluation of UAQTE but also presented a robust framework that could potentially integrate quantitative and qualitative analysis as a precedent for future assessments of educational policies.

II. Methodology

This paper employed explainable machine learning to pinpoint the factors that best influence the satisfaction of the recipients of UAQTE. To add more context to these findings, topic modelling, which was used in earlier studies [1]-[3], was also utilized to understand the gaps that UAQTE may have. To achieve these objectives, the following regression models were employed: (1) k-Nearest Neighbors (kNN), (2) linear (LR), (3) decision tree (DT), (4) random forest (RF), and (5) gradient boosting machines (GBM). Explanations on how the best performing regression model predicts satisfaction of respondents with UAQTE was facilitated through the use of shap values as described in [4]-[7]. Moreover, topic modelling was done through the use of the open-source package BERTopic [7].

A. Development of survey instrument

A comprehensive consultation process involved approximately 10 pairs of parents, 8 faculty members, 40 tertiary students, and representatives from local universities and colleges (LUCs) and private schools. This inclusive approach comprised three rounds of focus-group discussions, ensuring diverse perspectives from key stakeholders, including implementers and school administrators, to enrich the research findings. The goal was to identify pertinent questions for evaluating the impact and effectiveness of the UAQTE implementation. Following the conclusion of these focused dialogues, a total of 30 questions were thoughtfully crafted based on the rich engagement with the stakeholders. Of these 24 questions, half of these were dedicated to identifying the demographics of the target respondents while the remaining questions were strategically formulated to delve into the assessment of the impact and efficacy of UAQTE on its recipients. In addition, within this framework, a Likert scale question was ultimately incorporated to prompt respondents to rate their satisfaction on a scale from 1 to 5 (1 being the lowest and 5 the highest). This addition aimed to provide a nuanced understanding of the sentiment among recipients regarding their satisfaction with the implementation of UAQTE.

B. Pilot-testing of developed survey

Before its full deployment, the developed survey underwent a pilot-testing phase to ensure its validity in capturing the sentiments of the intended respondents regarding the implementation of UAQTE. For this pilot test, 35 volunteer graduating tertiary students, all beneficiaries of UAQTE, were enlisted to evaluate the effectiveness of the developed survey. Feedback from the volunteers primarily centered on enhancing the clarity of the survey questions. Additionally, participants in the pilot test recommended the inclusion of open-ended questions aimed at eliciting suggestions on how to broaden the scope of UAQTE to better support struggling students. Following a process of iterative feedback collection from the pilot-test participants and subsequent revisions to the survey, the survey deployment proceeded once the final version met the necessary recommendations and was deemed satisfactory. The survey was conducted through the use of BOSESKO, a digital citizen participatory toolkit.

C. Model Training and Selection Phase

Five machine learning regressions models were used to develop a model that would extract patterns in the survey data (data frame with 1462 records containing 13 features, inclusive of the satisfaction ratings given by the respondents) that would be useful in assessing the impact and effectiveness of the implementation of UAQTE. Particularly, the kNN, LR, DT, RF, and GBM regressors, as detailed and implemented in [8]-[9] at default parameters, were adopted in this work. For the creation and training of these models, 75% of the survey data was designated as the training and validation set while the remaining 25% utilized were used as the holdout set. Lastly, the evaluation of model performance was based on the calculation of root-mean-squared errors for each model.

D. Model Interpretation Phase

To extract insights from the best performing regression model, this study used Shapley values. These values were instrumental in comprehending the influence of individual features within the trained model on the satisfaction levels of UAQTE recipients following the implementation framework as outlined in [4]-[7].

III. Result and Discussions

A. Development of survey instrument

This study focused on three key regions in the Philippines—specifically, CAR, NCR, and Region V—that are recognized to have the highest concentration of UAQTE recipients. Regrettably, owing to political unrest in the Visayas and Mindanao regions due to the barangay elections, this study narrowed its geographical scope to the Luzon archipelago alone to ensure the safety of the survey deployment and collection team. The demographic profile of the respondents is summarized in Table I. Overall, a total of 1462 respondents were surveyed, with 60% of the respondents being women. Notably, the predominant baccalaureate degree among respondents was identified as Bachelor of Science in

Nursing which falls under the science, technology, engineering, and mathematics (STEM) specialization.

TABLE 1. DEMOGRAPHICS OF RESPONDENTS

Demographics	Frequency
Average Age (years)	21.8 ± 1.17
Gender	
Man	408
Woman	897
LGBTQIA+	79
Prefer not to say	78
*Course**	
BS Secondary Education	118
BS Nursing	112
BS Entrepreneurship	64
BS Information Technology	58
BS Accountancy	56
BS Agribusiness	54
BS Psychology	50
BS Business Administration (Management)	46
BS Business Administration (Financial)	45
BS Biology	42

**Due to limited space, only the top 10 courses are shown.*

Table 2 presents the root-mean-squared-error of the trained regression models. Comparison of the training, validation, and holdout scores revealed that the decision tree regressor at a max depth of 4 gave the lowest holdout RMSE at 0.6881. Remarkably, the decision tree regressor consistently outperformed all other models including robust alternatives like the random forest and gradient boosting machines regressors across all scoring metrics. These compelling findings unequivocally establish the decision tree regressor as the optimal performer on the dataset. With these results, the decision tree regressor was chosen as the best performing model to be interpreted using Shapley values.

TABLE 2. PERFORMANCE OF TRAINED MODELS

Model	Training RMSE	Validation RMSE	Holdout RMSE
k-Nearest Neighbors	0.9398	0.9112	0.9447
Linear	0.9374	0.9304	0.9555
Random Forest	0.9644	0.9321	0.9332
Decision Tree	0.7230	0.7123	0.6881
Gradient Boosting Machnie	0.9714	0.9455	0.9543

In the context of machine learning models, lower RMSE values indicate better predictive performance. Based on the provided values, the Decision Tree model appears to be the most effective, as it consistently achieves the lowest RMSE across all three phases (training, validation, and holdout), making it the optimal performer in this study.

B. Interpretation of Best Performing Model

The visual representation of the impact of various features on the prediction of the best-performing model towards respondent satisfaction with UAQTE is shown in Figure 1 and Figure 2 (through bar and beeswarm plots). Specifically, it was found that the primary determinant of satisfaction with UAQTE implementation was the inclination of the UAQTE recipient towards engaging in international employment post-graduation. This discernible trend in the results indicated that the best performing model excels in predicting a satisfied recipient if there is a lack of desire to pursue international employment.

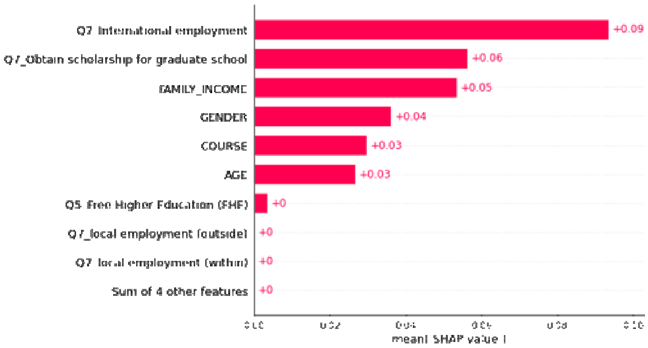


Figure 1. A sample line graph using colors which contrast well both on screen and on a black-and-white hardcopy

Conversely, dissatisfaction is predicted with higher accuracy when the respondent expresses a desire for overseas work. While this work acknowledges the need for further validation to establish this prospect causality, one plausible speculation is that UAQTE recipients may perceive the educational subsidy provided by the Philippine government as insufficient to confer a competitive advantage in securing employment opportunities abroad. For instance, the popularity of the BS Nursing course in the Philippines, known for its high employment prospects overseas, highly underscores this point.

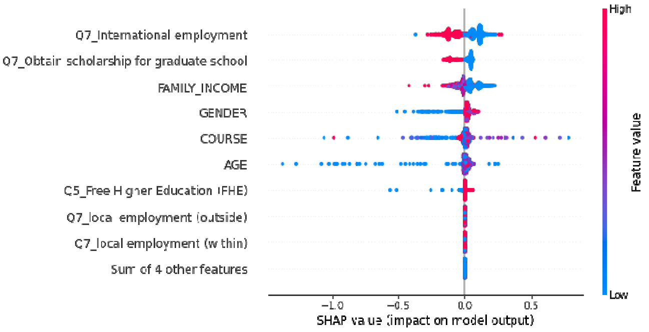


Figure 2. A sample line graph using colors which contrast well both on screen and on a black-and-white hardcopy

Given that UAQTE does not cover overhead costs associated with employment abroad, students aspiring to work internationally may seek additional subsidies to sustain themselves during the waiting period for overseas employment and to cover costs for processing their work visas.

This perceived situation suggests a nuanced interplay between the perceived adequacy of the subsidy and the specific career aspirations of UAQTE recipients especially to those with ambitions for international employment. Further exploration and validation of these nuanced dynamics could deepen our understanding of recipient satisfaction within the context of UAQTE implementation.

C. Results of Topic Modelling

The thematic analysis of the Universal Access to Quality Tertiary Education Act, depicted in Figure 3, highlighted four key themes from student feedback. In conjunction with Figures 1 and 2 which leveraged Shapley values for predictive modeling, it is evident that post-graduation employment preferences are pivotal to student satisfaction. Delving into these themes, "Financial Transparency and Concerns" (Topic 0) addresses the balance of gratitude for fee elimination and worry over hidden costs. "Infrastructure and Resource Adequacy" (Topic 1) reflects anxieties about the physical quality of educational facilities. "Emotive and Temporal Reactions" (Topic 2) capture the immediate feelings of students towards the implementation of the legislation, and "Beneficiary Experience and Challenges" (Topic 3) speaks to the personal hurdles students face despite being primary beneficiaries. Each theme identified contributed to a comprehensive understanding that while the Act is beneficial, it also surfaced areas for policy enhancement to align with preferred career trajectories of students with a noticeable emphasis on global employment aspirations.

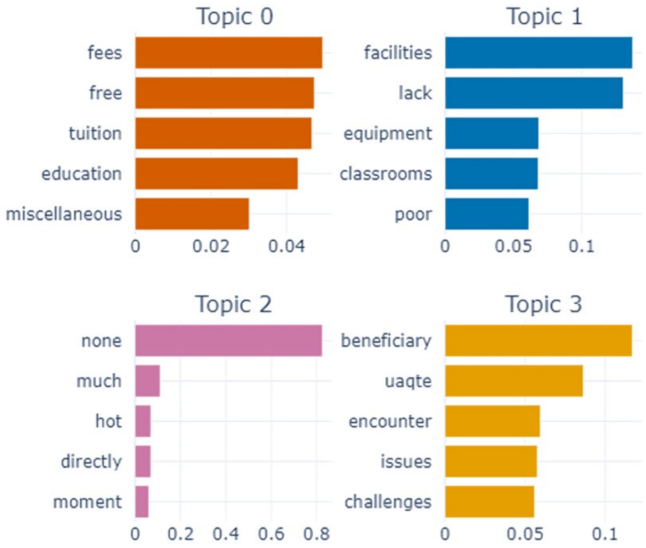


Figure 3. Topic Modelling Scores of the Top 4 Topics Observed.

D. Practical Implications of Findings

The revelation of this study that the inclination towards international employment significantly influences satisfaction of UAQTE recipients has noteworthy practical implications. Educational institutions and policymakers could use these insights to tailor support mechanisms, allocate resources more efficiently, and enhance career guidance services. For example, consider a scenario where a university that is based on the insights from this study, identifies a high proportion of engineering students within their UAQTE program expressing a desire for international employment post-graduation. Recognizing this trend, the university could tailor its career services to provide specialized guidance workshops for these engineering students that would offer (1) information on job markets abroad, (2) assistance in preparing for international job interviews, and (3) guidance on navigating the complexities of work visas. Simultaneously, the university might also allocate additional resources to create targeted networking events with global employers or industry professionals who can share insights about international career paths in this aspect. This approach not only addresses the specific needs of the students but also optimizes resource allocation to enhance their global employability and overall satisfaction within the UAQTE program. Policymakers may also consider adjustments to the UAQTE program to better align with the career goals of students aspiring to work abroad – specifically, to address the perceived inadequacy of educational subsidies for global competitiveness. Furthermore, the findings of this work highlight the importance of continuous research to ensure adaptive and responsive educational policies that support students in their diverse career aspirations which would ultimately contribute to their satisfaction and success in both local and international professional contexts.

IV. Conclusions

While the UAQTE Act is recognized as beneficial, this study underscores the need for policy enhancements to better align with students' preferred career trajectories, especially those aspiring to work globally. The findings emphasize the significance of understanding and addressing nuanced dynamics related to perceived subsidy adequacy and career aspirations to enhance recipient satisfaction within the UAQTE program. This research provides actionable insights for educational institutions and policymakers to refine support mechanisms and optimize resource allocation, fostering overall satisfaction and success for UAQTE recipients in both local and international professional contexts.

V. Future Work

A crucial avenue for future work would involve further validation of the identified correlation between satisfaction and the desire for international employment through longitudinal studies or qualitative assessments. Additionally, exploring the specific challenges faced by UAQTE recipients aspiring to work abroad, such as financial constraints or bureaucratic hurdles, would also provide a more nuanced understanding of their needs of the UAQTE recipient. Further investigations could also delve into the effectiveness of potential UAQTE policy adjustments to specifically examine whether tailored support mechanisms and enhanced career guidance would have a measurable impact on the satisfaction and success of students who have global career aspirations. As the educational landscape continues to evolve, continuous research efforts would be essential to ensure that policies such as UAQTE remain responsive to the dynamic needs of students to foster an environment that optimally supports diverse career goals within their respective specializations.

ACKNOWLEDGMENT

The authors gratefully acknowledge the generous support provided by Commission on Higher Education (CHED) through the LAKAS program. The authors would also like to extend their gratitude to Bicol University for the administrative support given during the conduct of this study. Moreover, the opinions expressed in this manuscript are solely those of the authors and do not reflect the position of CHED, Bicol University, and their respective offices.

REFERENCES

[1] Vayansky, Ike, and Sathish AP Kumar. "A review of topic modeling methods." Information Systems 94 (2020): 101582.

[2] Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv preprint arXiv:2203.05794 (2022).S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions." arXiv, 2017. doi: 10.48550/ARXIV.1705.07874.

[3] L. D. Austero, C. Y. Sy and M. J. P. Canon, "Discovering Themes from Online News Articles on the 2018 Mt. Mayon Eruption," 2018 International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 2018, pp. 242-245, doi: 10.1109/IS3C.2018.00068.

[4] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1. Springer Science and Business Media LLC, pp. 56–67, Jan. 17, 2020. doi: 10.1038/s42256-019-0138-9.

[5] R. Mitchell, E. Frank, and G. Holmes, "GPUTreeShap: Massively Parallel Exact Calculation of SHAP Scores for Tree Ensembles." arXiv, 2020. doi: 10.48550/ARXIV.2010.13972.

[6] S. M. Lundberg et al., "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," Nature Biomedical Engineering, vol. 2, no. 10. Springer Science and Business Media LLC, pp. 749–760, Oct. 10, 2018. doi: 10.1038/s41551-018-0304-0.

[7] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011. [Online]. Available: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

[8] P. Virtanen et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python," Nature Methods, vol. 17, no. 3. Springer Science and Business Media LLC, pp. 261–272, Feb. 03, 2020. doi: 10.1038/s41592-019-0686-2.

[9] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv, 2022. doi: 10.48550/ARXIV.2203.05794.

John Raymund B. Barajas has earned a Master of Science in Chemical Engineering from De La Salle University - Manila, as well as a Master of Science in Data Science from the Asian Institute of Management. His expertise spans a range of interdisciplinary fields, including data science, data analytics, and natural language processing. Furthermore, he possesses specialized knowledge in material science, with a particular focus on water and wastewater treatment research.

Lea D. Austero is an ASSISTANT PROFESSOR at Bicol University, Legazpi City, Albay, Philippines. She received her Doctor in Information Technology at the Technological Institute of the Philippines Quezon City just last May 2023. Her published works include Determining resource capacity in disaster assistance using a model-driven decision support system; Discovering themes from internet news articles on the 2018 Mount Mayon Eruption; and Solving course timetabling problem using Whale Optimization Algorithm.

Dr. Jennifer L. Llovido is a faculty member of the Computer Science and Information Technology Department at Bicol University College of Science, Legazpi City, Philippines, with an academic rank of Associate Professor V. She completed her Doctor in Information Technology (DIT) at the University of the Cordilleras, Baguio City, Philippines. Her published research works are centered on the fields of natural language processing, data mining, and system design and development. She can be reached at jllovido@bicol-u.edu.ph.

Lany L. Maceda earned her Doctorate in Information Technology from University of the Cordilleras, Baguio City, Philippines, in 2020. She is a faculty member of the Department of Computer Science and Information Technology, holding an academic rank of Associate Professor V at Bicol University. Moreover, she also serves as the Director of the Research, Development and Management Division at the same institution.

She has been actively promoting data-driven policy-making through her research papers published in reputable international journals and conferences with research interests on machine learning particularly on natural language processing and data mining. She can be reached at llmaceda@bicol-u.edu.ph.

Mideth B. Abisado is an Associate Member of the National Research Council of the Philippines and a Board Member of the Computing Society of the Philippines Special Interest Group for Women in Computing. She is the Director of the CCIT Graduate Programs. She completed her Doctor in Information Technology (DIT) at the Technological Institute of the Philippines. Her research focuses on Emphatic Computing, Social Computing, Human-Computer Interaction, and Human Language Technology. She can be reached at mbabisado@national-u.edu.ph.

Classifying gastric cancer carcinoma stages with deep semantic features and GLCM texture features

Sikandar Ali *, Samman Fatima *, Ali Hussain *, Maisam Ali*, Muhammad Yaseen*, Tagne Poupi Theodore Armand*, Hee-Cheol Kim**

* Dept. of Digital Anti-Aging Healthcare, Inje University, Gimhae 50834, Republic of Korea

** College of AI Convergence, Institute of Digital Anti-Aging Healthcare, u-AHRC, Inje University, Gimhae 50834, Korea

sikandarshigri77@gmail.com, samman.1511@gmail.com, alihussainnrana@gmail.com, maisamali053@gmail.com, shigriyaseen@gmail.com, poupiarmand2@gmail.com, heeki@inje.ac.kr

***Abstract*— Gastric cancer is one of the leading health issues that contributes to cancer related deaths. The tricky thing about cancer is that it often goes undetected until at higher stages, which makes treatment less effective. The significant death rate from gastric cancer highlights the importance of a precise and prompt diagnosis. This paper aims to tackle this problem by proposing an approach to classify the early and advanced stages of gastric cancer. This importance of this study stems from its two-pronged strategy, which provides a deeper understanding of stomach cancer stages using texture analysis and deep learning. We take advantage of the strengths of deep learning features, Gray Level Co-occurrence Matrix (GLCM) features, and machine learning algorithm to create a diagnostic tool that is more precise and accurate. Medical images from gastric cancer dataset showing early and advanced stages of gastric cancers carcinoma are included to develop this model. Our method combines the effectiveness of texture features extracted from GLCM combined with deep semantic features and classify the stages with machine learning model. We carefully evaluated Machine learning classifiers namely Support Vector Machine (SVM), Decision Tree (DT), and K-nearest neighbour (KNN) to classify the early and advanced stages. Each classifier was evaluated with different performance measures. The Support Vector Machine (SVM) classifier demonstrated the best performance with an accuracy of 96.93%. This highlights the potential of SVM for diagnosing different cancer stages, which could have positive implications, for clinical practice.**

***Keywords*— Gastric Cancer, Machine Learning, Support Vector Machine (SVM), Classification, GLCM (Gray Level Co-occurrence Matrix) Texture Features, deep semantic features**

I. Introduction

Gastric cancer is one of the significant global deaths causing cancer diseases. After lung cancer, it is the second leading cause of mortality from cancer [1, 2]. The cause is carcinomas (Stomach cancer with an adenocarcinoma origin in mucus-producing cells). Among stomach cancers, this one is the most prevalent. Carcinoma gastric cancers make up the majority of malignancies that develop in the stomach, which make up more than 90% of tumours. The typical 5-year survival rate is less than 20%, which is a poor prognosis primarily due to late detection since the early stages are clinically quiet. Only a few nations, notably Japan, have established large early detection programs. The 5-year survival rate can reach 90% if the tumour is found and treated before it invades the muscle layer of the stomach.[3]

Every year, more than 989,600 new diagnoses are made. 738,000 deaths are estimated to have occurred in 2008. 2012 saw an estimated 720 000 cases of stomach cancer fatalities [1]. There were 691 000 new instances of true gastric adenocarcinomas (non-cardia gastric cancers) in 2012, and 260 000 new cases of gastro-oesophageal junction adenocarcinomas (cardia gastric cancers), which are physically distinct. Despite significant improvements in our knowledge of epidemiology, pathology, molecular causes, and therapeutic options and tactics, as well as a drop in incidence and death, the burden of disease remains high [4]. Following lung and liver cancers, which, respectively, account for 23% and 28% of the total global burden of disability-adjusted life-years from cancer in men, gastric cancer accounts for 20% of the total [5]. Over 70% of new cases and deaths occur in developing countries. East Asia (Korea, Mongolia, Japan, and China) has the highest incidence rates, with yearly incidence rates ranging between 40 and 60 per 100,000 people. Contrary to the significantly lower rates recorded for the coastal and river valley regions, pockets of high risk are reported in Latin America's Andes Mountains, with rates between 20 and 30 per 100,000 [6]. Affluent populations in North America and Africa have lower rates (between 0.3 and 3 per 100,000). The rates of infection among African Americans are around twice as high as those among white Americans. In general, men have twice as many cases as women do. Different food habits, especially in European nations, and the incidence of Helicobacter pylori infection contribute to regional variations. The long-term low and steady prevalence of Helicobacter pylori infection in these nations likely explains this stalling of progress [7].

This study investigates to diagnose the early and advanced stages of gastric cancer using cutting edge Artificial intelligence-based techniques. Some of the researchers have used GLCM and deep semantic features for breast cancer diagnosis [8]. To improve diagnostic precision of early and advanced stages of gastric cancer carcinoma, we make also use of Gray-Level Co-occurrence Matrix (GLCM) texture features, deep semantic features and use machine learning classifier for the classification purpose. Our research hopes to aid in the diagnosis of stomach cancer, resulting in more efficient treatment plans and higher patient survival rates.

II. Materials and methods

In this section, we elucidate the materials and methods used in this research study which includes data collection, data processing and our proposed approach for gastric cancer classification model.

A. *Dataset Description*

In our study, we used a publicly available gastric cancer carcinoma dataset. The dataset contains more than 1800 patch images with image size 512 X 512. The dataset contains images of early stages and advances stages cancer. We chose 977 images of early and advanced stages.

B. *Data Preprocessing*

Data preparation is one of the key methods used in deep learning and machine learning. The degree to which the data has been pre-processed typically determines how accurate your model will be. In other words, the quality of your model is strongly correlated with the preprocessing of the data. We pre-processed the dataset for early and advanced stages. We used around 977 patch images of size 512 X 512. First, we convert the class labels into numerical labels representing 0 as early stage and 1 as advanced stage. The numerical representation is required for the model to be used during the training and prediction phases. This is an integral step while training the model where class labels are non-numeric. Second, we extracted Gray level co-occurrence Matrix (GLCM). Third, we applied CNN based feature extractor on the data and extracted deep semantic features from the data. then, we combined these two kinds of features. The dataset was split into 80 training and 20% testing and finally, we trained the machine learning classifiers for early and advanced stages of gastric cancer carcinoma.

C. *Proposed model*

Features play a vital role in the performance of any machine learning or deep learning model. In this study, we aim to develop a model by combining the gray level co-occurrence matrix features and deep semantic features. The nature of both features is different. Mostly researchers use Convolutional Neural Networks (CNN) models as feature extractors, doing so the texture features are get less attention and are extracted from shallow layers though they are also very important for model training. Both features i.e., the semantic features which comes from CNN based extractors and the GLCM features which are basically the texture features representing the shapes, edges of images etc, have their own importance. The CNN features, which are often called deep semantic features, contain more spatial

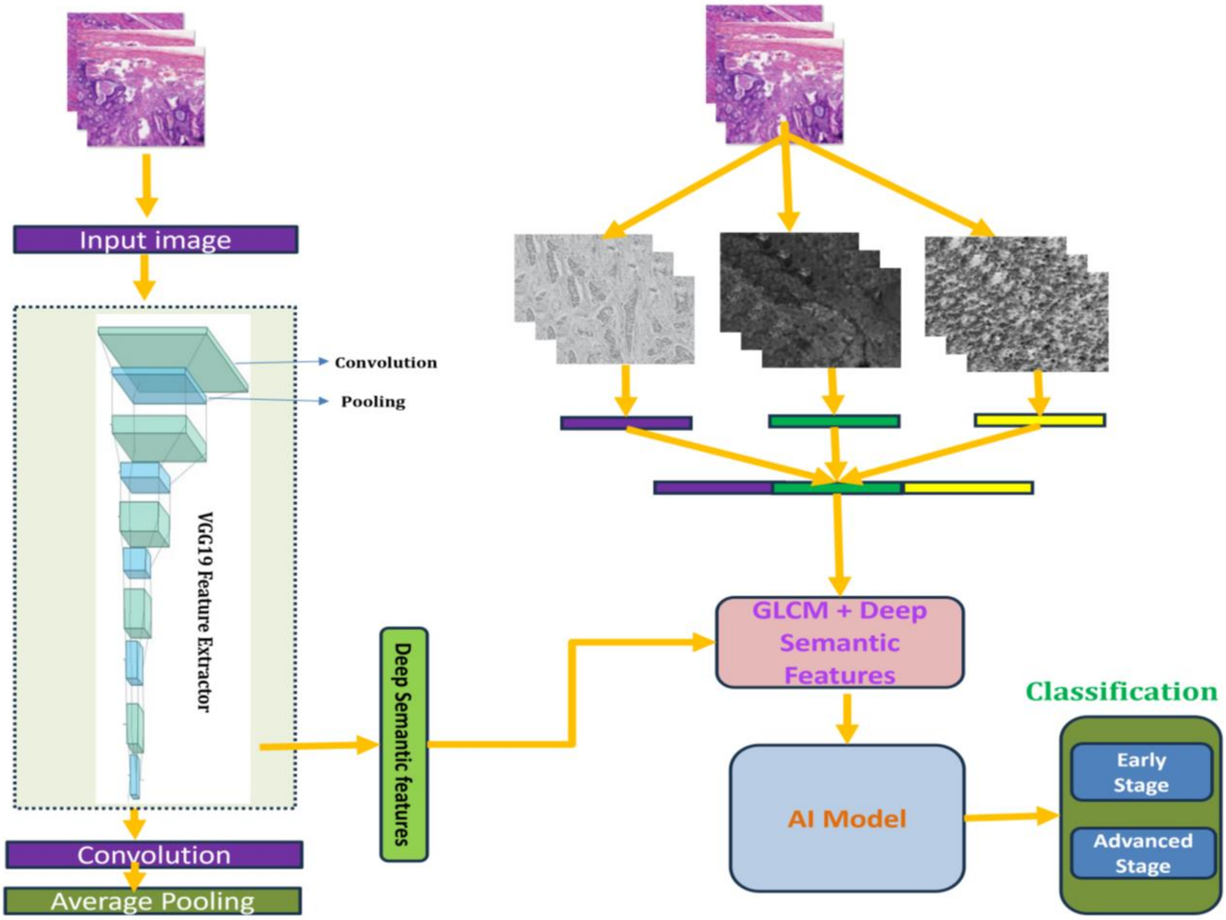


Figure 1. The architecture of proposed model.

information and gives more detailed information about the distribution of tissues. GLCM illustrates the texture of images. Heralick et al. [9] proposed GLCM to describe texture features. We calculated 15 different GLCM features including energy, contrast homogeneity, correlation, dissimilarity etc. For extracting deep semantic features, we used VGG19 and extracted 131087 features. We concatenated both the GLCM features and deep semantic features. Then we trained machine learning models with these combined features. The architecture of our proposed model has been shown in figure 1.

III. Experimental Results and discussion

This section explains about the experimental results about this research work. We extracted deep semantic features and GCLM features and concatenated both features and used three machine learning classifiers for the classification purpose of the different stages of gastric carcinoma. Our experimental results revealed that SVM outperformed other two models and demonstrated 96.93 % accuracy. Table 1 shows the performance metrics of all the applied models.

TABLE 1. Performance measures of the models

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
SVM	**96.93**	**96.01**	**96.96**	**96.05**
DT	94.89	94.56	94.01	94.86
KNN	94.16	94.00	94.23	94.11

Note: SVM denotes support vector machine; DT denotes decision tree and KNN denotes K nearest neighbours.

There are several metrics for the evaluation of the model. One of them is confusion matrix. Figure 2 shows the confusion matrix of our model showing the distribution of all the predicted responses to their true classes. Out of 196 test samples, 190 test samples as shown on the diagonal of the confusion matrix are correctly predicted and 6 samples are misclassified by the model. It means the model accuracy is quite encouraging for classifying the stages of gastric cancer carcinoma. Likewise figure 3 shows the ROC curve.

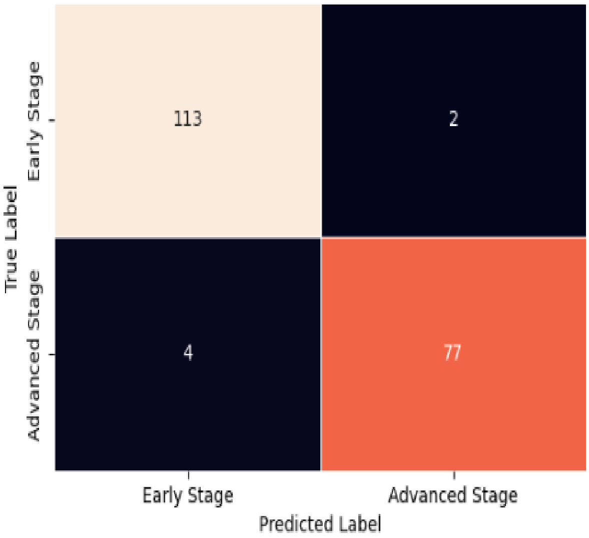


Figure 2. Confusion matrix of proposed model.

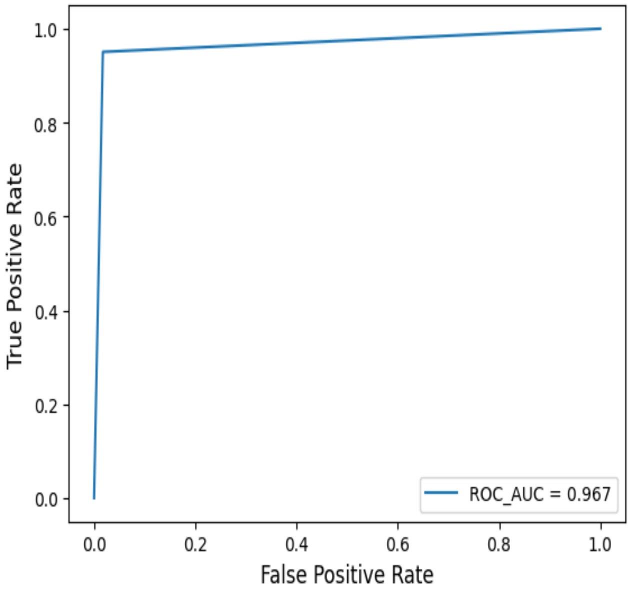


Figure 3. ROC curve of proposed model

Stomach cancer has been investigated by using different approaches over the years. For example Sakai al et.[10] proposed an automatic detection method using convolutional neural networks to help identify early stomach cancer in endoscopic images. Using two classes (cancer and normal) of image datasets with comprehensive texture information on lesions acquired from a limited number of annotated images, he performed transfer learning. The trained model's accuracy was 87.6%. Furthermore, he produced a heat map of unknown images that appeared to be a candidate zone of early stomach cancer. The accuracy of detection was 82.8%.

Sharma et al.[11] proposed a model where H&E stained histological whole slide images of gastric cancer are used to investigate deep learning techniques. In order to classify cancers based on immunohistochemical reaction and necrosis detection, a convolutional neural network is introduced. Handcrafted features are used to compare the deep learning methodology with conventional methods. The findings surpass conventional methods with an overall classification accuracy of 0.6990 for malignancy and 0.8144 for necrosis detection.

Li et al.[12] presented a study, a deep learning radiomics nomogram based on DECT which demonstrated high predictive value in identifying the location of lymph node metastases (LNM) in gastric cancer patients. This nomogram achieved a noteworthy accuracy of 0.77 (training set) and 0.76 (test set), outperforming both single-energy and clinical models. The best-performing model for the radiomics nomogram was chosen after a comparison of four models. The nomogram demonstrated strong predictive power for patient survival by including CT-reported lymph node status, radiomics signatures, and clinical data. This work enhanced the field of radiomics research on gastric cancer by offering a very precise method for prognosticating patients and predicting LNM.

Yang et al. [13] presented a study, based on publicly accessible gene expression data from TCGA-STAD, they compared and assessed three machine learning models to predict metastasis of Lymph nodes. The features were chosen based on their associations with the LN status using the Pearson correlation coefficient (PCC) technique. The accuracy and F1 score were used to evaluate the model's performance. Using 26 specifically selected gene features, the Naive Bayesian model performed better, with an accuracy of 0.72 in the test set and 0.741 in the training set. In both the training and test sets, the F1 score was 0.597 and 0.652, respectively.

Mortezagholi et al. [14] presented a comparative analysis between the two groups of participants the healthy and the sick, they chose 405 samples. 11 traits and risk variables in all were looked at. To categorize the patients with stomach cancer, they employed four machine learning techniques: k-nearest-neighborhood (KNN), naïve Bayesian model, decision tree (DT), and support vector machine (SVM). The accuracy rates of the SVM, DT, naïve Bayesian model, and KNN algorithms were 90.08, 87.89, 87.60, and 87.60 percent, respectively, based on the outcomes of the evaluation of the four approaches. The results demonstrated that the KNN algorithm (87.17) had the lowest rate and the SVM (91.99) had the highest level of F-Score.

IV. CONCLUSIONS

In the past few years, gastric cancer remained one of the world's most difficult health problems. Artificial technologies have been developed for the prediction, diagnosis, and prognosis of this disease aiming to address this issue. In this study, the early and advanced stages of gastric cancer carcinoma was investigated by combining GLCM features and deep semantic features. For the classification of the different cancer stages, machine learning classifiers were used while training the models with combined features. We employed Support vector machine (SVM), Decision Tree (DT), and K-nearest neighbour (KNN) machine learning models for this purpose to determine the cancer stage outcomes of gastric cancer patients. SVM outperformed other two classifiers demonstrated higher results for identifying gastric cancer carcinoma stages. This approach would aid in the exact and accurate diagnosis of the gastric cancer in clinical applications. This research study is a part of our major research project. We aim to investigate more approaches and techniques and explore the diagnosis of gastric cancer carcinoma.

ACKNOWLEDGMENT

This research was supported by a grant from Korea Health Technology R&D Project through the Korea Health Industry Development Institute validation (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant No: HI21C0977).

REFERENCES

[1] Ferlay, J., et al., Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. European journal of cancer, 2013. 49(6): p. 1374-1403.

[2] Jemal, A., et al., Global cancer statistics. CA: a cancer journal for clinicians, 2011. 61(2): p. 69-90.

[3] Miyahara, R., et al., Prevalence and prognosis of gastric cancer detected by screening in a large Japanese population: data from a single institute over 30 years. Journal of gastroenterology and hepatology, 2007. 22(9): p. 1435-1442.

[4] Colquhoun, A., et al., Global patterns of cardia and non-cardia gastric cancer incidence in 2012. Gut, 2015. 64(12): p. 1881-1888.

[5] Soerjomataram, I., et al., Global burden of cancer in 2008: a systematic analysis of disability-adjusted life-years in 12 world regions. The Lancet, 2012. 380(9856): p. 1840-1850.

[6] Correa, P., et al., Gastric cancer in Colombia. III. Natural history of precursor lesions. Journal of the National Cancer Institute, 1976. 57(5): p. 1027-1035.

[7] Ferro, A., et al., Worldwide trends in gastric cancer mortality (1980–2011), with predictions to 2015, and incidence by subtype. European journal of cancer, 2014. 50(7): p. 1330-1344.

[8] Hao, Yan, et al. "Breast cancer histopathological images classification based on deep semantic features and gray level co-occurrence matrix." Plos one 17.5 (2022): e0267955.

[9] Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics, 1973; 6, 610–621.

[10] Sakai, Y., et al. Automatic detection of early gastric cancer in endoscopic images using a transferring convolutional neural network. in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2018. IEEE.

[11] Sharma, H., et al., Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. Computerized Medical Imaging and Graphics, 2017. 61: p. 2-13.

[12] Li, J., et al., Dual-energy CT–based deep learning radiomics can improve lymph node metastasis risk prediction for gastric cancer. European Radiology, 2020. 30: p. 2324-2333.

[13] Yang, Y., et al., An immune-related gene panel for preoperative lymph node status evaluation in advanced gastric cancer. BioMed Research International, 2020. 2020.

[14] Mortezagholi, A., et al., Make intelligent of gastric cancer diagnosis error in Qazvin’s medical centers: Using data mining method. Asian Pacific journal of cancer prevention: APJCP, 2019. 20(9): p. 2607.

Sikandar Ali received his B.E. degree in Computer Engineering from Mehran University of Engineering & Technology, Pakistan. He got his MS from the Department of Computer Science from Chungbuk National University, the Republic of Korea. Furthermore, he is now a Ph.D. candidate at Inje University South Korea majoring in Artificial Intelligence in healthcare. His research interests include artificial intelligence, data science, big data, machine learning, deep learning, reinforcement learning, Computer vision, and medical imaging.

Samman Fatima received her bachelor's degree in software engineering from Fatima Jinnah women University Pakistan. Now she is pursuing her Master's degree from Inje University. Her research interests are medical imaging, artificial intelligence, Healthcare data, machine learning, deep learning.

Ali Hussain received the B.S. degree in computer science from Government College University Faisalabad (GCUF), Pakistan, in 2019. He is currently pursuing the master's degree with the Department of Digital Anti-Aging Healthcare, Inje University, South Korea. His research interests include artificial intelligence, data science, big data, machine learning, deep learning, computer vision, reinforcement learning, and medical imaging.

Maisam Ali received his B.E degree in Electrical and communication Engineering from Hamdard University, Pakistan. He is currently pursuing his master's degree from Inje University. His research interests are artificial intelligence, machine learning, deep learning, computer vision.

Muhammad Yaseen received his B.E degree in Electrical Engineering from Hamdard University, Pakistan. He is currently pursuing his master's degree from Inje University. His research interests are artificial intelligence, machine learning, deep learning, computer vision and medical imaging.

Tagne Poupi Theodore Armand was born in Cameroon, received Msc in information System and networking at ICT University USA, Cameroon Campus in 2021. Currently, he is a Ph.D. research scholar at the Institute of Digital Anti-aging and healthcare at Inje University. His research interest field includes image processing with a focus on medical image analysis, Deep Learning, Machine Learning and Metaverse.

Hee-Cheol Kim received his BSc at the Department of Mathematics, MSc at the Department of Computer Science in SoGang University in Korea, and Ph.D. at Numerical Analysis and Computing Science, Stockholm University in Sweden in 2001. He is a Professor at the Department of Computer Engineering and Head of the Institute of Digital Anti-aging Healthcare, Inje University in Korea. His research interests include machine learning, deep learning, Computer vision.

Enhanced Experiences: Benefits of AI-powered Recommendation Systems.

Kouayep Sonia Carole*, Tagne Poupi Theodore Armand**, Hee Cheol Kim*

* Institute of Digital Anti-Aging Healthcare, Inje University, Gimhae 50834, Republic of Korea
** Institute of Digital Anti-Aging Healthcare, Inje University, Gimhae 50834, Republic of Korea
carolesonia39@gmail.com, poupiarmand2@gmail.com, heeki@inje.ac.kr

***Abstract*— Today, information technology has brought various innovations and developments in almost every field of advancement. Recommendation Systems (RS) have achieved a significant milestone in the service business during the information systems era. Regarding online services, RS has been crucial in enhancing product availability and offering prospective customers a wider range of luxurious options. Conversely, online retailers face increasing their sales volume and achieving greater product prices than their rivals. One solution is to use recommendation systems that leverage artificial intelligence (AI) to provide personalized recommendations to users. These systems employ machine learning algorithms to examine user data, including search history, purchasing patterns, and preferences, to anticipate the products that consumers are most likely to be interested in. AI-powered recommendation systems have demonstrated their immense value as tools for decision-making, enhancing user experience, and fostering corporate success. This comprehensive review explores the multifaceted world of recommendation systems, delving into their mechanisms, applications, and transformative impact across diverse domains. From e-commerce to content streaming and beyond, these systems have redefined how we discover, choose, and engage with products, content, and services.**

***Keywords*— Artificial Intelligence (AI) collaborative filtering (CB), content-based recommendation, hybrid recommendation system, Recommendation System (RS), user-based recommendation**

I. Introduction

Recommendation systems emerged soon after the development of the World Wide Web, and related technologies were thoroughly studied and applied in both academic and corporate contexts. Recommendation systems are powerful web programs that provide daily suggestions for many types of material, such as news feeds, videos, e-commerce products, music, movies, books, games, friends, and work. They are offering services to an immense multitude of individuals, amounting to billions. Recommendation systems are artificial intelligence (AI) technologies that suggest items and services to users based on their interests and decisions. It assists users in discovering neglected products, provides personalized recommendations based on user preferences, and ultimately improves the efficiency and enjoyment of the shopping experience. This technology can significantly improve customer satisfaction and increase revenue by automating the search process and saving clients time. Recommendation systems utilize machine learning algorithms to offer personalized products, services, or information recommendations, considering users' behavior, interests, and history. According to statistics, 80% of customers are more likely to buy from a brand that provides a personalized experience.

Additionally, businesses that utilize recommendation engines experience a 150% surge in click-through rates, boosting sales and revenue. Netflix is an excellent example of a company that has harnessed the power of recommendation engines to revolutionize streaming. Netflix reduced its churn rate and annual savings of almost $1 billion by delivering customized content to its viewers.

AI-powered recommendation systems have become indispensable in the digital landscape as they facilitate customized experiences and shape user behavior. These systems use artificial intelligence to examine extensive datasets, detect user preferences, and provide customized recommendations. This paper provides an in-depth analysis of the fundamental elements of recommendation systems, elucidating their mechanisms and the various applications that derive advantages from their capabilities.

II. Types of Recommendation Systems

Recommender systems can be constructed using several methods, ranging from algorithmic and formulaic approaches to model-centric ones. The methodologies encompass page rank, collaborative filtering, content-based, and link prediction. However, it is important to note that complexity does not necessarily translate to better performance, and simple solutions and implementations often produce the best results. Establishing clear criteria for "good" recommendations is essential for assessing the effectiveness of the recommender system that has been constructed. The efficacy of a recommendation can be evaluated using diverse strategies that gauge its comprehensiveness and precision. Accuracy refers to the ratio of the right recommendations to the total number of possible recommendations.

In contrast, coverage quantifies the fraction of items in the search area that the system can provide recommendations for. The evaluation of suggestions is contingent exclusively upon the dataset and the methodology employed to generate these recommendations. Figure1 illustrates two distinct recommendation system approaches: one is a content-based strategy, while the other is a collaborative filtering method. A hybrid recommendation system is a modern recommendation [9].

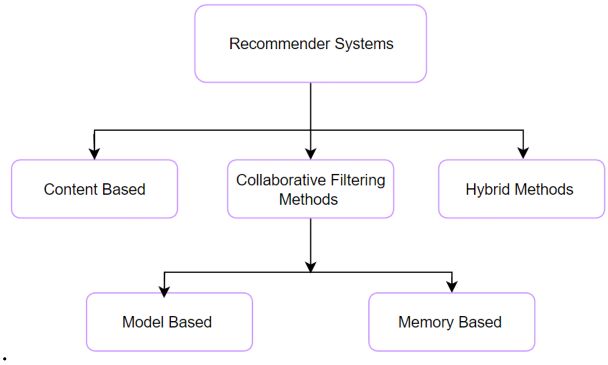


Figure 1. Different recommendation techniques.

A. *Collaborative filtering*

Collaborative filtering is a technique used in recommendation systems to predict user interests and preferences based on data and patterns from many users. The basic principle of collaborative filtering is that two users with similar preferences in one product are likely to have similar preferences in other products. The two primary categories of collaborative filtering methods are memory-based, which involves neighborhood computation, and model-based, which utilizes data mining techniques such as Bayesian networks, clustering, and semantic analysis

B. *Content-based*

The content-based approach suggests recommending an item to a user by considering their previous interactions with related items. A content recommender system generally consists of three main steps: user-profile generation, item-profile generation, and model-building. These steps are used to provide personalized recommendations for an active user. A content-based recommender system utilizes the characteristics or attributes of objects and user profiles to suggest items to users. Both the user and item qualities hold equal significance in creating a prediction. Let's take the example of a news recommender. To determine the similarity between news articles, it is necessary to consider features such as categories (Finance et al., Entertainment, etc.) and Location (local, national, or international).

C. *Hybrid recommendation*

In modern times, organizations employ a combination of many recommendation approaches, known as a hybrid recommendation technique [1]. The hybrid approach [2, 3] leverages the benefits of both content-based and collaborative filtering techniques, combining their efforts to develop a system that maximizes their individual strengths. The system employs accuracy metric methodologies to provide user-centered evaluation and give recommendations. It enhances the overall efficacy of the recommendation.

III. AI-POWERED RECOMMMENDED SYSTEMS

Recommendation systems use AI techniques to analyse data and make suggestions

A. *AI methods used in recommender systems*

Machine learning and deep learning techniques play the main role in recommender systems that help with prediction and recommendation. The primary determinant in recommender systems is the combination of information security and privacy [6]. The recommender system employs machine learning techniques, including Matrix factorization, singular value decomposition, and variable-weighted singular value decompositions.

1) Matrix Factorization

This method correlated both user and item to latent factors which help to hide the internal information behind the data [12]. Matrix factorization algorithms can be used to make predictions about user preferences by mapping users and items to these latent factors and using the relationships between the factors to make recommendations

2) Singular Value Decomposition

This fancy math method makes the big table from Matrix factorization easier to work with. It helps the computer to find pattern in the data even faster [13,14].

3) Variable Weighted

This method is the improved version of the Singular Decomposition method. The variable weight plays a main role in the method [10]

B. *Deep learning method used in recommender system*

Deep learning is the subset of machine learning field where different models of deep learning approach can be used in prediction and recommendation of recommender system such as Multilayer Perceptron (MLP), Restricted Boltzmann Machine (RBM), Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) [6,7,8]

a. MLP with auto-encoder

MLP is based on a feed forward neural network that uses multiple hidden layers. Each layer uses a different transfer function as activation function arbitrarily. Auto-encoder is a deep learning model, i.e., based in back propagation technique that finds the gradient of the error function w.r.t. the weights od the neural network

b. Convolution neural network

CNN is a deep learning model that based on feed-forward network. This neural network used a different convolution layer which finds the local and global features of the input data.

c. Restricted Boltzmann Machine

This model is usually working on probability distribution over the inputs, which consists of two layers such as visible and hidden layer [11]. No layer is communicated among the visible and hidden layer. A matrix of weights W represents the strength of the connections among hidden and visible units.

Performance of the model mainly depends on the number of hidden units

d. Adversarial networks

Adversarial Network is a kind of neural network, which is generative in nature and this network has two parts such discriminator and generator.

IV. BENEFITS OF USING AI-POWERED RECOMMENDED SYSTEMS

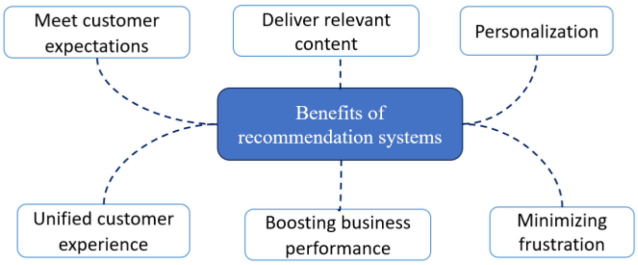

Meet customer expectations
Deliver relevant content
Personalization
Benefits of recommendation systems
Unified customer experience
Boosting business performance
Minimizing frustration

Figure 2. Benefits of Recommendation systems

1. PERSONALIZATION

In today's world of information overload, customers are inundated with choices and information, making it challenging to capture their attention. Personalization enables businesses to differentiate themselves by offering clients customized recommendations that align with their distinct requirements and preferences. Consequently, this can increase engagement, drive sales, and foster long-term customer loyalty. A recommendation engine is a very efficient tool for providing customized consumer experiences. A recommendation engine utilizes consumer data on prior behavior and preferences to propose products or services that are highly likely to interest the individual. This results in a smoother and more pleasant customer experience where customers feel understood, valued, and appreciated.

2. UNIFIED CUSTOMER EXPERIENCE ACROSS CHANNELS

An AI-driven recommendation engine has the potential to revolutionize the delivery of a uniform experience across diverse channels such as social media, websites, and mobile apps. This engine ensures uniformity by collecting and utilizing client interaction data from several channels. When a consumer searches for a product on your website, the recommendation engine can use this information to provide individualized product recommendations via email or other channels, creating a smooth and consistent user experience.

3. DELIVER RELEVANT CONTENT

In today's digital age, the delivery of pertinent content is paramount for engaging customers. Brands that provide customized content have a greater likelihood of attracting and retaining their client base. Recommendation engines are crucial in helping organizations achieve this goal by providing customized and relevant material to their target audience.

4. MINIMIZING FRUSTRATION IN THE CUSTOMER EXPERIENCE

An AI recommendation engine can reduce user frustration by offering customized and relevant content. The engine employs consumer behavior and preferences analysis to deliver tailored recommendations that effectively capture customer attention. This not only enhances the overall consumer experience but also results in increased sales and profitability for enterprises. By employing an AI recommendation engine, enterprises may prevent consumer dissatisfaction by delivering a tailored and engaging experience.

5. MEET CUSTOMER EXPERIENCE EXPECTATIONS

Providing personalized and seamless experiences is now more crucial than ever, emerging as a pivotal factor in building brand loyalty. By examining client behavior and preferences, these engines provide customized product recommendations that anticipate individual wants and aspirations. This tailored approach enhances the strength of customer connections and fosters enduring dedication. Furthermore, recommendation engines contribute to maintaining consistency across several touchpoints, ensuring a uniform experience from the website to social media and other channels. By offering customized and relevant content, businesses can engage customers on their preferred platforms, ensuring a seamless and uninterrupted experience throughout their whole contact.

6. BOOSTING BUSINESS PERFORMANCE WITH RECOMMENDATIONS ENGINES

In order to enhance the likelihood of a purchase, offering personalized product recommendations to your customers is recommended. Utilizing an AI-powered recommendation engine can optimize the efficiency of your up-selling and cross-selling strategies, simplifying the buying process and improving customer convenience.

V. AI-POWERED RECOMMENDATIONS SYSTEMS: ONLINE RESERVATION HEALTH-CARE: STUDY CASE

A recommendation engine utilizes machine learning and data analysis to generate customized recommendations. Data is the fundamental basis of a recommendation engine, as it supplies the essential information required for extracting patterns. This technique can potentially transform how people receive medical treatments by simplifying the procedure according to their symptoms and preferences. Here is a potential implementation of the concept:

1. **Data Collection**: AI-powered recommendation systems acquire user data through explicit methods, such as user ratings and comments, and implicit

ways, such as examining order history, return history, and search logs. After gathering substantial data, the system offers more relevant recommendations, enhancing the probability of attracting consumer interest. Individuals can enroll themselves to access the functionalities offered by the health interface.

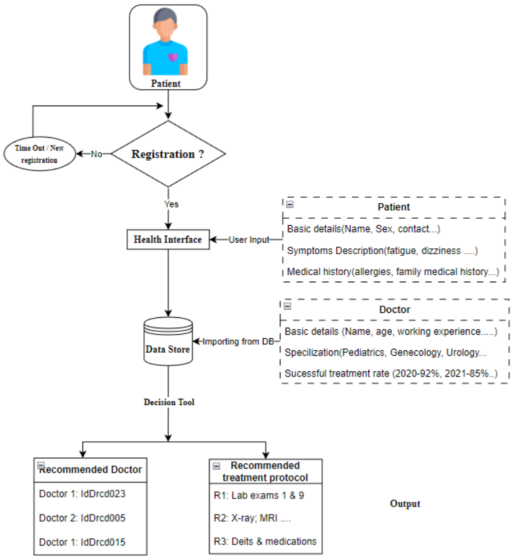

Patient
Registration ?
No
Time Out / New registration
Yes
Health Interface
User Input
Patient
Basic details(Name, Sex, contact...)
Symptoms Description(fatigue, dizziness ....)
Medical history(allergies, family medical history...)
Data Store
Importing from DB
Doctor
Basic details (Name, age, working experience.....)
Specilization(Pediatrics, Genecology, Urology...
Sucessful treatment rate (2020-92%, 2021-85%..)
Decision Tool
Recommended Doctor
Doctor 1: IdDrcd023
Doctor 2: IdDrcd005
Doctor 1: IdDrcd015
Recommended treatment protocol
R1: Lab exams 1 & 9
R2: X-ray; MRI ....
R3: Deits & medications
Output

Figure 3. Online reservation recommender system for health-care

2. **Data storage**: After a successful registration, individuals are given access to the health interface, where they can input their personal information, symptoms, and other pertinent details.
3. **Data Analysis**: The patient's specific information is stored in a data repository. Decision logic refers to making decisions based on various methods such as classification, fuzzy logic, IF-THEN rules, and decision trees. The system employs AI algorithms to analyze the symptoms the user provides, categorizing and evaluating them based on a database of medical conditions and their corresponding departments.
4. **Data filtration**: Once the system has accumulated and analyzed sufficient data to generate pertinent recommendations, the final stage involves filtering the data. Data can be filtered using various techniques, such as content-based, cluster-based, or collaborative filtering. The system utilizes symptom analysis to recommend the most suitable department(s) where the patient should seek medical care.

VI. Challenge in Recommendation Systems

Cold start problem: occurs when a system is used by a new client or when new items are introduced into the system.

Synonym: two or more different words which represents to same object or meaning is known as synonym. Recommendation algorithms cannot distinguish between these words. For instance, "comedy movies" and "comedy films" are perceived as distinct by a memory-based system—collaborative filtering technique.

Privacy: Sharing personal data with recommender systems might enhance system performance, but it poses risks to data privacy and security. Privacy concerns require Users to be more confident in data to recommendation systems. Hence, content-based and collaborative filtering recommender systems must establish trust among users. However, collaborative filtering recommender systems are particularly susceptible to privacy concerns. In collaborative filtering systems, user data, including reviews, is maintained in a central repository, which might be compromised and lead to data misuse.

Sparsity: occurs when a user-item metric becomes sparse due to a very big item set. It results in subpar recommendations for certain things. The recommendation of new things relies on the evaluation provided by previous users. The absence of a record of user preferences results in subpar recommendations for new users.

Latency problem: Collaborative filtering-based recommendation systems encounter latency problems when the knowledge base is often updated with new items. In such cases, the system promotes previously scored products, while newly added products have not yet been ranked. Using the CB filtering approach might decrease the waiting time for goods, while it may result in excessive specialization.

VII. Conclusions

In today's ever-changing business environment, companies face intense competition and must do everything they can to stay ahead of their competitors. Recommendation systems are a way to stay ahead of the curve and achieve larger business goals like increasing sales, ad revenue, and user engagement. However, to be successful with a recommendation system, you must carefully consider its need and agility. Agility is also essential for recommendation systems. Recommendation systems must adapt to remain relevant and effective as user behaviors, preferences, and needs change. Agile recommendation systems evolve, considering what works and what doesn't and additional data sources that help improve recommendations. By continually evaluating and enhancing AI- powered recommendation systems, businesses can achieve their goals while providing their users with the best possible experience.

Acknowledgment

This research was supported by the MSIT (Ministry of Science ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) in 2022 (2022-0-01091, 1711175863).

REFERENCES

[1] Erion, C. and Maurizio, M., Hybrid Recommender Systems: A Systematic Literature Review. Intell. Data Anal., 21, 6, 1487–1524, 2017.

[2] J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[3] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.

[4] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in *Proc. ECOC'00*, 2000, paper 11.3.4, p. 109.

[5] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[6] Cheng, H.T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., Anil, R., Wide & deep learn- ing for recommender systems, in: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, ACM, pp. 7–10, September, 2016.

[7] *Portugal, I., Alencar, P., Cowan, D., The use of machine learning algorithms in recommender systems: A systematic review. Expert Syst.* "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland *Appl., 97, 205– 227, 2018*.

[8] Mu, R., A survey of recommender systems based on deep learning. IEEE Access, 6, 69009–69022, 2018 J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[9] Zhong, Z. and Li, Y., A Recommender System for Healthcare Based on Human-Centric Modeling, in: 2016 IEEE 13th International Conference on e-Business Engineering (ICEBE), IEEE, pp. 282–286, 2, November, 2016.

[10] Wu, J., Yang, L., Li, Z., Variable weighted BSVD-based privacy-preserving collaborative filtering, in: 10th International Conference *on Intelligent Systems and Knowledge Engineering (ISKE), IEEE, pp. 144–148, November, 2015*.

[11] Yedder, H.B., Zakia, U., Ahmed, A., Trajković, L., Modeling prediction in recommender systems using restricted Boltzmann machine, in: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, pp. 2063–2068, October ,2017.

[12] Ortega, F., Hernando, A., Bobadilla, J., Kang, J.H., Recommending items to group of users using matrix factorization based collaborative filtering. Inform. Sci., 345, 313–324, 2016

[13] Ponnam, L.T., Punyasamudram, S.D., Nallagulla, S.N., Yellamati, S., Movie recommender system using item based collaborative filtering technique, in: 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), IEEE, 1–5, 2016

[14] Sahoo, A.K., Pradhan, C., Mishra, B.S.P., SVD based Privacy Preserving Recommendation Model using Optimized Hybrid Item-based Collaborative Filtering, in: 2019 International Conference on Communication and Signal Processing (ICCSP), IEEE, pp. 0294–0298, April, 2019

Kouayep Sonia Carole is Ph.D. candidate in digital anti-aging and healthcare at Inje University, South Korea. She received her Master at the Department of Information Technology (IT) and applications Engineering at Pukyong University in Busan, South Korea. Previously, she earned her B.S degree in Computer Science from Dschang in Cameroon. Her research interests include image processing, computer vision, Artificial intelligence and Business intelligence

Tagne Poupi Theodore Armand is a Ph.D. research scholar at the Institute of Digital Anti-aging and healthcare at Inje University. He received his M.Sc. in information System and Networking at ICT University USA, Cameroon Campus. His research interest field includes Artificial Intelligence in healthcare; image processing with a focus on medical image analysis, Deep Learning, Machine Learning, and Metaverse.

Hee-Cheol Kim received his BSc at the Department of Mathematics, MSc at the Department of Computer Science at SoGang University in Korea, and Ph.D. in Numerical Analysis and Computing Science at Stockholm University in Sweden in 2001. He is a professor at the Department of Computer Engineering and Head of the Institute of Digital Anti-aging Healthcare Inje University in Korea. His research interests include machine learning, deep learning, Computer vision, and medical informatics.

Session 3C: System, Software Engineering

Chair: Dr. Hyunho PARK, Electronics Telecommunications Research Institute (ETRI), Korea

1 Paper ID: 20240330, 221~225

Evaluation of |Y> Magic State Distillation Circuit

Mr. Youngchul Kim, Dr. Soo-Cheol Oh, Ms. Sangmin Lee, Mr. Ki-Sung Jin, Mr. Gyuil Cha,

ETRI. Korea(South)

2 Paper ID: 20240293, 226~231

DB Workload Management through Characterization and Idleness Detection

Dr. Abdul Mateen, Mr. Khawaja Tahir Mahmood, Dr. Seung Yeob Nam,

Federal Urdu University of Arts, Science & Technol. Pakistan

3 Paper ID: 20240417, 232~240

Micro-services internal load balancing for Ultra Reliable Low Latency 5G Online charging system

Mr. Ngoc Tien Nguyen, Mr. Thanh Son Pham, Mr. Van Duong Nguyen, Dr. Cong Dan Pham, Mr. Duc Hai Nguyen,

Viettel High Technology. Viet Nam

4 Paper ID: 20240471, 241~247

AppTest: Assessing the Usability and Performance Efficiency of BOSESKO for Digital Participation

Dr. Jennifer Llovido, Dr. Michael Angelo Brogada, Dr. Lany Maceda, Dr. Mideth Abisado,

Bicol University. Philippines

5 Paper ID: 20240456, 248~251

Research on the transformation path of DevOps in the Digital Era

Ms. XIAOLING NIU, Ms. LINGLING YANG, Ms. KAILING LIU, Mr. ZHAOWEI LIU,

The China Academy of Information and Communication. China

Evaluation of |*Y*> Magic State Distillation Circuit

Youngchul Kim*, Soo-Cheol Oh*, Sangmin Lee*, Ki-Sung Jin*, Gyuil Cha*

*Future Computing Research Division, ETRI, Daejeon, Republic of Korea

kimyc@etri.re.kr, ponylife@etri.re.kr, sanglee@etri.re.kr, ksjin@etri.re.kr, gicha@etri.re.kr

***Abstract*— For a universal quantum computer, surface code-protected logical Clifford and non-Clifford gates must be supported fault-tolerantly. However, to implement non-Clifford gates, magic states are required, and since these magic states are faulty, distillation circuits are used to obtain high-fidelity magic states by utilizing multiple low-fidelity states. It is not easy to implement and simulate the operations of a distillation circuit because it requires many resources. This paper presents a resource-efficient implementation and evaluation of a |*Y*> magic state distillation circuit on a quantum simulator.**

***Keywords*— fault-tolerant quantum computing, quantum error correction, surface code, lattice surgery, magic state distillation**

I. Introduction

Developing a fault-tolerant quantum computer with too noisy qubits and gates on a quantum device is challenging. Much research is on quantum error correction to support fault tolerance with topological codes such as surface code [1][2]. Surface code provides the most prominent quantum error correction method to implement fault-tolerant quantum computation due to its high error threshold of around 1% and simple two-dimensional structure with only nearest-neighbour connectivity [1]. However, it takes many physical qubits to encode a logical qubit with surface code. Also, the surface code should be encoded with more physical qubits to provide a lower error rate. Various techniques, such as braiding [1], lattice surgery [2][3], and twist [4], can implement logical operations between logical qubits protected by surface code. Out of these techniques, lattice surgery can reduce cost and complexity while maintaining two-dimensional layout.

We need to provide a quantum computer that supports universal quantum gates such as Clifford and non-Clifford gates to reap the benefits of quantum computing. Clifford gates, including Pauli gates, can be effectively simulated on a classical computer [7] and implemented by lattice surgery. However, it takes complex work to implement non-Clifford gates logically. In particular, the logical *S* and *T* gates require the magic state that can be prepared with the state injection process. Since state injection is not fault-tolerant, the injected state has low fidelity and needs to be distilled by the magic state distillation procedure. However, obtaining a higher-fidelity magic state from multiple lower-fidelity states requires many resources and time. Previous studies [3][6] have proposed several magic state distillation protocols and have calculated their costs numerically.

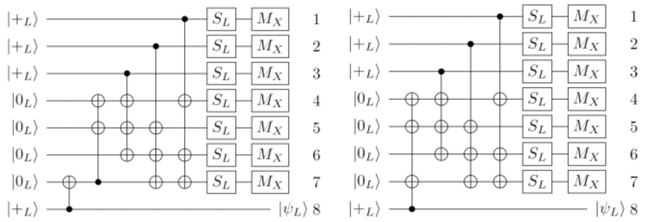
The common distillation protocols are to implement the 7-qubit Steane code for the magic state $|Y\rangle = \frac{1}{\sqrt{2}}(|0\rangle + i|1\rangle)$ consumed at the logical *S* gate to purify the state and the 15-qubit Reed-Muller code for the magic state $|A\rangle = \frac{1}{\sqrt{2}}(|0\rangle + e^{i\frac{\pi}{4}}|1\rangle)$ consumed at the logical *T* gate [1].

Using the quantum simulator QPlayer [8][9], we have implemented the |*Y*> state distillation circuit in two different ways: a multi-target CNOT implementation in lattice surgery and an implementation based on lattice surgery translation [5], evaluated the cost of the resources used in them. Here, we considered the placement of logical qubits to minimize the resources required to perform multi-target CNOT in lattice surgery.

This paper is organized as follows. In section II, we first introduce the multi-target CNOT-based distillation and lattice surgery translation-based distillation process for |*Y*> state and explain how |*Y*> state is purified through these processes. Then, in section III, we implemented the above circuits on different placements of logical qubits and evaluated the cost of the resources.

II. Distillation processes for |*Y*> state

We need a high-fidelity magic state |*Y*> to implement a logical *S* gate using gate teleportation. A magic state is first created by a process called state injection. However, the injected state can be imperfect, so it is then purified to a high-fidelity state using a distillation process.



(a) (b)

Figure 1. |*Y*> state distillation circuit implementing 7-qubit Steane code.

A. *Multi-target CNOT-based |Y> state distillation*

The distillation circuit implementing 7-qubit Steane code for |*Y*> magic state is shown in Figure 1. This circuit consists of initializing |0> and |+> states, multi-target CNOTs, logical *S* gate teleportation, and *X*-basis measurement. Figure 1(b) circuit is translated from Figure 1(a) according to the commutation relations between different non-commuting

CNOTs to minimize the number of CNOTs and implement the lattice surgery translation-based technique.

For |*Y*> state distillation circuit, it first creates a logical Bell pair. Then, one logical qubit from the Bell pair with the other six logical qubits is encoded in Steane code to entangle eight logical qubits using the multi-target CNOTs. After that, seven encoded qubits are each rotated with an *S* gate using |*Y*> state, which is injected in the first round of distillation or is distilled in a prior round. The logical *S* gate can be implemented with lattice surgery-based operations as shown in Figure 3(b).

The distillation circuit produces a purified |*Y*> state from the other logical qubit of the Bell pair. This procedure is performed iteratively to obtain a higher fidelity output state based on the *X*-basis measurement outcomes of the seven qubits. The output state is interpreted by evaluating the eigenvalues of the Steane code *X* stabilizers such as $S_1=M_{X1}M_{X4}M_{X6}M_{X7}$, $S_2=M_{X2}M_{X5}M_{X6}M_{X7}$, $S_3=M_{X3}M_{X4}M_{X5}M_{X6}$. If the eigenvalues of all three stabilizers are +1, the output is purified |*Y*> state and will be kept. Otherwise, the state is discarded, and the subsequent distillation round must be performed. Furthermore, if the product of all measurements is 1, the output state will include a *Z*-error so that it needs to be tracked.

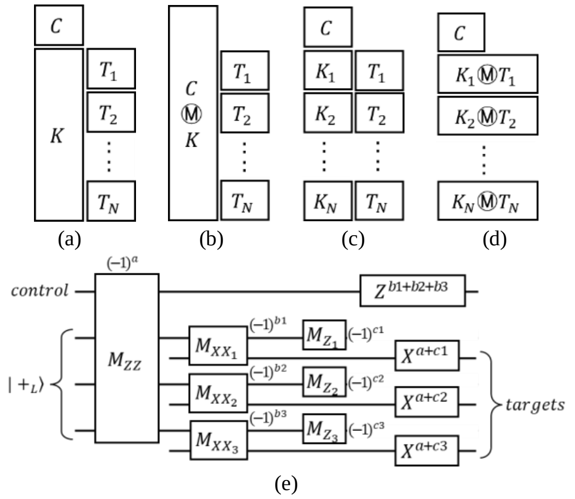


Figure 2. Multi-target CNOT in lattice surgery. *C* and $T_{1..N}$ refer to control and target qubits. *K* indicates the vertical rectangular ancilla qubits. (a) The vertical rectangular ancilla qubit is initialized to the |+> state. (b) The control qubit performs logical joint measurements, M_{ZZ}, with the vertical rectangular ancilla qubit. (c)(d) It is then split to perform logical joint measurements, M_{XX}, with the target qubits. (e) Lattice surgery-based multi-target CNOT circuit that implements the previously outlined the multi-target CNOT.

Lattice surgery logical operations between logical qubits allow us to perform state teleportation and gate teleportation of logical qubits, as shown in Figures 3(a) and 3(b), respectively.

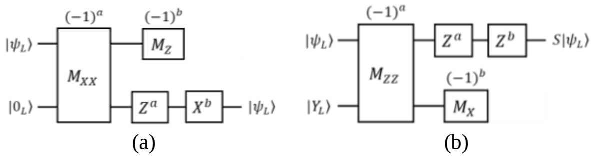


Figure 3. State teleportation and logical *S* gate in lattice surgery.

B. Lattice surgery translation-based |Y> state distillation

Figure 4 shows that CNOTs with different controls and the same target can be implemented using lattice surgery. It first initializes the two patches corresponding to the control qubits of each CNOT with |+> in Figure 4(b). It then performs a smooth split operation on each patch in Figure 4(b) and a rough merge operation on the two split patches. Eventually, the state of the three patches will be the same as the result in Figure 4(a) [5].

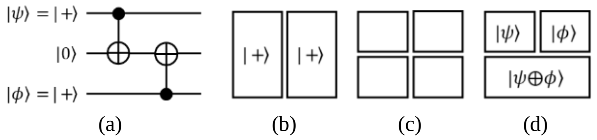

$|\psi\rangle = |+\rangle$
$|0\rangle$
$|\phi\rangle = |+\rangle$
$|+\rangle$
$|+\rangle$
$|\psi\rangle$
$|\phi\rangle$
$|\psi \oplus \phi\rangle$
(a)
(b)
(c)
(d)

Figure 4. Implementation of CNOTs with the same target and different controls to extend to multi-target CNOTs using lattice surgery.

Lattice surgery-based CNOT with different controls and the same target can be extended to a |*Y*> state distillation circuit consisting of multi-target CNOTs, as shown in Figure 1(b).

III. Evaluation of |Y> State Distillation Processes

In this section, we have implemented the |*Y*> state distillation circuits mentioned in Section II using the QPlayer quantum simulator and evaluated each circuit.

A. Logical qubit architecture

For this work, we have used surface code logical qubits with a distance of 2, as shown in Figure 5(a) [10]. Lattice surgery merging and splitting operations along the *X*(*Z*)-boundary between two logical qubits are shown in Figures 5(b) and 5(c).

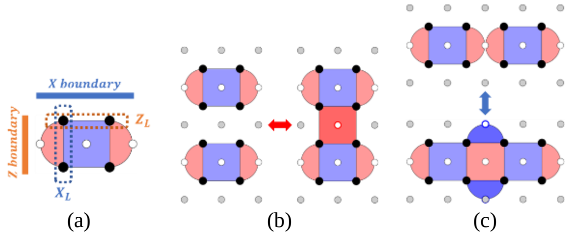


Figure 5. Layout of surface code logical qubit and lattice surgery between two logical qubits. (a) Surface code logical qubit with distance 2. Physical data qubits are represented by black circles, and ancilla qubits by white circles. The purple(pink) plaquettes represent the *X*(*Z*) stabilizers, respectively. (b) Logical joint measurements along the *X*-boundary, M_{ZZ}. (c) Logical joint measurements along the *Z*-boundary, M_{XX}.

The |*Y*> state distillation circuits have been implemented in three logical qubit architectures, as shown in Figure 6. The logical qubit architecture with 24 logical qubits in Figure 6(a) and the logical qubit architecture with 18 logical qubits in Figure 6(b) perform a multi-target CNOT-based |*Y*> state distillation circuit. The logical qubit architecture with 20 logical qubits in Figure 6(c) performs a lattice surgery-based |*Y*> state distillation circuit. Logical operations of |*Y*> state distillation circuits do not use some logical qubits in the logical qubit architectures, but we consider a checkerboard form for simplicity.

The two logical qubits must be placed in the nearest neighboring locations to perform lattice surgery merging and splitting operations, which involve multi-target CNOT, state teleportation, and gate teleportation in a logical qubit architecture. Therefore, it is essential to properly arrange the control and target qubits for multi-target CNOT in the |*Y*> state distillation circuit.

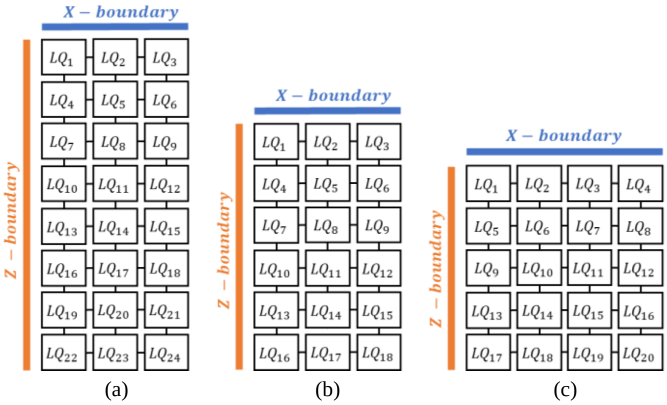


Figure 6. Logical qubit architectures for implementing |*Y*> state distillation circuits.

B. Multi-target CNOT-based |Y> state distillation

We have implemented the multi-target CNOT-based |*Y*> state distillation circuit on the logical qubit architecture in Figures 6(a) and 6(b), arranged in three forms: MT-CNOT-1, MT-CNOT-2, and MT-CNOT-3, respectively, as shown in Figure 7. In Figure 7, the patch numbers denote the logical qubits in the circuit, and *K* indicates the vertical rectangular ancilla qubits required to perform the lattice surgery-based multi-target CNOTs of Figure 2.

In Figure 7(MT-CNOT-1), the vertical rectangular ancilla qubit is initialized to the |+> state first and performs logical joint measurements, M_{ZZ}, with the control qubits 1, 2, 3, and 8. It is then split and shrunk to perform logical joint measurements, M_{XX}, with the target qubits 4, 5, 6, and 7. After the multi-target CNOTs, we inject the |*Y*> state into the neighboring qubits and perform a logical *S* gate, followed by an *X*-basis measurement, as shown in Figure 1(b). However, since a vertical rectangular ancilla qubit consisting of up to six patches is required to perform multi-target CNOTs and too much memory is required to organize it, this paper excludes it from the list of possible implementations and only considers logical qubit placement and computation.

Therefore, in Figure 7(MT-CNOT-2), the size of the vertical rectangular ancilla qubit is reduced to four patches to minimize the amount of memory required to construct the vertical rectangular ancilla qubit. To this end, the control qubits 1 and 8 locations are adjusted, as shown in Figure 7(MT-CNOT-2(a)). In this placement, the distillation circuit's first and fourth multi-target CNOTs are performed by decomposing them into single-target CNOTs and multi-target CNOTs, respectively. For the first multi-target CNOT, the ancilla qubit H_8 is first used to perform the CNOT with target qubit 7, and the right vertical rectangular ancilla qubit *K* is used to perform the CNOT with target qubits 4 and 5. For the fourth multi-target CNOT, the ancilla qubit H_1 is used to perform the CNOT with target qubit 4, and the left vertical rectangular qubit *K* is used to perform the CNOT with target qubits 6 and 7.

In Figure 7(MT-CNOT-3), the logical qubit architecture of Figure 6(b) is used to optimize space for ancilla qubits for multi-target CNOT operations and |*Y*> state injected qubit for logical *S* gates. In Figure 7(MT-CNOT-3(a)), control qubits 1, 2, 3, and 8 perform CNOT with target qubits 4, 5, 6, and 7 using the left and right vertical rectangular ancilla qubits *K*. The vertical rectangular ancilla qubit, split to perform logical joint measurement M_{XX} with the target qubits, is used as a space to inject the |*Y*> state for logical *S* gates. Logical qubits 5 and 6 move to perform the logical joint measurement, M_{ZZ}, with the neighboring qubit where the |*Y*> state was injected.

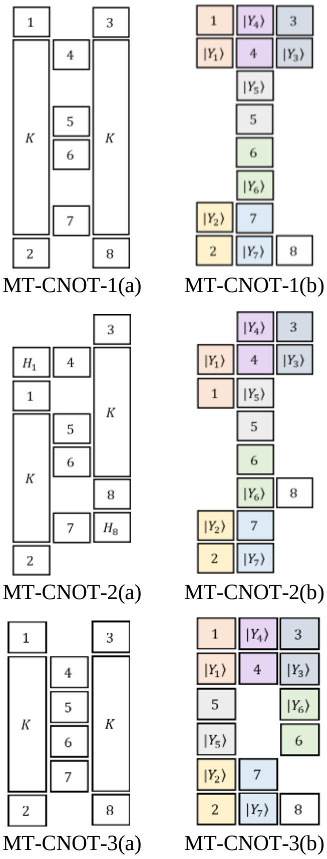


Figure 7. This illustrates how |*Y*> state distillation circuits are implemented with multi-target CNOTs.

C. Lattice surgery translation-based |Y> state distillation

The lattice surgery translation-based distillation process [5] is implemented in a logical qubit architecture consisting of the 20 logical qubits of Figure 6(c), as illustrated in Figure 8. The lattice surgery translation-based distillation process is performed as follows. Firstly, for each multi-target CNOT in the circuit in Figure 1(b), the vertical rectangular patches are initialized with the |+> state in Figure 8(a). Each vertical rectangular patch is then smoothly split in Figure 8(b). The numbers in Figure 8(b) indicate which patch contributes to which qubit of the circuit in Figure 1(b). In Figure 8(c), patches with different control qubits and the identical target qubits are moved to neighboring locations, and rough merge operations are performed. And then, the rough-merged patches are reduced to a single patch size. We inject |*Y*> state into the patches adjacent to the logical qubit patch to perform the logical S gate. We merge smoothly with the |*Y*> state-injected patch and take *X*-basis measurements of logical qubits 1 through 7. Therefore, according to the measurement outcomes, logical qubit 8 has the distilled state.

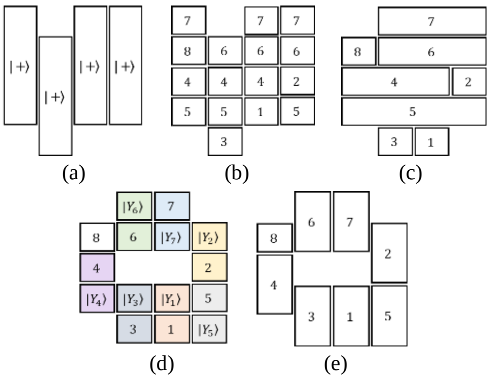


Figure 8. This illustrates how |*Y*> state distillation circuits are implemented using lattice surgery translation [5].

We have evaluated the space-time cost of the implementation methods for |*Y*> state distillation circuit in Table 1. Since the surface code patch with distance-2 used in this paper requires d^2 data qubits and (d^2 - 1) ancilla qubits, the total physical qubits are $2d^2$ to the leading order. Thus, the space cost requirements are equal to the number of logical qubits in the logical qubit architecture multiplied by the number of physical qubits for one patch.

The time cost requirements are given by code-cycles. The implementations based on multi-target CNOTs have the following time costs.

- [2*d*] Initialization of the vertical rectangular ancilla qubits with states |+> for smooth merging and splitting of control qubits and vertical rectangular ancilla qubits, twice each.
- [2*d*] Smooth merging and splitting for CNOT(1, 4) and CNOT(8, 7) (for MT-CNOT-2 only).
- [2*d*] Rough merging and splitting for CNOT(1, 4) and CNOT(8, 7) (for MT-CNOT-2 only).
- [4*d*] Smooth merging and splitting of control qubits and vertical rectangular ancilla qubits.
- [2*d*] Shrinking the split ancilla qubits.
- [2*d*] Rough merging and splitting of target qubits and split ancilla qubits.
- [1*d*] Movement of qubits 5 and 6 (for MT-CNOT-3 only).
- [1*d*] Injection of |*Y*> states.
- [1*d*] Smooth merging with |*Y*> state injected qubits.

The lattice surgery translation-based implementation has the following time costs [5].

- [1*d*] Initialization of the vertical rectangular qubits with states |+>.
- [1*d*] Smooth splitting of the vertical rectangular qubits.
- [1*d*] Movement of qubits 1 and 7.
- [1*d*] Rough merging of the same target qubits.
- [1*d*] Shrinking the merged qubits.
- [1*d*] Injection of |*Y*> states.
- [1*d*] Smooth merging with |*Y*> state injected qubits.

Although the time cost of the lattice surgery translation-based implementation is less than the multi-target CNOT-based one, we can simulate the |*Y*> state distillation circuit using the multi-target CNOT-based implementation with fewer logical qubits.

TABLE 1. EVALUATION OF THE SPACE-TIME COST OF THE IMPLEMENTATION METHODS FOR |*Y*> STATE DISTILLATION CIRCUIT

Cost	Implementation methods for |*Y*> state distillation		
	MT-CNOT-2	**MT-CNOT-3**	**LST**
Space	24 x 2 x d^2	18 x 2 x d^2	20 x 2 x d^2
Time	16*d*	13*d*	7*d*
Total	$768d^3$	$468d^3$	$280d^3$

IV. CONCLUSIONS

We have implemented and evaluated the |*Y*> state distillation circuits using the QPlayer quantum simulator on logical qubit architectures. The logical qubits are encoded in a surface code with distance 2. The logical qubits are arranged to use fewer resources to perform the multi-target CNOTs comprising the |*Y*> state distillation circuit. We also have implemented and compared how CNOT operations with different controls and the same targets can achieve the same results using lattice surgery-based methods and extend them to multi-target CNOT. The |*Y*> state distillation circuit requires many logical qubits and, therefore, many physical qubits, which is challenging to implement in practice and complex to verify by simulation. This paper verifies |*Y*> state distillation using more efficient resources. As a next step, we will evaluate |*A*> state distillation circuits for logical *T* gate.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-00014), A Technology Development of Quantum OS for Fault-tolerant Logical Qubit Computing Environment).

REFERENCES

[1] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, “Surface codes: Towards practical large-scale quantum computation,” Physical Review A. 86, 2012.

[2] C. Horsman, A. G. Fowler, S. Devitt, and R. V. Meter, “Surface code quantum computing by lattice surgery,” New Journal of Physics. 14, 2012.

[3] D. Litinski, “A Game of Surface Codes: Large-Scale Quantum Computing with Lattice Surgery,” Quantum 3. 128, 2019.

[4] D. Litinski and F. V. Oppen, “Lattice Surgery with a Twist: Simplifying Clifford Gates of Surface Codes,” Quantum 2. 62, 2017.

[5] D. Herr, F. Nori, and S. Devitt, “Lattice surgery translation for quantum computation,” New Journal of Physics. 19, 2017.

[6] S. Bravyi and A. Kitaev, “Universal quantum computation with ideal Clifford gates and noisy ancillas,” Physical Review A. 71, 2005.

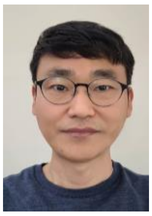
[7] D. Gottesman, “The Heisenberg Representation of Quantum computers,” arXiv:quant-ph/9807006, 1998.

[8] K. S. Jin and G. I. Cha, “QPlayer: Lightweight, scalable, and fast quantum simulator,” ETRI Journal. 45, (2023), 304–317.

[9] K. S. Jin and G. I. Cha, “Multilayered logical qubits and synthesized quantum bits,” Quantum Science and Technology. 8 035008, 2023.

[10] A. Erhard, H. P. Nautrup, M. Meth, L. Postler, R. Stricker, M. Stadler, V. Negnevitsky, M. Ringbauer, P. Schindler, H. J. Briegel, R. Blatt, N. Friis, and T. Monz, “Entangling logical qubits with lattice surgery,” Nature. 589, 2021.

Youngchul Kim is currently a principal researcher at ETRI, Daejeon, Republic of Korea since 2000. He received his BS and MS degrees in Computer Science from the Kangwon National University, Chuncheon, Republic of Korea in 1996 and 1999, respectively. His research interests include fault-tolerant quantum computing, distributed systems, and cloud systems.



Soo-Cheol Oh received his BS, MS, and PhD degrees in Computer Engineering in 1995, 1997, and 2003, respectively, from Pusan National University, Pusan, Republic of Korea. From 1997 to 1998, he worked as a research engineer at the LG Multimedia Research Laboratory. Since 2005, he has been working as a principal researcher at the ETRI, Daejeon, Republic of Korea. His current research interests are in quantum computing and cloud systems.

Sangmin Lee received her BS degree in Computer Engineering at the Inha University, Incheon, Republic of Korea in 1991. She has been with the ETRI, Daejeon, Republic of Korea since 1991, where she has worked on developing the SCSI and FC RAID system, distributed parallel file system, dual-mode big data platform, and simulation technology for the digital twin. Currently, she is working as a principal researcher. Her current research interests include distributed systems, extreme storage systems, and quantum operating systems.

Ki-Sung Jin received his BS and MS degrees in Computer Engineering from Jeonbuk National University, Jeonju, Republic of Korea, in 1999 and 2001, respectively. Since 2001, he has been with the ETRI, Daejeon, Republic of Korea, where he has worked on developing the cluster database, distributed parallel filesystem, dual-mode big data platform and simulation technology for the digital twin. He is currently a principal researcher. His current research interests include distributed systems, extreme storage systems, and quantum operating systems.

Gyuil Cha received his BS and MS degrees in Computer Science from Korea University, Seoul, Republic of Korea, in 1998 and 2000, respectively. Since 2000, he has been with the ETRI, Daejeon, Republic of Korea, and is currently a principal researcher. His research interest is a quantum operating system for fault-tolerant quantum computing. He has been involved in the technology development of operating systems, memory virtualization, supercomputing, microservice architectures, and extreme storage systems.

DB Workload Management through Characterization and Idleness Detection

Abdul Mateen
Department of Computer Science,
Federal Urdu University of Arts,
Science & Technology, Islamabad
45570, Pakistan;
abdulmateen@fuuastisb.edu.pk

Khawaja Tahir Mahmood
Department of Computer Science,
Federal Urdu University of Arts,
Science & Technology, Islamabad
45570, Pakistan;
khawajatahirmahmood@gmail.com

Seung Yeob Nam*
Department of Information and
Communication Engineering
Yeungnam University
Gyeongsan, South Korea;
synam@ynu.ac.kr

***Abstract*—It is difficult to handle the database (DB) workload due to the huge increase in data, the functionality demand from the user, and the rapid changes in data. It is not easy to manage the DB workload, which therefore leads to malnourishment. To get efficient results, there must be complete knowledge about the type and changes in workload. The versatility and complexity of DBMSs led the DB researchers towards new philosophy and thoughts. A novel approach is introduced for DB workload management through characterization, scheduling, and database idleness detection. In workload characterization, workload is observed, and effective workload characterization parameters are selected. After that, scheduling is performed in order to arrange the DB workload to reduce the waiting time for each workload. Lastly, database idleness is identified at run-time and exploited for system as well as user-initiated workloads to improve efficiency. The proposed approach for workload management is validated through experiments using benchmark workloads.**

Keywords—Workload, Autonomic, Characterization, Idleness Detection, Scheduling

I. Introduction and Background

The traditional DBMS are managing the data in from decays, and due to this reason, it is one of the core components of any company or business. The workload in DBMS has been grown due to large amount of data that cannot be handled by the human. The worth of DBMS was more than 15 billion dollars in 2005, with 10% annual growth that is also one of the major causes to increase of ownership [1]. Many techniques and models are presented in order to manage the workload efficiently. The main challenge in DBMS is to manage the workload efficiently which is complex, tricky, and comprised of complex queries. These workloads use various resources and process huge data. Usually, important workloads require an immediate response as compared to routine workloads. On the other hand, DBMS also have to manage the system-oriented tasks that run regularly (e.g., index, statistics management etc.).

The proposed research is being conducted to manage the complex and tricky workloads efficiently. The number of tasks assigned to a worker in a given amount of time is referred to as their workload. In computer science, it is defined as the processing time required to complete a task, which may include some programming [2]. The amount of work that a database management system (DBMS) produces or can produce in a given time interval is referred to as the database workload. A more concrete definition of the DB workload could be the set of structured query language statements, which could be OLTP or DSS, sessions, commands, and other regular operations. The OLTP workload is composed of tinny transactions that may be insert, delete, and so on, on the other hand the DSS workload requires long time for completion, many resources and involves much computations. Prior to the 1980s, capacity planning was introduced with the goal of cost sharing and as a method of forecasting future organizational demand. It is used to forecast the organization's future needs. Capacity refers to the maximum amount of work that an organization can complete in a given period of time. Workload management capacity planning always ensures that functions meet their production targets while staying within budget. In terms of workload change, capacity planning's primary concern is with resources, their management, configuration, and impact. Characterization and performance are used to plan capacity [2-3]. Later, the capacity planning research was converted to Resource-Oriented Workload, which maintained the service level concept by allocating resources to applications while maximizing resource utilization.

The goal of a performance-oriented task is to maximize resource efficiency while minimizing laborious effort. It executes the adaptation by configuring and optimizing without any human input [4], and is sometimes referred to as autonomic workload management. Finally, the evolution of the workload has moved from capacity planning to resource orientation to performance orientation. Technology-enabled cost sharing and Service Level Objectives (SLOs) are the cornerstones of the evolution of workload. The SLO, which details the agreement's objective, is the most important part of a service level agreement (SLA) between a service provider and a client. By comparing the proposed solution to these well-known methods, the effectiveness of the solution is evaluated. This approach is used for two reasons: first, none of the DBMS providers (Oracle, SQL Server, DB2) disclose information about internal workings of their products, and second, it is challenging to change the internals of the DBMS because the source code is not readily available. This approach was also used by earlier researchers, such as Mumtaz et al [17], to demonstrate the efficacy of their suggested strategy. Another well-known and effective method for comparison is time sharing, in which DB workloads are run for specified intervals of time and may be completed, or they may be briefly halted and resumed when it is their turn. This article is organized as following: Section II discusses the related work. Section III describes the steps to characterize, schedule and identification of idleness detection in the DB workload. In section IV, experimental results are illustrated, and finally section V concludes the article with future directions.

II. State of the Art

Many researchers are working to increase the self-management, dependability, and efficiency of DBMSs. Here, the literature review is divided into two areas, namely, the sections on characterization and scheduling strategies. A number of these are covered in the section that follows and are pertinent to the solution for workload characterization and scheduling in DBMSs and DWs.

Two utility functions (dynamic resource allocation and query scheduling) are introduced [13] to optimize the autonomic systems according to the specific goals. It is discussed how these two utility functions can incorporated in DBMSs. These functions convert the BI importance policies into low-level policies through which performance objectives of the organization are achieved. It also advice better technique is for some workload executing on the DBMS under particular circumstance. DBMS congestion indicators were identified and discussed by Zhang et. al. that consists of DB monitor metrics. The system should have the ability to detect the congestion and then work according to the situation. The approach is then validated by performing various experiments and these indicators can vary according to the congestion. However, the research did not identify or predict system performance problems in detail. Moreover, the experiment is performed only taking the specific scenario. A model is proposed [5] to identify and characterize the workload; and effective workload parameters are identified. [6, 7] proposed the classifier that works on the basis of workload characteristics and manages the workload characteristics through decision tree induction. However, there is no discussion about the distinct workload parameters which are most affective to distinguish the workload in the proposed classification. Research in [8] examine the SQL statements and views with respect to their composition; and classifies the data. The efficiency is increased by finding out the criteria for DB workload. However, there is no discussion about the generation of instructions for the physical design automatically. An analysis of characterization technique for Business Intelligence (BI) workload is discussed through clustering technique [9] where the Singular Value Decomposition (SVD) and SemiDiscrete Decomposition (SDD) clustering algorithms are applied. The proposed approach uses resource related parameters that include number of joins, CPU usage and Input/ Output rate. For each workload class data is validated and resource demand is identified and finally a report is generated that reveals a priority wise list of hardware configuration. The characterization of workload is done by performing five steps which are components identification, selection of characterizing parameters, input data normalization, workload grouping into classes and finally identification of each class. The proposed characterization technique is performed on TPC-H benchmark like data but with some assumption based parameter. The technique is limited as for characterization it only considers one parameter i.e. user resource demand. A technique is proposed [10] where the State Transition graph was introduced for characterizing the workload. It shows the same pattern for those customers who are doing similar type of activities. After finding this information, clustering is used to characterize the workload. However, the proposed technique had some problems such as maximum session drops when there large number of customers and no mechanism is describe to extract these session drops. An approach [11] is introduced that is used to observe the DB workload by using the n-Gram model. It consists of an API that determines workload shifts with two major steps; first observing the workload and comparing with the workload model and configuration analysis is done with the shift. Experiments are performed on the DVD shop dataset where different workload shift scenarios are taken. The DB workload in E-commerce workload is characterized by using the black box attributes [12] where parameters such as disk usage, response time and CPU time are used. It is observed through experiments that dynamic cache has the ability to reduce response time. It was also highlighted in the research that efficiency of the Hybrid cache is better than Table and Query cache. The experiments are performed on the TPC-W benchmark to validate the black box approach. A framework [13] for autonomic workload management is proposed which monitor and control the flow of the DB workloads dynamically without any human involvement. The framework mainly comprises of multiple workload management techniques and performance monitor functions. Whenever the performance degrades, the proposed approach find out that issue and handle it intelligently by the help of indicators. The experiments and results are verified by using a prototype implemented in DB2. One issue in this research is that the workload is assumed to be constant which is heterogeneous in real cases. The other problem with the research is that prototype is built instead of complete system. A framework [14] is proposed to handle workload by stopping and restarting queries. When a query I resumed it always restart from the stopped position. The proposed framework has ability to manage single Query Execution Plan (QEP) and fails to restart query when records are updated. A query suspension and resumption [15] technique is introduced where asynchronous checkpoints are used. In case of query suspension all the resources are released while in case of resumption the required resources are resumed. The technique is also implemented in a tool named as PREDATOR and proved that this tool has better results than others. It allows to suspend the query as compared to previous techniques where switching is performed between individual operators. However the proposed technique does not re-optimize the given query and memory wastage is also less. A technique to manage the OLTP workload is discussed [16] where MPL is set by using the parameters (CPU and disk) and used feedback controller. [17] discussed a framework Query Shuffler (Qshuffler) where DB workload is managed by finding out dependency of query. This proposed research works without knowing database and causes of query dependency. It is implemented in DB2 and experiments are performed over the TPC-H benchmark workload. As per author’s claim, it is four times better than the FIFO scheduling. As larger jobs have to wait for long time therefore average execution time is larger. The Mixed Workload Scheduler (MWS) is introduced [18] where a non-preemptive scheduling was used in order to manage the BI workload. It takes assumption-based parameters and considers queries at the same level. Moreover, the work does not consider the interaction between various queries and have no ability to set the MPL dynamically. There are three parts of the manager, first is Admission control that make batches of queries according to the available memory. In this research, priority is given to that query which requires maximum memory. The underload and overload problem is managed by using the Priority Gradient Multiprogramming (PGM). The manager cannot manage the interactive and ad-hoc queries. Moreover, smaller queries remain suspended until the completion of long queries. The Query Progress Indicator (PI) for single query with a graphical user interface [19] is introduced to represent the status of query and provides information about the remaining execution time. It is limited to single query and does not consider the query dependency. A Multi-query PI [20] displays the time of running queries with the consideration of query dependency over others. The technique is the first attempt for a multi-query PI. It has also the ability to predict future queries with an efficient workload management. The proposed PI is implemented in POSTGRE SQL and various experiments are performed to show its effectiveness. A Query Patroller (QP) for IBM DB2 is discussed in [4] which manage the database requests from the users as per availability of resources. It divides the workload into classes on the basis of their cost and Multi-Programming Level (MPL) threshold. Usually, DBA assign privileges to different resources in QP that analyze and schedule the workload after observation. Saturation is avoided by limiting the workload with maximum resource usage. The proposed advisor for tuning and optimization of query [21] is used to manage the workload in Oracle. The SQL Tuning Advisor is responsible to generate the recommendations for the given workload that is further used to generate best Query Execution Plan (QEP). It assists the query optimizer when fails due to heavy workload. The proposed manager for handling resources in Oracle [22] has the ability to allocate resources to given workload. OLTP workload execute before the DSS workload when there is an enormous number of requests from the users. ODRM also facilitate administrator by defining and scheduling policies for different requests. There are three main components Resource of the ORDM (Consumer Group, Resource Plan and Resource Plan Directive). Economic model is introduced by [23] that is used to manage the workload proactively. The workload is divided into various classes on the basis of cost and time limit. Experiments are performed with a claim that the proposed model is scalable with the workload size. A Priority Adaptation Query Resource Scheduler [24] manages the multiclass queries in an efficient way. The MPL is set according to the given workload by using the miss ratio projection and resource utilization heuristics. The proposed research has the ability to handle static workload; however, fails when there is rapid change in queries. A framework for the adaptation of workload [2] is introduced to manage the workload by using four modules and the workload is predicted by the Kalman Filter. The experiments of the proposed research are performed on the stable workload and can handle only the linear workload.

There is no identification of effective workload parameters in previous research. Some of these used the values in various calculations which are either estimated or static. Due to this reason, there results are inaccurate. Moreover, none of the above strategy or tool finds out the proportion of the OLTP and DSS by using the Fuzzy Logic. In this situation, a novel workload management approach is required where the database requests must be handled in an efficient way.

III. DB Workload Management

The proposed research work consists of three major modules that include workload characterization, scheduling and idleness detection. These modules are discussed as:

A. Characterization of DB Workload

Main components of the DB workload management are workload, resources and objectives which work in a loop to facilitate users. Workload has evolved from capacity planning to resource sharing and finally performance oriented [2]. Steps of the proposed workload management approach are shown in Fig 1. A Case Base Reasoning (CBR) algorithm is adopted that uses the parameters to find out the workload type. It is explained in our previous research [24].

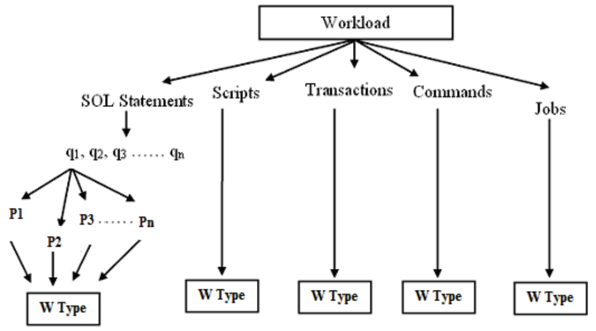


Fig. 1. Identification of DB Workload Type [24]

The suggested database workload characterization technique is verified and shown to be effective in classifying the workload type. For our experiments, we have employed the TPC-W benchmark workloads (ordering profile as OLTP and browsing profiles as DSS workloads). The values of the selected MySql status variables are retrieved using the display status statement of MySql and recorded in the database to help identify the type of workload. The workload is built using threads created by the client computers, and while it is running, values for the status variable are logged every 5 seconds. According to the workload characterization for the browsing workload, DSS accounts for 89.3 percent of the burden and OLTP for 11.7 percent. When it comes to ordering workload, it indicates that 8.5% of the workload is DSS and the rest is OLTP. The labor associated with the shopping-related activities is then carried out. Following its execution and analysis, it was discovered that the shopping profile has a workload split of 77% DSS and 23% OLTP. Due to the definition of the benchmark and our awareness of particular workload attributes, the outcomes of these trials are in line with what we had anticipated. To obtain accurate findings, the same tests are also run repeatedly.

B. Fuzzy Based (FB) Scheduler

The process of scheduling involves arranging the workloads produced by users and the system in a particular way for quicker execution. It is crucial to address all underlying problems and to set up the workload in the database so that each request is handled with the greatest possible CPU efficiency. Both internal and external scheduling can be done in DBMSs [16]. It is not required for DBMSs to have solely OLTP or DSS workload, i.e. workload that is 100% OLTP and 0% DSS or 100% DSS and 100% OLTP. If the workload is thought of as crisp, higher performance from the DBMS cannot be attained due to the mix of workload. The workload behavior in DBMS is typically non-deterministic; for instance, w1 and w2 are composed of 60% OLTP and 40% DSS transactions, respectively, and 20% OLTP and 80% DSS transactions. There is a need of a scheduler that can differentiate the OLTPness and DSSness of the incoming workload. The Fuzzy logic is adopted here to find out the workload type proportion from the database requests. Fuzzy logic is based on the Fuzzy Set Theory [25] where the value is always from 0 to 1. The crisp logic has only the Boolean value however in case of Fuzzy logic degree of the truthiness or falseness is identified. Following section discusses the architecture of the proposed scheduler for DBMS that uses the Fuzzy logic to calculate the proportion of OLTP and DSS. Finally, Fuzzy rules and membership functions are applied and database requests are arranged accordingly.

Fuzzification, Fuzzification Rules and Defuzzification are the main steps of the FB scheduler. Membership function is used to calculate value in the range of 0 and 1. The Gaussian method is adopted in this research due to its simplicity and easy conversion to the Fuzzy set. User Priority (UP), DSS and OLTP are found by the membership function.

Consider a DB workload 'w' with various database queries is executed over MySQL 5.1. Proportion of both workload types is identified for all database requests. The OLTPness and DSSness of each workload is calculated. The proportion of OLTP workload is 0.473% and DSS workload is 0.527. Subsequently, these proportions are calculated for some specific time. A particular workload that is generated by the users during some specific time t. There were six sub-workloads from different uses that are given to DBMS with different priorities. OLTP and DSS proportion is identified by adopting the same procedure. Finally, these sub-workloads are executed on the basis of these calculations to minimize the waiting time. After the fuzzification, the Fuzzy Rules in the form of "If Condition then Action" are designed and expressed in terms of Fuzzy words. In this scenario, eight (8) priorities (0 to 7) may be assigned to each user. The user with 0 priority has the highest priority and 7 for the lowest priority. There are 8 classes (Class 0 to 7) of the workload for execution on the basis of their size and importance. The execution of workloads starts from class 0 to class 7. It has been observed after experiments that much time is required for the calculation. So, number of classes minimized. The High class consists of the workloads with smallest time and high priority. In defuzzification step, final output is produced by conversion of Fuzzy Set to their corresponding output values. Workload type is find out by considering the above OLTP and DSS proportion.

C. *Idleness Detection*

The idleness detection approach constructs and stores sample data in the CBR as depicted in Table 1. The threads create a variety of workloads, and each thread is in charge of executing workloads that include OLTP, DSS, or a mix of query types. The Workload ID, OLTP Transactions, DSS Transactions, and DBMS Idleness are among the parameters saved in the CBR. The column Workload ID represents the automatically generated unique number, OLTP Transaction column shows the number of OLTP queries in the workload, DSS Transaction column shows the number of DSS queries in the workload, and the DBMS Idleness column represents the amount of time the given database workload was idle.

TABLE 1. CBR DATA FOR DBMS IDLENESS DETECTION

Workload ID	OLTP Transaction	DSS Transaction	DBMS Idleness (%)
1	12	5	28
2	5	3	57
3	7	11	19
4	15	5	33
5	31	3	26
6	2	5	19
.	.	.	.
.	.	.	.

In order to find the DBMS idleness, four steps are adopted that is Retrieve, Reuse, Revise and Retain. The initial step to identifying the DBMS inactivity for the incoming workload is matching the OLTP and DSS Transactions of the currently running workload with cases that have previously been saved in the CBR.

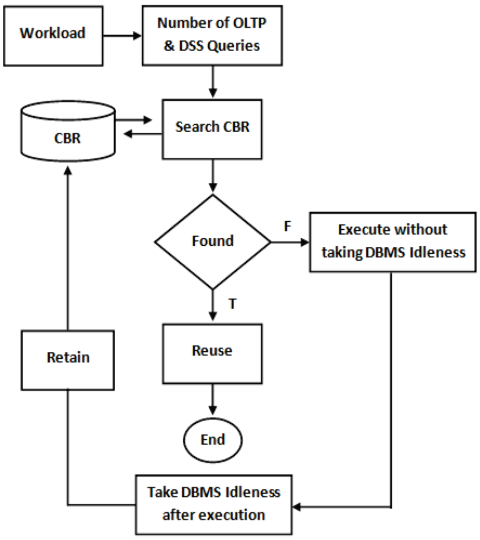

Workload
Number of OLTP & DSS Queries
CBR
Search CBR
Found
F
Execute without taking DBMS Idleness
T
Retain
Reuse
End
Take DBMS Idleness after execution

Fig. 2. Steps to Identify the DBMS Idleness

The retrieval is carried out using the problem description. It is vital to keep in mind that if equivalent instances are not found, cases may be incorrectly matched, leading to wrong results and judgments. This is essential since accuracy depends on how many cases are saved and how many matched cases can be retrieved. It is possible to gauge how similar two cases are by calculating the distance between them using similarity matrices that can be offered by experience or experts, for instance. Four further sub-steps are included in the retrieval stage: finding related features, making an initial match, searching, and selecting the matched instances from the CBR. The values of the selected features are then extracted from the incoming database workload. The workload in this scenario is used to determine the number of OLTP and DSS transactions or requests. The gap between the CBR recordings and the incoming workload's characteristics is calculated. Here, the Euclidean Distance technique is used to redetermine the matched or nearest cases [26, 27]. It is one of the easiest and quickest methods to locate the matched scenarios. The distance between the characteristics of the incoming workload and the stored examples is determined using the Euclidean Distance equation. After determining the distance between each CBR record and the matching record, the related DBMS idleness is used to determine the number of free cycles. However, if there isn't a match, the workload runs regardless of the DBMS's idleness, and thereafter, the DBMS's idleness throughout execution is recorded in the CBR along with the quantity of OLTP and DSS requests as a new case for later usage as shown in Fig 2.

Let's assume the DB workload consists of nine OLTP and thirteen DSS transactions. To find the DBMS idleness against this workload, we must now employ the match cases from the CBR. As indicated in Table 2, the values of the CBR and new workload for the workload parameter in this instance are not translated into the range [0, 1]. The gap between the new workload and every record stored in the CBR is calculated using the aforementioned Euclidean Distance calculation. The DBMS inactivity for the new job originates from the row/record with the lowest distance value. In this example, the minimum Euclidean distance is 2.82842712, and the corresponding DBMS idleness is 19 (highlighted row 3). Consequently, the DBMS Idleness of the incoming workload will also be 19. The stated DBMS idleness is used by system processes. However, the workload continues to run even while the DBMS is not idle if neither an exact nor a close match can be obtained. After the task is over, it is maintained with the actual DBMS inactivity.

TABLE 2. EUCLIDEAN DIFFERENCE CALCULATION

CBR Data (in the range [0, 1])			New Workload (in the range [0, 1])		Euclidean Difference
OLTP	*DSS*	*CPU Idleness (%)*	*OLTP*	*DSS*	
12	5	28	9	13.0000	8.54400375
5	3	57	9	13.0000	10.77032961
7	11	19	9	13.0000	2.82842712
15	5	33	9	13.0000	10.00000000
31	3	26	9	13.0000	24.16609195
2	5	19	9	13.0000	10.63014581
.	.	.	.	.	.
.	.	.	.	.	.

After the Retrieve step is finished, the Reuse step begins, and it consists of two smaller steps. the distinctions between the calculations for the present and retrieved instances, followed by the identification of the cases that need to be added as new cases to the CBR. The revise step is required when some inaccuracy is found where domain knowledge is used to correct the results. Here, the Substitution adaptation method is adopted, where the transformation adaptation alters the structure of the solution. Finally in retain phase, only valuable information of the given case is inserted into the CBR for future use.

IV. RESULTS AND DISCUSSION

The results of this work are compared with FIFO, and PB techniques where average waiting time (AWT) is calculated for each by executing the DB workloads (w1, w2, w3, w4, w5, and w6). FIFO technique works on first in first out basis, SJF scheduling technique executes from smaller to larger request while PB executes on the basis of user priority. The results achieved from the proposed workload management are analyzed with other techniques. The wait time of proposed and others approaches is calculated as shown in following Table 3 and 4 (execution time of the selected workloads is 0 to 100 minutes).

TABLE 3. RESULT OF THE FIFO AND PROPOSED APPROACH (PA)

AO	Priority	OLT P %	DSS %	OLTP ET	DSS ET	Total Time	FIFO WT	PA WT
2	2	0.27	0.73	0.9	11.3	12.2	7.2	45.4
4	6	0.56	0.44	3.2	5.4	8.6	25	36.8
1	2	0.34	0.66	1.9	5.3	7.2	0	5.6
3	0	1	0	5.6	0	5.6	19.4	0
6	1	0.23	0.77	0.2	6.9	7.1	40.4	12.8
5	3	0.3	0.7	0.6	6.2	6.8	33.6	18.4
						Average Wait Time	**22.1**	**19.83**

where AO are representing arrival order of the workloads.

TABLE 4. RESULTS OF THE PS AND PA

AO	Priority	OLTP %	DSS %	OLT P ET	DSS ET	Total Time	PS WT	PA WT
2	2	0.27	0.73	0.9	11.3	12.2	14.3	45.4
4	6	0.56	0.44	3.2	5.4	8.6	40.8	36.8
1	2	0.34	0.66	1.9	5.3	7.2	23.7	5.6
3	0	1	0	5.6	0	5.6	0	0
6	1	0.23	0.77	0.2	6.9	7.1	7.2	12.8
5	3	0.3	0.7	0.6	6.2	6.8	34.6	18.4
						Average Wait Time	**20.11**	**19.83**

TABLE 5. AWT OF FIFO, PB AND PROPOSED APPROACH (PA)

Scenario	FIFO AWT	PB AWT	PA AWT
1	20.93	20.11	19.83
2	16.5	19.8	13.8
3	21.8	17.6	14.9

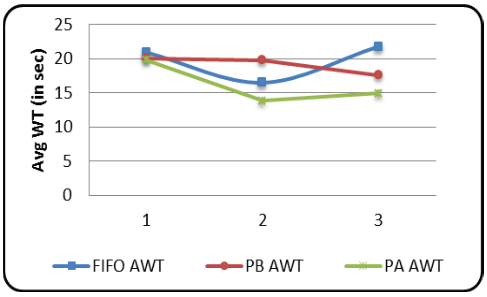


Fig. 3. AWT of FIFO, PB and Proposed Approach

After these experiments, two main problems are identified as shown in Table 5 and Fig. 3; first, smallest workloads have to wait for a long time in FIFO as well as PB; and second, both does not consider the importance i.e., whether the workload is generated from the higher management or not. In FIFO and SJF all the requests are considered at same level and executes from smallest to longest request without considering its importance for higher management. On the other hand, PB executes the workloads by only considering its priority. The proposed research work in all scenarios is better than other existing and well-known workload management techniques.

V. CONCLUSION AND FUTURE WORK

The study recommended a strategy for managing DB workloads that estimates each job's percentage of OLTP and DSS to determine how it is carried out. The workload parameters that are more helpful for classifying and characterizing the task are identified. The workload type is determined using the CBR approach, and the percentage of OLTP and DSS is determined using fuzzy logic. Once the OLTP and DSS percentages are established, the complete workload is arranged for better execution. Results from experiments with workloads that are similar to those for OLTP and DSS show that the suggested characterization and scheduling procedures are effective and take less time, taking workload size and importance into account. The assessments also provide a comparison with other well-known workload management programs. The suggested strategy's results are compared to First in First Out, Priority

Based, and Smallest Job First strategies. The main objective of the analysis and comparison with other commonly used techniques is to show the merits of the proposed study. The logical makeup of the workload, a clustering strategy to combine query templates, and any learning method that may forecast the rate at which the DB workload will arrive will all be considered in future workload.

ACKNOWLEDGMENT

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korean Government [Ministry of Science and ICT (MSIT)] under Grant 2020R1A2C1010366, and in part by the Basic Science Research Program through the NRF funded by the Ministry of Education under Grant 2021R1A6A1A03039493.

REFERENCES

1. M. C. Huebscher, and J. A. McCann (2008). A survey of autonomic computing—degrees, models, and applications. ACM Computing Surveys (CSUR), vol. 40, no. 3 , 1-28.
2. B. Niu, P. Martin, W. Powley, R. Horman, and P{. Bird. (2006, October). Workload adaptation in autonomic DBMSs. In Proceedings of the 2006 conference of the Center for Advanced Studies on Collaborative research (pp. 13-es).
3. S. S. Lightstone, G. Lohman, and D. Zilio. (2002). Toward autonomic computing with DB2 universal database. ACM Sigmod Record, vol. 31, no. 3, 55-61.
4. Z. Zewdu, M. K. Denko, and M. Libsie. (2009, May). Workload characterization of autonomic DBMSs using statistical and data mining techniques. In 2009 International Conference on Advanced Information Networking and Applications Workshops (pp. 244-249). IEEE.
5. S. Elnaffar, P. Martin, and R. Horman. (2002, November). Automatically classifying database workloads. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 622-624).
6. S. S Elnaffar. and P. Martin. (2004, November). An intelligent framework for predicting shifts in the workloads of autonomic database management systems. In Proc of 2004 IEEE International Conference on Advances in Intelligent Systems–Theory and Applications (No. 15-18, pp. 1-8).
7. S. Y. Philip, H. U. Heiss, S. Lee, and M. S. Chen. (1992). On Workload Characterization of Relational Database Environments. IEEE Trans. Software Eng., vol. 18, no. 4, 347-355.
8. T. J. Wasserman, P. Martin, D. B. Skillicorn, and H. Rizvi. (2004, November). Developing a characterization of business intelligence workloads for sizing new database systems. In Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP (pp. 7-13).
9. D. A. Menascé, V. A. Almeida, R. Fonseca, and M. A. Mendes. (1999, November). A methodology for workload characterization of e-commerce sites. In Proceedings of the 1st ACM conference on Electronic commerce (pp. 119-128).
10. M. Holze, and N. Ritter. (2008). Autonomic databases: Detection of workload shifts with n-gram-models. In Advances in Databases and Information Systems: 12th East European Conference, ADBIS 2008, Pori, Finland, September 5-9, 2008. Proceedings 12 (pp. 127-142). Springer Berlin Heidelberg.
11. F. Liu, Y. Zhao, W. Wang, and D. Makaroff. (2004, October). Database server workload characterization in an e-commerce environment. In The IEEE Computer Society's 12th Annual International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems, 2004.(MASCOTS 2004). Proceedings. (pp. 475-483). IEEE.
12. M. Zhang. (2014). Autonomic workload management for database management systems (Doctoral dissertation, Queen's University), Kingston, Ontario, Canada.
13. S. Chaudhuri, R. Kaushik, and R. Ramamurthy. (2005, June). When can we trust progress estimators for SQL queries?. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data (pp. 575-586).
14. B. Chandramouli, C. N. Bond, S. Babu, and J. Yang. (2006, April). On suspending and resuming dataflows. In 2007 IEEE 23rd International Conference on Data Engineering (pp. 1289-1291). IEEE.
15. B. Schroeder, M. Harchol-Balter, A. Iyengar, E. Nahum, and A. Wierman. (2006, April). How to determine a good multi-programming level for external scheduling. In 22nd International Conference on Data Engineering (ICDE'06) (pp. 60-60). IEEE.
16. M. Ahmad, A. Aboulnaga, S. Babu, and K. Munagala.(2008, April). Qshuffler: Getting the query mix right. In 2008 IEEE 24th International Conference on Data Engineering (pp. 1415-1417). IEEE.
17. A. Mehta, C. Gupta, S. Wang, and U. Dayal. (2009, March). rFEED: a mixed workload scheduler for enterprise data warehouses. In 2009 IEEE 25th International Conference on Data Engineering (pp. 1455-1458). IEEE.
18. A. Mehta, C. Gupta, and U. Dayal. (2008, March). BI batch manager: a system for managing batch workloads on enterprise data-warehouses. In Proceedings of the 11th international conference on Extending database technology: Advances in database technology (pp. 640-651).
19. G. Luo, J. F. Naughton, and P. S. Yu. (2006). Multi-query SQL progress indicators. In Advances in Database Technology-EDBT 2006: 10th International Conference on Extending Database Technology, Munich, Germany, March 26-31, 2006 10 (pp. 921-941). Springer Berlin Heidelberg.
20. B. Dageville, D. Das, K. Dias, K. Yagoub, M. Zait, and M. Ziauddin. (2004, August). Automatic sql tuning in oracle 10g. In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30 (pp. 1098-1109).
21. A. Rhee, S. Chatterjee, and T. Lahiri. (2001). The Oracle database resource manager: Scheduling CPU resources at the application level. HPTS, vol. 1, no. 2, 2-4.
22. S. Krompass, H. Kuno, J. L. Wiener, K. Wilkinson, U. Dayal, and A. Kemper. (2009, March). Managing long-running queries. In Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (pp. 132-143).
23. H. Pang, M. J., Carey, and M. Livny. (1995). Multiclass query scheduling in real-time database systems. IEEE Transactions on knowledge and data engineering, 7(4), 533-551.
24. M. Abdul, A. M. Muhammad, N. Mustapha, S. Muhammad, and N. Ahmad, (2014). Database workload management through CBR and fuzzy based characterization. Applied Soft Computing, vol. 22, 605-621.
25. L. Zadeh, (1965). Fuzzy Sets and Systems, Fox Journal, editor. System Theory, Brooklyn, NY: Polytechnic Press, pp. 29–39.
26. E. Deza, M. M. Deza, M. M. Deza, and E. Deza. (2009). Encyclopedia of distances (pp. 1-583). Springer Berlin Heidelberg, ISBN 978-3-642-00234-2/ebook, Springer.
27. P. E. Danielsson. (1980). Euclidean distance mapping. Computer Graphics and image processing, vol. 14, no. 3, 227-248.

Micro-services internal load balancing for Ultra Reliable Low Latency 5G Online charging system

Ngoc Tien Nguyen*, Thanh Son Pham*, Van Duong Nguyen*, Cong Dan Pham*, Duc Hai Nguyen*
*OCS Research Center, Viettel High Technology, Viettel Group, Hanoi, Vietnam
(tiennn18, sonpt26, duongnv21, danpc, haind13)@viettel.com.vn

Abstract—**Connection mesh plays an important role in micro-service architecture and greatly influences system performance. 3GPP proposed a 1ms requirement for 1-way transmit activities on the 5G URLLC data plane. As a 5G component, our Online Charging System needs to manage sub-milliseconds operations. In this paper, we model the request flows in the mesh as a max flow problem with back-pressure. Inspired by the connection lib approach, we propose our connection library Microchassis. We select basic routing (front-pressure) methods including round-robin (RR), least concurrent (LCC), and proposed two more routing methods using *back-pressure flow control policy based on concurrent status* (BCC) and *back-pressure flow control policy based on throughput combined with concurrent status* (BTC) for the library. Then we evaluate the performance and latency among four methods of Microchassis and validate our modeled formulas. Finally, we simulate the OCS architecture of the main charging flow and qualify it against the URLLC latency standard. The result shows that back-pressure strategies can maintain great QoS under extreme conditions, superior to the front-pressure algorithm family. BTC performed well in all cases with the highest throughput in the incident. Besides, BCC only shows its advantages on high throughput systems. Front-pressure strategies are better choices for internal well-managed service layers, and LCC is outperformed on RR.**

Keywords—**Service mesh, Connection library, Internal load balancing, Benchmark, 5G, URLLC, Online-charging system**

I. Introduction

Connection lib shows its important role in a software system when millions of customers from billions of devices demand high-quality network services. Especially in the telecom sector, the explosion of new services and end-user devices in the 5G era is required as eMBB, mMTC, and URLLC standards [1], widely known in Telecom, Medical, Industry, Banking, and Automotive industries [2]. Our OCS decides the availability of service to end users. This calls for the need for robust message forwarding in micro-service networks. Internal client routing algorithms need to be fast and efficient, so this paper evaluates different strategies.

In particular, we discuss the connection lib approach in micro-service mesh architecture with four routing Internal Load-balancing methods relevant to OCS, two popular and our two proposed: simple Round robin (RR), metric-based Least concurrent (LCC), and Back-pressure flow control methods, based on concurrent state (BCC), and based on flow speed and concurrent state (BTC). First, we introduce the mesh problem in section I and the connection lib approach in section II. Then we model the max flow and back-pressure problem in section III. In section IV, we demonstrate our adoption of the Microchassis in OCS. In the same section, we propose our back-pressure flow control policies BCC and BTC. Then we analyse the performance of each method in section V with further discussion in section VI.

The main focus of the paper is the performance and latency of the four methods in various scenarios on the same input. Our three contributions are: the request flow as a max flow problem with back-pressure; a guide on implementing back-pressure strategies BCC and BTC, as well as benchmark results to compare strategies.

A. Background

1) Overhead in service mesh: Service mesh with an independent sidecar is favored for its flexible capability and cost-effectiveness in a variety of purposes with the ability to be transparent under network layers. Sidecar greatly influences the response time, hence the overall system's performance which was studied in previous research. Istio in their guide [3] stated factors of latency in the data plane and showed a positive estimate. Istio measured each network I/O event gain 1ms more. Envoy [4] is the most popular proxy sidecar, especially in Istio. While [5] assesses the overhead impact of using Envoy, Zhu et al. considered the mesh overhead as an inevitable aspect and predicted it by message size, network speed, business process, and zero-copy factor. Zero-copy shows a minor impact on latency but a big trade-off for the effort of managing the buffer programmatically.

2) Routing strategy: A survey [6] mentions some well-known LB methods in switching. Envoy [4], F5 load balancer [7], IPVS [8] and Nginx all support popular methods like Round-robin, Least connection, Destination or Source hashing, Shortest delay, and Maglev. NGINX uses P2R combined with the least connection or least time strategy [9] to reduce the complexity and probability of errors. However, HAProxy refuted the idea that P2R is universally efficient, contrary to theoretical reasoning [10]. This reinforced our motivation to compare complex back-pressure algorithms against simple algorithms such as least-X, and RR.

3) Back-pressure: Originally a concept in dynamic fluid theory, back-pressure is used in computing to describe the resistance or force opposing the flow of data through services. It exists in any resource utilization process. The need to handle back-pressure in the max flow problem was mentioned in [11]. Instead of optimizing for QoS by each specific root cause, the back-pressure approach senses the high pressure and then adjusts the flow based on output metrics. The back-pressure approach deals with the case where the input rate exceeds the system's threshold - detailed below as the saturation point. This approach aims to maintain the highest throughput of the best quality. Theoretically, back-pressure algorithms rely on at least one of the following strategies:

Buffering: handles temporary spikes and flattens the distribution of incoming requests, but is ineffectual when the incoming requests consistently overwhelm the system. Little's Law dictates that the required queue length would reach infinity.

Drop: dropping is the last resort in any routing strategy. Dropping needs to respect the sequence of message events, specified by the business flow. As proved in III.5.3, overflowing messages must be dropped. Discarding messages and responding promptly is conducive to better responsiveness perceived by the end users, compared to letting messages wait in queues.

Control speed: Flow control mitigates request waves by cutting off requests to attain a sustainable flow. The producer does this based on the consumer's status, which can be implemented in two ways.

1) Push: the producer listens to the connection throughput or the consumer's capacity information to decide. The push model is used in RxJS, RxJava, RSocket, etc
2) Pull: consumer signals that it can handle more requests. Pull model is used in Apache Flink v.15+ cho TCP-based và Credit-based back-pressure [12].

Circuit breaker (CB) is considered semi-auto flow control with the ability to terminate the flow and then raise the speed slowly, according to an FSM configured by the developer. CB is most suitable for front-pressure routing methods. However, it can perform flow control only after incidents have happened, resulting in a latency graph with cycling peaks that resemble shark teeth, as illustrated in [11].

II. Connection lib approach

Detached sidecar cannot obtain business KPIs to make routing decisions. This lack of capability may prove to be detrimental in a distressed system. Integrated Internal client-side Load-Balancing library (ILB) controls the flows on a per-hop per-flow basis, solves the latency problem by bypassing external proxies and sidecars, maintains quick, continuous, self-managed communications in complex misinformation cases, adapts to the local context without compounding the network delay. The ILB can run independently or interact with the centralized control plane. With the ILB lib, capacity planning and auto-scaling rules are temporary solutions that rely on a rate limiter. Those solutions only reduce the probability of a cascading failure without full-time protection. The priority of a per-flow rate limiter is to maximize the throughput under a threshold. Furthermore, one of the minor problems of ILB is that it worsens the computing overhead for flow control.

Uber Hyperbahn [13], Java Maven library Netflix Ribbon [14], and Meta ServiceRouter Lib (SRLib) [15] also adopt this architecture.

A. Drawbacks of connection lib

The ILB connection lib handles all messages indiscriminately, without paying attention to the message semantics including:

1) Time dimensions: (a) ILB needs to handle any range of messages time (e.g. minutes, days, etc.) (b) Timeout is defined by the business process. Many related research leveraged time dimensions: (1) as metrics of service mesh routing [11], [16], (2) to minimize the latency [17], and to predict the latency [5].
2) Size
3) API Destination or path

In OCS, we usually trade system utilization (around 60%) for consistently low latency in sensitive and high-performance businesses (also shown in [2]). Since the need for an adaptive ILB to serve over 100 million profiles in OCS, we extended the problem of routing and monitoring beyond micro-services, in our connection lib to fulfill the 5G URLLC requirements. The lib routes continuously and flexibly while taking business KPIs into account. We have designed Microchassis - OCS ILB connection lib, implemented in Java, in place of the service mesh sidecar.

III. Modeling

A. Max flow problem modeling

We model the L7 max flow problem, as the following:

Definition III.1 (Service layer). *Let $\mathcal{L}_i$ denote a service layer ith of a q service layers system $\mathcal{S}_{\gamma,\delta,q}$ or a subsystem since services pth to qth denoted by $\mathcal{S}_{\gamma,\delta,p,q}$, describes γ incoming gate and δ outgoing gate. Each layer $\mathcal{L}_i$ at time t contain n_i homologous instances ζ_i*

$$|\mathcal{L}_i(t)| = \left|\left\{\zeta_i^k \middle| k \in [1, n]\right\}\right| = n_i \tag{1}$$

Iff tail layer $\mathcal{L}_j$ has its unneglectable pressure responses, then consider it as mid of the response flow (2). Else it should be considered as the tail of the flow (3) and separated into two, request (3a) and response (3b) flow.

$$\mathcal{L}_0 \overset{\xrightarrow{X_1}}{\xleftarrow[Y_{2j-1}]{Y_0}} \quad \mathcal{L}_1 \overset{\xrightarrow{X_2}}{\xleftarrow[Y_{2j-2}]{Y_1}} \quad ... \overset{\xrightarrow{X_{j-1}}}{\xleftarrow[Y_{j+1}]{Y_{j-2}}} \quad \mathcal{L}_{j-1} \overset{\xrightarrow{X_j}}{\xleftarrow[Y_j]{Y_{j-1}}} \quad \mathcal{L}_j \tag{2}$$

$$\begin{cases} \mathcal{L}_0 \xrightarrow{X_1} \quad \mathcal{L}_1 \xrightarrow{X_2} \quad ... \xrightarrow{X_{j-1}} \quad \mathcal{L}_{j-1} \xrightarrow{X_j} \quad \mathcal{L}_j & \text{(3a)} \\ \mathcal{L}_0 \xleftarrow{Y_{2j-1}} \quad \mathcal{L}_1 \xleftarrow{Y_{2j-2}} \quad ... \xleftarrow{Y_{j+1}} \quad \mathcal{L}_{j-1} \xleftarrow{Y_j} \quad \mathcal{L}_j & \text{(3b)} \end{cases}$$

Definition III.2 (Compound instance)**.** *A service instance is a combination of H handling functions $\mathcal{H}$ share the same set of resources with at least 1 of the following differences in signature:*

- *incoming processor*
- *one of a function of inner instance processing chain*
- *outgoing processor*

, each chain of handling summarized a handle f_h

$$\mathcal{H}_i = \bigcup_{h:h\in[1,H]} f_h, \quad \forall h_1 \neq h_2 \in [1,H] \begin{cases} f_{h_1} \cup f_{h_2} \neq f_{h_1} \\ f_{h_1} \cup f_{h_2} \neq f_{h_2} \end{cases} \tag{4}$$

Definition III.3 (Discrete time)**.** *On segment time $\mathcal{T}_m$, or window T_k is the amount of time which has absolute duration time Δ_{t_k} bounded with time t_k since beacon time t_0. Notice that when mentioning segment time $\mathcal{T}_m$, it is a range where many inclusive requests fit in, and when mentioned window T_k, it illustrates ordered windows occur exactly $k+1$ inclusive requests. Minor time ϵ_{t_k} is the idle duration time between two adjacent requests. By the order of processing messages in the network buffer, messages come in order, there aren't any two requests coming at the same time.*

$$\begin{aligned} &\forall t_k \succ t_{k-1}; \quad t_k \succ t_0 \\ &\qquad \Delta_{t_k} = \Delta(t_0, t_k) = |t_k - t_0| \\ &\qquad \epsilon_{t_k} = \Delta(t_{k-1}, t_k) = |t_k - t_{k-1}| \\ &\qquad t_k = t_0 + \Delta_{t_k} = t_{k-1} + \epsilon_{t_k} \\ &\qquad |T_k| = |\mathcal{T}_k| = \Delta_{t_k}, \quad T_0 = t_0 \textit{ and } |T_0| = 0 \end{aligned} \tag{5}$$

for a window slide from t_a to t_b, both count since t_0, we have

$$\forall t_a \prec t_b, \quad |T_a^b| = t_b - t_a \tag{6}$$

Definition III.4 (Capacity varying)**.** *Representative utilizable $\mathcal{U}$ is used for forwarding request x_{t_0}, consumed demand $d(x_{t_0})$. Then, the utilizable resources amount is formalized as:*

$$\mathcal{U} = \frac{d(X_{\mathcal{T}_m})}{\Delta t_m} = \frac{\sum_{j:j\in[0,m]} d(x_j)}{\Delta t_m} \tag{7}$$

Definition III.5 (Messages)**.** *X_{T_m} denotes set of $m+1$ requests of same type appears within the window T_m bounded inclusive by t_0 and t_m requests. X_{T_0} is not a valid range, although it defines 1 request x_{t_0} but without a time range.*

$$X_{T_m} = \{x_{t_0}, x_{t_1}, ..., x_{t_m}\} \tag{8}$$

for an inclusive sliced window that contains requests in group X, we denote

$$X_{T_a^b} = \{x_{t_a}, x_{t_{a+1}}, ..., x_{t_b}\} \tag{9}$$

Definition III.6 (Throughput state)**.** *Current throughput of layer i contain m_i messages*

$$\omega_{i,\mathcal{T}_m} = \lim_{t\to t_m} |X_i| = \frac{\mathrm{d}m}{\mathrm{d}t} = \frac{m_i}{\Delta t_m} \tag{10}$$

Thoughput over an instance a of layer i is number of message over window time T_m

$$\Theta(\zeta_i^a, X_{T_m}) = \frac{m+1}{\Delta t_m} \tag{11}$$

or segment time $\mathcal{T}_m$

$$\begin{aligned} \Theta(\zeta_i^a, X_{\mathcal{T}_m}) = \Theta(\zeta_i^a, X_{T_a^b}) &= \frac{|X_{\mathcal{T}_m}|}{|T_a^b|} \quad \forall x_j \in X \wedge x_j \in T_a^b \\ &= \frac{|X|}{|T_a^b|} = \frac{b-a+1}{t_b - t_a} \end{aligned} \tag{12}$$

Then throuhgput over layer $\mathcal{L}_i$ is

$$\Theta(\mathcal{L}_i, X_{\mathcal{T}_m}) = \sum_{a=1}^{n_i} \Theta(\zeta_i^a, X_{\mathcal{T}_m}) = \omega_{i,\mathcal{T}_m} \tag{13}$$

And we defined $\hat{\Theta}(X_{T_m})$ as configured throughput representing the route strategy allowed in window T_m. $\hat{\Theta} = 0$ is a disabled connection.

Definition III.7 (Throughput capacity)**.** *For each layer, throughput capacity $\phi_{i,\mathcal{T}_m}$ represents for max number of messages servable in duration $\mathcal{T}_m$.*

$$\tilde{\Theta}(\zeta_i^a, X_{\mathcal{T}_m}) = \max_m \Theta(X_{\mathcal{T}_m}) \tag{14}$$

Throughput over a layer is the sum current throughput over all instances.

$$\tilde{\Theta}(\mathcal{L}_i, X_{\mathcal{T}_m}) = \sum_{a=1}^{n_i} \tilde{\Theta}(\zeta_i^a, X_{\mathcal{T}_m}) = \phi_{i,\mathcal{T}_m} \tag{15}$$

Definition III.8 (Concurrent in process state)**.** *Denote that client ILB ψ_i value represents in-flight messages of $\mathcal{L}_i$ client. In not jamming network and network, time assumed nearly 0, according to definition III.1, then minimum ψ=1 shown underlying server there was a single thread continuous processing requests. Ψ_i is the number of max allow to send, $\psi_i \in [0, \Psi_i]$. τ_i is constant number of threads in thread pool of $\mathcal{L}_i$ server, υ is number of processing messages, and π_i is number of message in server queue. Active threads will handle the message, which means if the number of messages is less than the number of threads, the queue will be empty. Therefore number of message locate at instance k of layer i, $m_{\zeta_i^k}$ is:*

$$m_{\zeta_i^k} = \min(\upsilon_{\zeta_i^k}, \pi_{\zeta_i^k} + \tau_{\zeta_i^k}) \tag{16}$$

$$\psi_i = m_{\zeta_i^k} + \psi_{i+1} \tag{17}$$

Definition III.9 (Saturation point)**.** *The system defines a performance rating ρ of current average flow demand within a segment time $\mathcal{T}_m$ on standard average demand on the ideal system.*

$$\rho = \lim_{\Delta t_m \to \epsilon_1} \frac{\hat{d}(X_{\mathcal{T}_m})}{d(X_{\mathcal{T}_m})} = \frac{|\mathcal{T}_m|}{m+1} \tag{18}$$

If the performance rating drops under a predefined threshold or it makes the flow change the handle function, mark it as saturation point ϱ. Since $\rho < \varrho$, the number of messages over segment $\mathcal{T}_m$ decreased, so did the throughput.

The max flow problem is formalized as single data flow through service mesh, based on two Proposition III.1.1 and III.1.2. This proposition models the problem as a steady flow problem, which means the fluid of data is retained after layers.

Proposition III.1. *Single presentation in time:*
Proposition III.1.1 *Time representation: Any represent of the service layer is based on its path order on the processing flow, instead of its connected graph order. If the request traverse through the instance multiple time into different handle f_h that $f_{h_1} \neq f_{h_2}$, it is considered as two different steps to calculate back-pressure, as illustrated in fig 1.*

$$\forall t_0 \prec t_1 \Leftrightarrow, x_{t_0} \mapsto x_{t_1} \Rightarrow f_{h_1}(x_{t_0}, t_0) \prec f_{h_2}(x_{t_1}, t_1) \quad (19)$$

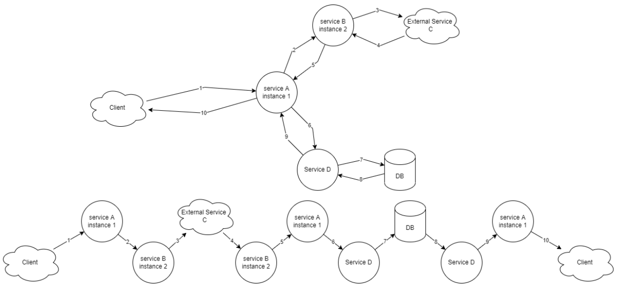


Figure 1. Flow stretching

Proposition III.1.2 *Split flow: (a) Any fan-out stage on the flow at the outbound of $\mathcal{L}_i$ which duplicates an output message is divided into different flows due to Definition III.2. (b) By the nature of sending time, if there were multiple different asynchronous outputs sent to multiple destinations, those are considered as multiple separated flows.*

Corollary III.0.1. *Forward once: Based on proposition III.1.2, there are not multiple outputs from any $\mathcal{L}_i$ at any point of time. Then outbound of the upstream layer is the direct inbound of the downstream layer.*

$$Y_i \equiv X_{i+1} \quad (20)$$

Definition III.10 (Request type)**.** *At any time t that system is not in pressure, set of requests x that go through same $\mathcal{H}_i$ handle, or any two different output formats of same handle $\mathcal{H}_{i-1}$ to the same next layer handle $\mathcal{H}_i$ is considered as the same message type $\tilde{X}$.*

$$\tilde{X} = \{x | \forall t \forall x_1, x_2, \mathcal{H}_i(x_1, t) = \mathcal{H}_i(x_2, t)\} \quad (21)$$

$$\mathcal{L}_i(\tilde{X}_i, T_m) = \tilde{Y}_i \quad (22)$$

Lemma III.1. *Discrete counter: The request counter of request flow at any point of time for whole layer $\mathcal{L}_i$ is a discrete non-negative integer, hence 3 following components share the same features:*

- *request sliding-windows counter ($|X_{\mathcal{T}_m}|$)*
- *concurrent value (ψ and Ψ)*
- *speed counter and rate limiter (Θ and $\hat{\Theta}$)*

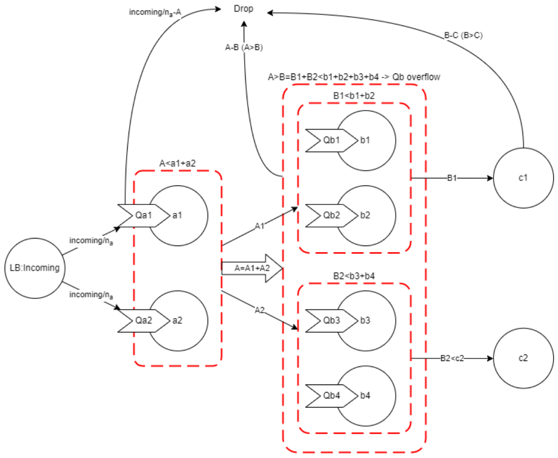


Figure 2. Internal queue analysis

Proof. By the definition III.3, III.5, a sliding windows count request represents a discrete value, then any sliding window-based counter and rate limiter display discrete non-negative integer value. □

Theorem III.2. *Average resource demand: From equation 24 and definition III.4, we have*

$$\hat{d}(\tilde{X}_{T_m}) = \frac{\mathcal{U}\Delta t_m}{|\tilde{X}_{T_m}|} = \frac{\mathcal{U}\Delta t_m}{m+1} = \frac{\mathcal{U}}{\Theta(\tilde{X}, T_m)} \quad (23)$$

For combined load, have:

$$\hat{d}(X_{T_m}) = \frac{\mathcal{U}\Delta t_m}{|X_{T_m}|} = \frac{\mathcal{U}}{\Theta(X, T_m)} \quad (24)$$

Utilizable load $\mathcal{U}$ dynamically changed. Therefore $\hat{d}(X)$ is a function of t.

$$\hat{d}(X) = f(t) \quad (25)$$

Corollary III.2.1. *If max concurrent in process message reduced, then $\hat{d}(X_{\mathcal{T}_m})$ reduce, hence reduce throughput $\Theta(X, T_m)$ due to the deterministic of $\mathcal{U}$*

Theorem III.3. *Drop throughput: $\bar{\Theta}(\mathcal{S}, X_{\mathcal{T}}) = \omega_{\mathcal{S},\mathcal{T}} - \Theta(\mathcal{S}, X_{\mathcal{T}})$*

Definition III.11 (Front-pressure algorithms)**.** *We classify ILB routing algorithms are* ***Front-pressure*** *if they route incoming messages without the ability to slow down the total forwarding function throughput of the producer instance.*

By drop theorem III.3, Front-pressure must be controlled by capacity planning carefully and strictly work with auto-scaling, unless an infinite amount of drop and timeout will occur.

B. Back-pressure modeling

Theorem III.4 (Cascading concurrent condition)**.** *The number of threads in the pool is advised by the business and can be overridden in run-time, therefore dynamic flow control Ψ_i*

parameter can be greater than τ_{i+1}. This denotes concurrent cascading. At any time we have:

$$\forall i \le j, \psi_i \le \Psi_j \tag{26}$$

$$\sup \Psi_{i+1} = \Psi_i \tag{27}$$

Proof.

$$\begin{aligned} \psi_i = m_{\zeta_i^k} + \psi_{i+1} \le \Psi_i \\ \Rightarrow \text{if } \Psi_{i+1} > \Psi_i \text{ , then } \psi_{i+1} \le \Psi_i \end{aligned} \tag{28}$$

On flow with throughput, generalizing we have

$$\Psi_i = \begin{cases} \psi_{i+1} = \upsilon_{i+1} & \Psi_i < \tau_{i+1} \\ \psi_{i+1} + \pi_{i+1} + \tau_{i+1} & \Psi_i \ge \tau_{i+1} \end{cases} \tag{29}$$

Expand to all layers within the flow, we have Cascading concurrent condition □

Theorem III.5 (Cascading throughput condition)**.** *From corollary III.0.1 have cascading throughput formula:*

$$\Theta(\mathcal{L}_i, X_{\mathcal{T}_m}) = \Theta(\mathcal{L}_{i+1}, X_{T'_m}) \tag{30}$$

with T'_m is the time window $\mathcal{L}_i$ finished transforming load X_i into Y_i without any other load, then

$$\Rightarrow \Theta(\mathcal{L}_i, X_{\mathcal{T}_m}) = \begin{cases} \Theta(\mathcal{L}_{i+1}, Y_{\mathcal{T}^s_m}) & = \frac{m_i}{\Delta t} \textit{ if } m_i < \phi_{i,\mathcal{T}_m} \\ \phi_{i,\mathcal{T}_m} & = \frac{m_r}{\Delta t} \textit{ if } m_i \ge \phi_{i,\mathcal{T}_m} \end{cases} \tag{31}$$

$$\Leftrightarrow \Theta(\mathcal{L}_i, X_{\mathcal{T}_m}) = \frac{\min(m_i, m_r)}{\Delta t} = \Omega_{i,\mathcal{T}_m} \tag{32}$$

that $\Omega_{i,\mathcal{T}_m}$ is response throughput of layer i. From equation 30 and corollary III.0.1, have explicit and recurrent relation formula of response throughput:

$$\Rightarrow \Theta(\mathcal{L}_{i+1}, X_{\mathcal{T}_m}) = \min(\Omega_{i,\mathcal{T}_m}, \Omega_{i+1,\mathcal{T}_m}) \tag{33}$$

Therefore layer throughput if formalized as

$$\Rightarrow \phi_{\mathcal{L}_i,\mathcal{T}_m} = \min(\Theta(\mathcal{L}_i, \mathcal{T}_{m_1}), \phi_{\mathcal{L}_{i+1},T^m_{m_1}}) \tag{34}$$

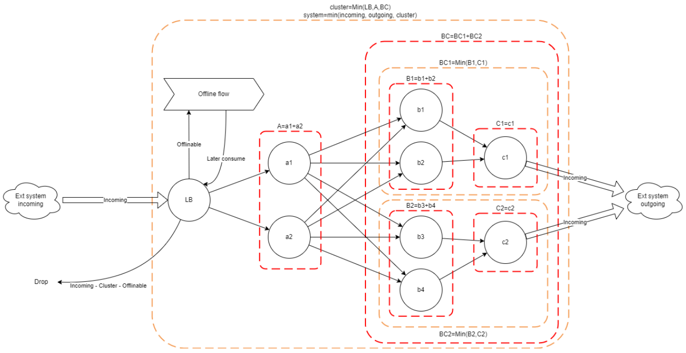


Figure 3. System throughput cascading

Corollary III.5.1. *Expand equation 34 to multiple load system, throughput over a back-pressure controlled system equivalent to current throughput to slowest layer. We have the cascading capacity formula:*

$$\phi_{\mathcal{S}_{.,..,p,q},T_m} = \min_{i\in[p,q]} \phi_{\mathcal{L}_i,X_{\mathcal{T}^{m_j}_{m_i}}} \quad \forall 0 < m_i < m_j \le m \tag{35}$$

This proved the cascading characteristic of back-pressure controlled flow that propagates the pressure (the $\min(\Omega_i)$ function) along the services chain.

Corollary III.5.2. *Congestion only occurs at the weakest layer which is the bottleneck of the current system state.*

Definition III.12. *Define pressure $\mathcal{P}^w_i$ of service flow w at a layer i is the rating between the overhead ξ of resource time actual demand over natural time on a single load. Pressure is based on a single load so it is identified by location of service $\mathcal{L}_i$, more specific is ζ^w_i. Then denote Pressure function $\mathcal{P}$ of a requests handled by function w at time t as:*

$$\xi(x^w_i, t) = d(x^w_i, t) - \hat{d}(x^w_i, t) \Rightarrow \mathcal{P}(x^w_i, t) = \frac{\xi(x^w_i, t)}{\hat{d}(x^w_i, t)} \tag{36}$$

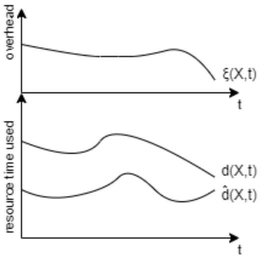


Figure 4. Pressure

From formula 24 and 36, we always have $d(x^w_i, t) \le \hat{d}(x^w_i, t)$, therefore any pressure is non-negative. Then the pressure of the whole flow is integral to the pressure function.

$$\mathcal{P}(X_{\mathcal{T}_m}) = \int_{\mathcal{T}_m} \frac{\xi(x,t)}{\hat{d}(x,t)} \mathrm{d}t = \int_{\mathcal{T}_m} \frac{d(x,t)}{\hat{d}(x,t)} \mathrm{d}t - t \tag{37}$$

Corollary III.5.3. *Performance rating is a function of time, inversely proportional to pressure.*

$$\begin{gathered} \rho(X) = \frac{1}{\frac{\partial \mathcal{P}(X)}{\partial t} + 1} \in (0, 1] \\ \rho \to 1 \Leftrightarrow \mathcal{P}(X) \to 0 \end{gathered} \tag{38}$$

When $\rho \ge \varrho$, throughput decreases, hence maximum throughput performs on saturation point.

Corollary III.5.4. *When back-pressure is locked by reaching a configured limit, it propagates upward and makes HoL locking, nearly immediately reducing the throughput of the whole system.*

$$\begin{bmatrix} \Theta(\mathcal{L}_i) = \min\limits_{j\in(i,q]} \omega_j \\ \psi_i \le \Psi_j = 0, \forall i < j \end{bmatrix} \to \begin{cases} \Theta(\mathcal{L}_1) = 0 \\ \psi_1 = 0 \end{cases} \to \begin{cases} \Theta(\mathcal{S}) = 0 \\ \psi_{\mathcal{S}} = 0 \end{cases} \tag{39}$$

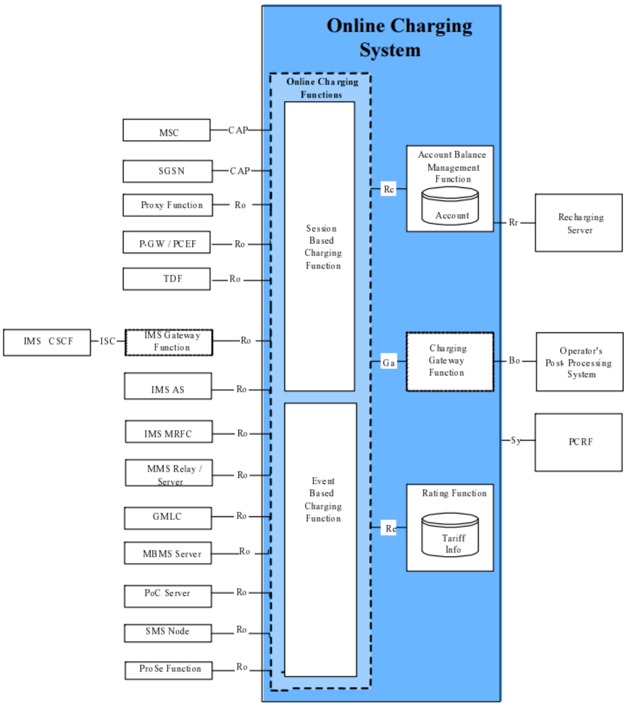


Figure 5. OCS architecture

Corollary III.5.5. *Back pressure is propagated directly when connections use synchronous or blocking actions. Asynchronous with enqueue and drop behavior or non-blocking actions lose pressure information, hence needed Cascading concurrent or throughput condition to form up an algorithm.*

Corollary III.5.6. *Back-pressure features: According to definition III.1, Back-pressure presents in both way of request and response. Back-pressure has a delay cascading effect due to queued layers. Losing cascading aid messages increases delay gaps. Back-pressure controller remove the need for sub-instance group divisions all can be connected in full mesh schema.*

Corollary III.5.7. *BTC control throughput policy: due to equation 33, the throughput of a layer is the minimum throughput of all under layers. The maximum of configuration throughput $\hat{\Theta}$ is:*

$$\max_{\mathcal{L}_i} \hat{\Theta} \leq \min_{j \in [i+1,q]} \phi(\mathcal{L}_j) \tag{40}$$

However on dynamic run-time, we cannot determine a strong $\phi(\mathcal{L}_j)$ for any layers, then we should use JIT value of this formula:

$$\Rightarrow \max_{\mathcal{L}_i} \hat{\Theta} \leq \min_{j \in [i+1,q]} \Theta(\mathcal{L}_j) \tag{41}$$

IV. vOCS Microchassis ILB

A. OCS features

OCS is a 24/7 service that receives billing orders from external systems. Messages incoming to the OCS are structured according to the OCS interface specification [18]. Knowing that network components are high-performance, HA, and transparent with applications. OCS needs to fastest handling charging requests in any conditions at the application layer. With the characteristic of reserving business flow, OCS shows clear large rising waves, instead of peaks.

B. ILB implementation

Microchassis is a wrapped layer of Netty connections, control per-flow with local pressure information. Applications embedded lib Microchassis communicate with each other by TCP/IP protocol and HTTP1.1. Microchassis manages sending, monitoring, service discovery, establishing connections, flow control, and Front-pressure CB, flow/packet tracing, interacting with VNF vMANO as the control plane of vOCS systems. Even if an instance makes multiple connections to a specific destination, each connection decides isolated. This brings the mesh ability to localize errors and isolate jammed instances. With cross-service (cascading) pressure, although we enable auto-read of Netty inbound handler chain, however on a robust physical network, servers still represent congestions, The client side needs to check channel buffer isWritable() continuously before writtenAndFlush().

At the scope of ILB at the application layer, some threshold metrics that could be used are (1) the number of requests/objects and (2) the size of something: total allocated buffer, some active queue size, CPU time, etc. As mentioned in 1, we do not implement time-based algorithms like shortest delay, response time, or any statistical strategy based on it.

According to corollary III.5.5, we implemented two methods BCC and BTC along with two basic RR and LCC for dynamic switching strategies on ILB:

- Round-robin: based on current destinations, the pointer shifts over the list each time gets a new target. The complexity of operating is O(1) and complexity of preparation is o(n). When the destination list changes, we have to re-classify active/writable connection as a projectile snapshot and then shift over this list. RR only combined with speed increase-adjusting when the instance boots up or re-opens after the circuit broke without slow-down control. By accepting any incoming and not slowing down actively, RR is classified as a front-pressure strategy. The weighted extension of RR faces the same problems.
- Least concurrent: based on current concurrent on all connections. LCC maintains the same balanced amount of concurrent among connections. A high value of concurrent reflex global jamming situation but not identify it exactly. The complexity of selection is o(n) and preparation is o(n). The result of LCC can illustrate for P2R[1] and any least-X strategy.
- Back-pressure flow control based on concurrent state (BCC): is a flow control that senses for connection's concurrent state and decides that the connection should increase or reduce the concurrent. Not directly reducing

[1] P2R only chooses better but does not prevent any pressure however adds unneglectable complexity to the implementation.

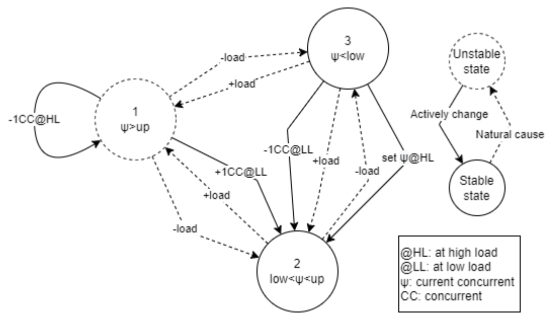


Figure 6. FSM of BCC

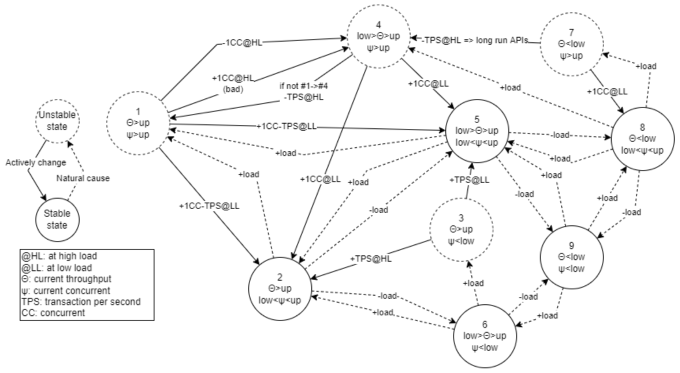


Figure 7. FSM of BTC

the speed but as shown in corollary III.2.1, BCC can control speed to be slow down by a decrease in configured max concurrent in-flight messages. The weakness of BCC is the chance to lock HoL. Transitions of BCC are illustrated in fig 6.

- Back-pressure flow control based on flow speed and concurrent state (BTC): a flow control utilized throughput and concurrent state as its score. Transitions of BTC are illustrated in fig 7. BTC shares the same complexity with BCC however direct control allows outcome throughput.

C. Back-pressure base on throughput and concurrent states

By two nature of a flow is unit speed and bandwidth, BTC propagated throughput and concurrent pressure to the HoL by sensing on each connection, on each cachable routing decision. BTC calculates scores of each flow based on an average of these two utilizations and selects a group of selectable destinations. To be sent, the message must wait through all conditions including concurrent, throughput, and itself computing time: $wait = max(wait_{\Psi_i}, wait_{\hat{\Theta}(\mathcal{L}_i)}, wait_{compute})$. Expand to uncountable request flow, if we raise Ψ_i and $\hat{\Theta}(\mathcal{L}_i)$ up on well affordable instance, the total wait for each request only be computed time. We use 20% Ψ gap to minimize the HoL blocking effect of BCC and BTC.

Resolve loop configure and state locking The following loops may occur in the mesh:

- State 7 was short-lived. If state 7 increases Ψ when the load is low, it transitions to state 8. If there were a high load then it returned to state 7. The 7-8-7 loop reduces throughput hence having a change to go to state 4. This case is self-solvable.
- State 1 when increasing Ψ leads to state 4 or state 5. State 5 is a stop state. When state 1 switches back and forth with state 4: we need to break to not lead to the configure adjustment loop.
  - The chain 1-4-1 results in increasing Ψ and reduced throughput that finally comes to state 2 self-end the loop with a suitable Ψ value. However in an incident, this loop increases Ψ while the consumer is overloading, this can lead to collapsing the whole service layer. So we remove increasing Ψ behave at state 1 in a high load situation.
  - The chain 1-4-1 reduces Ψ and $\hat{\Theta}$, which reduces configuration value continuously, cause: (1) throughput bottleneck then get to stop state 2, find out suitable throughput for instance based on preset Ψ or (2) bottleneck on Ψ then go to state 4, Ψ jamming won't trigger reduce throughput to make a race loop, but increase Ψ again to convergent to stop state 2 or 5.

V. Experimentation and discussion

A. Experiment

We benched performance for four methods: RR, LCC, BCC, and BTC. RR and LCC are the most popular methods in the group of front pressure methods, with RR being the naive, simplest, and most effective option for low-load systems. LCC is a method of grouping "least-X" over a metric X. BCC is the most basic for back pressure over a single metric method. In contrast, BTC is a testing method that combines back-pressure over multiple metrics and can be the base for tuning and learning flow control algorithms.

In this simulation, the database is simulated by an echo server. Inner applications are homogeneous single-flow instance layers. API gateway (LB) is a modified application that has generator threads, that generate random fields 4KB messages in ByteBuf format into LB incoming queue. The result at LB gives us a view of front-line queue status and send cycle time statistics (over client component of LB) as the latency a physical LB must wait: *RTT = InSQ:avgTime + SendCy:avgTime*. We divided our scenarios into 2 dimensions:

- Sizing dimention: 2x 3x 4x 6x 8x with 1x = 1 vCore CPU + 1.5GB physical memory
- Load dimention: underload/overload.

With each scale, application configures like handleThread, Xmx, CPU factor, ... scaled proportionally. Each service layer is deployed into 1 specific non-overlapping physical node to force all requests must travel over the physical network rather than some travel internal by localhost. Table 1 lists the configurations of 9 deployed servers in this study. To archive 99% of the confident level at 5% margin error of unlimited population, we collect 666 data samples with 1s windows since the first 100 skipped, matched timestamps from data points

TABLE 1. Hardware for Benchmark

	Servers
CPU	Dual Intel (R) Platinum 8260 @ 2.40GHZ
RAM	256GB 2933 MT/s
Network	BCM57414 NetXtreme-E 10Gb/25Gb RDMA
Disk	Dell Express Flash NVME P4610 1.6TB SFF
OS	RHEL 7.8
Virtualization	Kubernetes with Calico network

spanned among layers instances. Log4j logging is enabled for statistics and errors. Statistics is accounting per interval (1s) by Windows.

B. Results and evaluations

2 cores based multiple sizing underload test: Underload at 9kTPS in total 27x of 9 instances with each type spawn 3 pods. The result in table 2 shows that all methods are qualified with minor differences in underload test.

TABLE 2. Throughput test 9kTPS 27x=(2x3+3x3+4x3) 1 connections
(*avgTime: μs)

Strategy	RR	LCC	BCC	BTC
algo*	5.04	12.9	9.06	9.99
str*	7.54	15.66	12.18	14.57
InSQ*	6.64	6.96	6.99	6.72
SendCy*	1020	1095	1090	1067
InSQ/RTT	0.6%	0.6%	0.6%	0.6%

1) 4 cores based multiple sizing overload test:: Overload with 172kTPS on 36x total size for 6 instances in multiple types of base 4 cores show results in table 3

TABLE 3. Throughput test 172kTPS 36x=(4x2+6x2+8x2) 3 connections
(*avgTime μs, **avgTime ms)

Strategy	RR	LCC	BCC	BTC
algo*	5.02	15.34	8.08	10.68
str*	62.36	19.46	729.63	22.44
InSQ*	2820	58	21,237	518
SendCy*	29,966	5,977	10,664	5,088
RTT**	32.79	6.04	31.9	5.61
InSQ/RTT	8.6%	1%	66.6%	9.2 %

2) 4 cores based same sizing overload test:: We test again overload test with the same total sizing but divided into 6 identical 6x instances. LCC and BTC deliver outperforming latency, BTC is the same as we tested at minimum throughput (500TPS as a base comparison). This shows an undeniable ability to maintain the best quality at high throughput to obtain the vision of corollary III.5.3. Although resource utilization and most metrics are no different from BCC, BTC sacrifices about 5% throughput in exchange for a few times latency reduction showing superior performance in congestion control.

C. Discussion

URLLC defines 1-way latency for data plane less than 1ms. Within 5 layers, we have 4 inner connected layers, with 2-way connections to an external (source) system. Therefore, 10ms is the upper bound of the transmit latency. As we tested, the proposed back-pressure algorithm BTC and the basic front-pressure LCC fulfill URLLC requirements. The front-pressure algorithms outperformed back-pressure ones. Latency in the front-pressure setups absorbed incoming pressure, which was reflected in the unacceptable RTT, avgTime:SendCy, and avgTime:InSQ to tens of milliseconds. Although LCC tries to find the best target, in the cases of sending to degraded layers, the LCC strategy does not solve the problem adequately. In contrast, back-pressure strategies detect stronger instances in the underlying layers. BTC balances the quality among individual targets, thanks to the throughput-throttling mechanism instead of concurrent-throttling on Ψ value in BCC. BCC struggles with small streams when $\min \Psi_i = 2$. Among front-pressure results, LCC proved itself to be a great competitor to RR. If the load is mostly lower than the capacity, utilizing back-pressure anywhere in the mesh leads to resources wasted. Front-pressure strategies are thus the most effective. As shown in figure 8, there was a cross between increasing performance efficiency due to more flow control policy, and decreasing resource usage pre-sized by the front-pressure layers. Front-pressure layers rely on pre-allocating and auto-scaling instances to prevent incidents caused by any throughput changes. Besides, back-pressure can maintain the throughput near the instance's maximum capacity and guarantee QoS during scaling events, by mitigating message overflow into a manageable early drop reply stream, until more capacity is provisioned.

TABLE 4. Throughput test 172kTPS 36x=(6x6) 3 connections
(*avgTime μs, **avgTime ms)

Strategy	RR	LCC	BCC	BTC
algo*	4.97	15.52	7.58	10.95
str*	33.68	19.65	862.59	22.08
InSQ*	1,644	51	66,510	1,055
SendCy*	21,334	4,070	33,412	5,121
RTT**	22.98	4.12	99.92	6.18
InSQ/RTT	7.2%	1.2%	66.6%	17.1 %

Expansion in the next OCS generation will require more robust methods. In light of this, LCC is currently the best internal strategy for the mesh and BTC deserves to be placed at the head of flow.

Some limitations in the scope of the study are:

- Research has not considered the effect of queue size.
- Microchassis limitation: Currently ILB in Microchassis only supports HTTP and TCP messages for single requests, not yet support RPC-based connection and persistent HTTP stream.

VI. Conclusion

To obtain a smooth 5G real-time charging service, we use the most latency-effective approach - connection lib - as the integrated chassis for the mesh. In the paper, we evaluated connection lib with four routing strategies RR, LCC, BCC, and BTC and determined that the lib satisfied URLLC requirements under various conditions. Among tests, LCC is

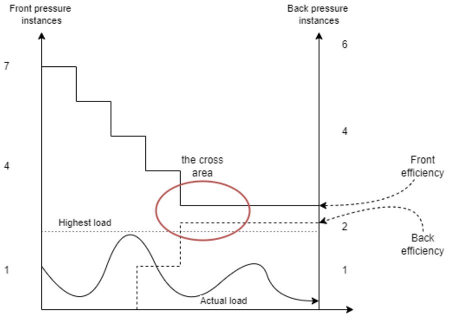


Figure 8. Effective trade-off cross

the most suitable for internal routing with higher resource efficiency, while BTC should be used for border gateways to maintain overall QoS. In the future, we will continue to enhance our strategies, along with tuning algorithms to better adapt edge cases. In particular, we will expand the problem to include geo-location and the ability to work with centralized solutions.

References

[1] I. T. Union. Minimum requirements related to technical performance for imt-2020 radio interface(s). [Online]. Available: https://www.itu.int/pub/R-REP-M.2410

[2] A. Nasrallah, A. S. Thyagaturu, Z. Alharbi, C. Wang, X. Shao, M. Reisslein, and H. ElBakoury, "Ultra-low latency (ull) networks: The ieee tsn and ietf detnet standards and related 5g ull research," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 88–145, 2018.

[3] Istio. Performance and scalability. [Online]. Available: https://istio.io/latest/docs/ops/deployment/performance-and-scalability/#data-plane-performance

[4] Envoy. Supported load balancers. [Online]. Available: https://www.envoyproxy.io/docs/envoy/latest/intro/arch_overview/upstream/load_balancing/load_balancers#arch-overview-load-balancing-types

[5] X. Zhu, G. She, B. Xue, Y. Zhang, Y. Zhang, X. K. Zou, X. Duan, P. He, A. Krishnamurthy, M. Lentz *et al.*, "Dissecting service mesh overheads," *arXiv preprint arXiv:2207.00592*, 2022.

[6] P. Goyal, P. Shah, N. K. Sharma, M. Alizadeh, and T. E. Anderson, "Backpressure flow control," in *Proceedings of the 2019 Workshop on Buffer Sizing*, 2019, pp. 1–3.

[7] F5. K42275060: There are several load balancing methods. which one is best for your environment? [Online]. Available: https://my.f5.com/manage/s/article/K42275060

[8] I. Alexandre Cassen. Ipvs scheduling algorithms. [Online]. Available: https://keepalived-pqa.readthedocs.io/en/latest/scheduling_algorithms.html

[9] N. Owen Garrett, F5. Nginx and the "power of two choices" load-balancing algorithm. [Online]. Available: https://www.nginx.com/blog/nginx-power-of-two-choices-load-balancing-algorithm/

[10] H. Willy Tarreau. Test driving "power of two random choices" load balancing. [Online]. Available: https://www.haproxy.com/blog/power-of-two-load-balancing

[11] M. Welsh and D. Culler, "Adaptive overload control for busy internet servers," in *4th USENIX Symposium on Internet Technologies and Systems (USITS 03)*, 2003.

[12] N. K. Apache Flink. A deep-dive into flink's network stack. [Online]. Available: https://flink.apache.org/2019/06/05/a-deep-dive-into-flinks-network-stack/#credit-based-flow-control

[13] Uber. Hyperbahn. [Online]. Available: https://github.com/uber-archive/hyperbahn

[14] Netflix. Ribbon. [Online]. Available: https://github.com/Netflix/ribbon

[15] H. Saokar, S. Demetriou, N. Magerko, M. Kontorovich, J. Kirstein, M. Leibold, D. Skarlatos, H. Khandelwal, and C. Tang, "{ServiceRouter}: Hyperscale and minimal cost service mesh at meta," in *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, 2023, pp. 969–985.

[16] M. R. S. Sedghpour, C. Klein, and J. Tordsson, "Service mesh circuit breaker: From panic button to performance management tool," in *Proceedings of the 1st Workshop on High Availability and Observability of Cloud Systems*, 2021, pp. 4–10.

[17] F. S. Samani and R. Stadler, "Dynamically meeting performance objectives for multiple services on a service mesh," in *2022 18th International Conference on Network and Service Management (CNSM)*. IEEE, 2022, pp. 219–225.

[18] 3GPP. Ts 32.296: Telecommunication management; charging management; online charging system (ocs): Applications and interfaces. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1913

Ngoc Tien Nguyen received his B.Sc. degree in Information System in 2020 from University of Engineering and Technology - Vietnam National University, Hanoi. He is currently a software engineer in Viettel High Technology, Viettel Group. He has industry experience with databases and high-performance computing, especially for telecommunications.

Thanh Son Pham received his B.Sc. degree in Mathematics and Informatics Engineering from Hanoi University of Science and Technology, Vietnam in 2018. He is currently a software engineer in Viettel High Technology, Viettel Group. He has industry experience with micro-service and high-performance computing, especially for telecommunications.

Van Duong Nguyen received his B.Sc. degree in Mathematics and Informatics Engineering from Hanoi University of Science and Technology, Vietnam in 2015. He is currently manage a database team of the Research Center for OCS (Online Charging Platform) of Viettel High-Tech Industry Corporation (VHT), Viettel Group. He has many years of experience in research and development of 5G core systems such as online charging system. His research interests include telecommunications core networks and database technologies.

Cong Dan Pham received his PhD. Degree in Probability and Statistics from Aix-Marseille University, France in Jun 2014. He has academic experience in universities like Aix-Marseille University, Hanoi University of Education, and industry experience companies like Viettel Group. He has published various journal and conference papers in IEEE and ScienceDirect. His research interests include probability and statistics, data science and machine learning.

Duc Hai Nguyen received his B.Sc. in Mathematics and Informatics Engineering from Hanoi University of Science and Technology, Vietnam in 2009. He is currently director of the Research Center for OCS (Online Charging Platform) of Viettel High-Tech Industry Corporation (VHT), Viettel Group. He has many years of experience in research and development of 5G core systems such as online charging system. His research interests include telecommunications core networks, Cloud Computing.

AppTest: Assessing the Usability and Performance Efficiency of BOSESKO for Digital Participation

Jennifer L. Llovido*, Michael Angelo D. Brogada*, Lany L. Maceda*, Mideth B. Abisado**

*Computer Science and Information Technology Department, Bicol University, Legazpi City, Philippines

**National University, Manila, Philippines

jllovido@bicol-u.edu.ph, madbrogada@bicol-u.edu.ph, llmaceda@bicol-u.edu.ph, mbabisado@national-u.edu.ph

Abstract— **Software systems need a proper evaluation to avoid errors and problems. This research aims to showcase an evaluation method to test an integrated mobile and web application for gathering valuable feedback from Filipinos nationwide. This application, named "BOSESKO: Building on Opinions and Sentiments for Sustainability and Knowledge Opportunities" – a multilingual, inclusive, deliberative, synoptic, digital participatory toolkit, is intended to act as a central hub to solicit insights on both existing and upcoming policies in the Philippines. AppTest pertains to a comprehensive evaluation of the BosesKo software system. This research hopes to identify the strengths and weaknesses of BOSESKO and improve it, which would serve as a tool to potentially enhance civic engagement in the country by actively involving the general public in policy-making. This study employed a mixed-method approach that seamlessly integrates manual and automated testing techniques. The assessment of BOSESKO is aligned with the ISO 25010 standard, focusing on evaluating usability and performance efficiency. To delve into the usability of BOSESKO, a usability survey was administered to gauge user perspectives on (a) learnability, (b) appropriateness, (c) recognizability, (d) operability, (e) error protection, (f) user interface aesthetics, and (g) accessibility. Additionally, in-depth performance evaluations were conducted using various automated tools, including WebPageTest, PageSpeed Insight, Google Lighthouse, Yellow Lab, and Pingdom. This combination of manual and computerized methodologies provided a robust and thorough analysis of BOSESKO and, in turn, offered valuable insights into the functionality and efficiency of the developed system. As of the time of this writing, about 10,500 users across the Philippines have already engaged with BOSESKO. Initial assessment results of BOSESKO revealed an acceptable user rating with an average weighted mean of 4.36 in terms of usability. Meanwhile, the applied automated tests yielded positive performance scores of 89% and 76% for the web and mobile platforms. However, the evaluation results suggest image and search engine optimization. Overall, these results underscore the robust foundation established by BOSESKO, indicating its considerable potential and strong starting point.**

Keywords— **usability test, efficiency test, automated testing, ISO25010, software performance**

I. Introduction

The completion of a fully developed system, inclusive of all its features and functionalities, does not mark the end of the entire software development process. It will further require extensive testing and evaluation to assess its readiness for deployment and public use. Software quality in industrial applications, or even in everyday life, is critical for both developers and users [1]. Software testing is essential for analyzing software quality, minimizing project costs, and ensuring timely delivery. In essence, software testing involves an in-depth evaluation, either manually or by automated methods, to ensure that the system meets outlined standards [2].

Manual testing includes usability inspections, which rely on professional analysis and theoretical concepts supported by study or observation. These inspections frequently use heuristic assessments, cognitive pathways, and rules/checklists to analyze specific navigational faults or behaviors. Furthermore, user-centered procedures that involve users in testing, interviews, and physiological exams are also part of this [3] where the quality of the developed system is gauged by customer satisfaction and approval [2]. Automated testing, on the other hand, addresses the constraints of manual testing by providing repeatable test cases, which is crucial in the increasing complexity of software development. Using automated testing methods to evaluate application quality and functionality has become critical, resulting in increased reliability, reusability, cost savings, greater coverage, and faster results when compared to manual testing [4].

Historically, product quality evaluation was largely concerned with functionality. As the complexity of robust system development has grown over time, an expansion of characteristics and sub-characteristics has been incorporated to comprehensively and rigorously evaluate software quality [5]. Software product quality models have progressed in their ability to adequately represent and convey the abstract concept of software quality since the 1970s [6]. One of these is the most recent standard, ISO 25010 Quality Model, which is deemed to have more features than the other quality models combined [7].

The ISO/IEC 25010 standard within the Quality Model division establishes quality requirements for software development products in eight domains, including functional suitability, performance efficiency, compatibility, usability, reliability, security, maintainability, and portability [1], [8]. This more sophisticated model provides a holistic view of software quality, supporting companies in obtaining a deeper grasp of user requirements. It supports well-informed decision-making by allowing for a more complete assessment of the product's strengths and weaknesses in order to improve software quality [5].

Furthermore, this research study seeks to assess the quality of the BOSESKO system—Building on Opinions and Sentiments for Sustainability and Knowledge Opportunities, by integrating the ISO 25010 standard quality model. The evaluation was conducted using a combination of manual and automated testing techniques, consolidating user feedback with statistical data obtained by the automated tools in use.

II. Literature Review

In the field of software development, software testing is widely considered as an essential phase for ensuring that the system conforms to user requirements and specifications [4]. The improvement of product quality is significantly influenced by the pivotal role of software testing. Consequently, various software testing models have been established to set a standard for assessing software products [9]. Quality software is free of errors, specifically bugs and vulnerabilities, such as missed or misinterpreted requirements and errors in design, functional logic, data linkages, process timing, validity checking, and coding errors. The second, more subtle, factor is that the system can handle its technical debt and meets various quality and quality attribute characteristics [10]. There are plenty of quality characteristics that describe the system and serve as the basis and cornerstone for the evaluation of a product's or software's quality. The quality model refers to the collection of qualities used as the foundation for evaluation [11].

An international standard known as ISO/IEC 25010 is used to assess or gauge the quality of software both internally (the software itself) and externally (the software as it relates to hardware, data, and other software integrated in a computer system) [12]. It has been utilized by several researchers in their research for its universality that makes it simple to adapt to the creation of numerous specialized software quality models [13]. This standard breaks software quality down into many sub-aspects, enabling developers to create custom metrics and measurements to precisely evaluate both functional and non-functional needs [14].

According to International Organization for Standardization (ISO) / International Electrotechnical Commission (IEC), there are eight quality attributes that make up the product quality dimension: Functional suitability, Performance efficiency, Compatibility, Usability, Reliability, Security, Maintainability, and Portability [15]. Functional suitability pertains to the degree to which software can offer features that are suitable for use in specific circumstances [16]. Efficiency is the capacity of a user to execute a given activity quickly and accurately, or within the allotted time task [17]. The extent to which a product or system may exchange information with other products, systems, or components while executing a function inside a common environment, hardware, or software is referred to as compatibility [18]. Meanwhile, usability is the extent to which a particular product may be utilized by certain users to accomplish particular goals while taking into consideration effectiveness, efficiency, and happiness in a specific usage context [19]. Reliability equates to how the software can be used according to the required functionality and the intended level of accuracy. Security is a product-level application that provides services to safeguard against unauthorized access, usage, modification, tampering, or other potentially harmful activities that may compromise confidentiality, non-repudiation, and authentication [20]. The capability of a software product to be effectively and efficiently modified is referred to as its maintainability while portability measures the capacity of a product or component to be transferred successfully and efficiently from one hardware, software, or user environment to another [21].

Various researchers such as [8], [22], [23], and [24] have employed ISO 25010 as the foundational model for conducting quality testing on their respective systems or applications. Comparatively, [25] used the ISO/IEC 25010:2011 product quality characteristics in their study, assessing their developed software from two distinct perspectives—user view and technical view. On a technical point of view, the approach to quality measurement is software-oriented, with automated testing and monitoring tools used to evaluate the system's functional suitability, performance efficiency, reliability, security, and portability. In contrast, measuring tends to be more user-oriented in quality test designs based on users' point of view. This entails using questionnaire data extraction methodologies and subjecting the survey tool to validity, reliability, and feasibility testing.

In the context of usability evaluation, both [26] and [27] employed a questionnaire-based method to assess the usability of their respective developed websites. In this approach, participants were required to complete a usability survey with questions or remarks regarding their experiences with the websites.

In contrast, [28] and [29] concentrated specifically on performance efficiency, utilizing automated testing tools to measure this aspect. Load testing, a type of performance testing, is a series of processes for determining load-related vulnerabilities in a system that is being tested [30]. One well-known and open source web performance testing tool is WebPageTest [31], which can be utilized to assess the website's speed, usability, and resilience [32]. Consequently, [33] used the WebPageTest tool in his study, which is acknowledged as the leading open-source tool for assessing three essential web metrics: loading time, page size, and the number of requests. This tool, which has been widely used by many researchers, facilitates the analysis of web traffic data from a large number of Internet users, making it a valuable resource for determining the efficiency of various web systems. Furthermore, this tool can present test findings that are specific and detailed, allowing researchers to analyze the advantages and disadvantages or elements of the website that must be refined [34]. Another automated tool is Google Lighthouse, as highlighted by [35], due to its advanced metric that assesses SEO, best practices, accessibility, and performance. The application of it was demonstrated in the study by [36], that produced comprehensive evaluations with ratings for several audit categories and doable recommendations for improvement. On the other hand, PageSpeed Insights evaluates webpage performance and optimization and offers speed scores for desktop and mobile versions [37]. The utilization of PageSpeed Insights in rating loading times, giving ratings, and making recommendations for speed optimization without sacrificing user experience was highlighted by [38]. Additionally, Pingdom is a useful tool for evaluating the effectiveness of websites. It assesses websites using important parameters such as page requests, load times, performance, and page sizes [39]. This is consistent with the methodology used by [40] in their study, which employed similar categories for evaluating websites. In addition, Pingdom provides useful advice to improve loading times. Its capacity to separate loading speeds for distinct website elements, such as images, CSS, JavaScript, and RSS is highlighted in the study by [39]. Furthermore, Pingdom is available as a free online trial edition [41], which makes it an easy-to-use tool for anyone looking to analyze and enhance the effectiveness of their website.

III. Methodology

The researchers employed the methodological approach implemented by [25] to conduct a thorough assessment of software quality, taking into account both user and technical points of view. The evaluation focused on two ISO 25010 characteristics - the usability and performance efficiency. Usability is categorized under the user perspective, whereas performance efficiency is classified under the technical point of view.

A. *Test Design Based on Technical Perspective*

To assess the performance efficiency of the system, various open-source automated tools such as WebPageTest, Google Lighthouse, PageSpeed Insights, and Yellow Lab were utilized. Additionally, a proprietary software, Pingdom, was also used to check if disparities would emerge in the results compared to those generated from free tools.

B. *Test Design Based on Users' Perspective*

A usability questionnaire was utilized to collect user feedback, encompassing statements that correspond to each sub-characteristic of usability. Specifically, there were six (6) statements addressing appropriateness recognizability, four (4) statements for learnability, five (5) statements for operability, another four (4) statements focusing on user error protection, six (6) statements covering user interface aesthetics, and seven (7) statements pertaining to accessibility. In total, the questionnaire comprised 32 items.

To administer the system testing based on the user's perspective, the usability questionnaire was adapted into an online survey instrument using Google Forms. A total of 346 respondents participated, specifically students aged between 19 and 24, coming from different state universities - who were one of the key users of the BOSESKO application. The demographic composition of the respondents comprised 48.1% women, 46.6% men, and the remaining 5.3% were from the LGBTQ+ community.

In conducting usability analysis, quantitative descriptive analysis was used. This involved calculation of the frequency distribution of scores based on the quantitative values assigned to each category on the five-point Likert scale. This was also used in the computation and interpretation of the weighted mean for each sub-characteristic as shown in Table 1.

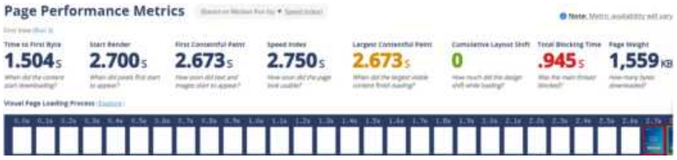
TABLE 1. Likert Rating Scale

Rating	Scale Range	Verbal Interpretation
5	4.50-5.00	Highly Acceptable
4	3.50-4.49	Acceptable
3	2.50-3.49	Moderately Acceptable
2	1.50-2.49	Unacceptable
1	1.00-1.49	Highly Unacceptable

IV. Discussion

A. *Performance Efficiency*

*1) **WebPageTest**:* Figure 1 illustrates the page performance metrics obtained through the WebPageTest tool. One notable observation was the system's slow first-byte time, which is 1.504 seconds, indicating a prolonged server connection and response time. Furthermore, pixels began appearing at 2.700 seconds, with the first contentful paint loading at 2.673 seconds. The speed index was notably slow, requiring 2.750 seconds for the page to appear usable. Additionally, the largest contentful paint (LCP) recorded a high value of 2.673 seconds, attributed to the element involved being an image. Recommendations for optimization were suggested to expedite the image fetching process. On a positive note, there were no instances of cumulative layout shift, and the main thread experienced no significant blockage, taking only 0.945 seconds. Lastly, the page weight was determined to be 1559 KB, highlighting the overall size of the webpage.



Page Performance Metrics

Time to First Byte 1.504s When did the content start downloading?

Start Render 2.700s When did pixels first start to appear?

First Contentful Paint 2.673s How soon did text and images start to appear?

Speed Index 2.750s How soon did the page look usable?

Largest Contentful Paint 2.673s When did the largest visible content finish loading?

Cumulative Layout Shift 0 How much did the design shift while loading?

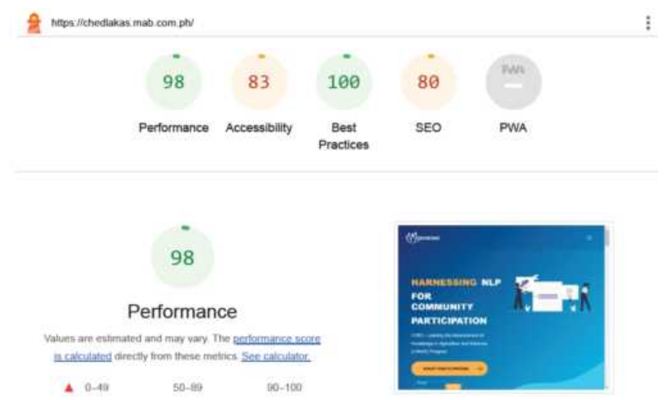
Total Blocking Time .945s Was the main thread blocked?

Page Weight 1,559KB How many bytes downloaded?

Visual Page Loading Process

Figure 1. Web Application Performance Summary Based on WebPageTest

*2) **Google Lighthouse:*** Four (4) metrics, namely performance, accessibility, best practices, and search engine optimization (SEO), were evaluated using this automated tool. Figure 2 shows that the system demonstrated satisfactory results, achieving 98% in performance, 83% in accessibility, and a perfect score of 100% in best practices. However, the SEO metric showed a slightly lower result at 80%. According to the tool's assessment criteria, the system's overall performance falls within the 90-100% range, categorizing it as a good score, while the mobile platform as presented in figure 3, got 76%, which needs improvement.



https://chedlakas.mab.com.ph/

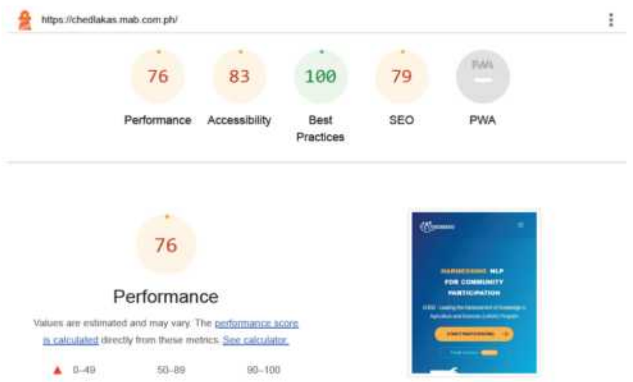
98 Performance
83 Accessibility
100 Best Practices
80 SEO
PWA

98

Performance

Values are estimated and may vary. The performance score is calculated directly from these metrics. See calculator.

Figure 2. Web Application Performance Summary Based on Google Lighthouse



https://chedlakas.mab.com.ph/

76 Performance
83 Accessibility
100 Best Practices
79 SEO
PWA

76

Performance

Values are estimated and may vary. The performance score is calculated directly from these metrics. See calculator.

Figure 3. Mobile Performance Summary Based on Google Lighthouse

3) ***PageSpeed Insights:*** Table 2 presents the results of using PageSpeedInsights which tested the system on both web and mobile platforms. The differences in performance are discernible. The system exhibited a robust performance using the web application with a score of 92, whereas for the mobile platform, it lagged behind with a score of 65. On the other hand, the system demonstrated higher accessibility on mobile devices, yielding an 83 accessibility score compared to the web application with a score of 82. Notably, both platforms excelled in adhering to best practices, securing a perfect score of 100. Examining SEO performance, the web application outperformed the mobile with scores of 80 and 79, respectively. Other metrics, such as first contentful paint, favored the web application, achieving a faster load time of 0.5 seconds compared to the mobile's 1.3 seconds. Moreover, the mobile platform showed considerably higher total blocking time, with 1710 milliseconds, while the web application experienced only 80 milliseconds. Similarly, the speed index favored the web application at 0.9 seconds, contrasting sharply with the mobile's 6.9 seconds. Analyzing the largest contentful paint, the web application achieved a loading time of 1.7 seconds, surpassing the mobile's 2.3 seconds. Meanwhile, no cumulative layout shift occurred for both platforms.

TABLE 2. PageSpeed Insights Results

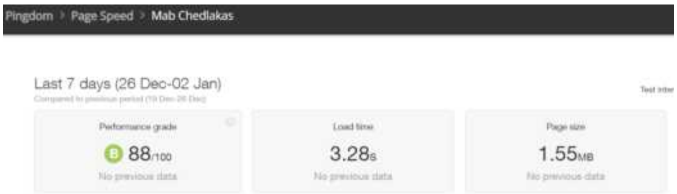
Metrics	BOSESKO Web Application	BOSESKO Mobile Application
Performance	92	65
Accessibility	82	83
Best Practices	100	100
SEO	80	79
First Contentful Paint	0.5 s	1.3 s
Total Blocking Time	80 ms	1710 ms
Speed Index	0.9 s	6.9 s
Largest Contentful Paint	1.7 s	2.3 s
Cumulative Layout Shift	0	0

4) ***Yellow Lab:*** Yellow Lab Tools was another tool utilized for software quality testing in the mobile platform. This automated tool as shown in Table 3 assessed the system in detail by following ten (10) metrics and assigning scores ranging from A as the highest to C as the lowest. The system's performance for the seven metrics, including Requests, DOM Complexity, JS Complexity, Bad JS, jQuery, CSS Complexity, and Web Fonts, was deemed excellent, achieving a mark of A. Additionally, page weight and server configuration received a grade of B. Regarding page weight, a notable inadequacy was observed in image optimization, a generally easy method for reducing page weight and load time. The system could save 255 KB by optimizing 8 images, where one oversized image was identified. Regarding server configuration, issues were detected, including one HTTP protocol offender and 11 caching instances being disabled. Lastly, the Bad CSS metric received a C rating. This was due to an unknown parsing error during testing, leading to miscalculations in other CSS metrics and scores.

TABLE 3. Yellow Lab Tools Score Details

Metrics	BOSESKO Web Application	BOSESKO Mobile Application
Page Weight	A	B
Requests	A	A
DOM Complexity	A	A
JS Complexity	A	A
Bad JS	A	A
jQuery	A	A
CSS Complexity	A	A
Bad CSS	C	C
Web Fonts	A	A
Server Configuration	B	B
Global Score	**A - 90/100**	**A - 89/100**

5) ***Pingdom:*** Aside from employing open-source software tools for testing, the researchers utilized Pingdom, a proprietary software, to assess whether there were significant differences in the results. Ultimately, the result generated from this tool as shown in figure 4, closely aligned with those obtained from other tools, yielding a performance score of 88 out of 100, corresponding to a descriptive mark of B which means efficient.



Pingdom › Page Speed › Mab Chedlakas

Last 7 days (26 Dec-02 Jan)

Compared to previous period (19 Dec-26 Dec)

Test inter

Performance grade	Load time	Page size
B 88/100	3.28s	1.55MB
No previous data	No previous data	No previous data

Figure 4. BOSESKO Web Application Performance Summary Based on Pingdom

6) Performance Efficiency Test Summary: After conducting a series of performance efficiency tests using five distinct automated tools, the generated results were consolidated in Table 4. For the web platform, five tools were considered: WebPageTest at 76%, Google Lighthouse at 98%, PageSpeed Insights at 92%, Yellow Lab Tools at 90%, and Pingdom at 88%, resulting in an average performance score of 88.33%. In contrast, the mobile platform involved only four tools - WebPageTest with a score of 74%, Google Lighthouse with 76%, PageSpeed Insights with a score of 65%, and Yellow Lab Tools with 89%, yielding an average performance score of 76%. The substantial gap between the web application and mobile efficiency is evident. Thus, further improvements may be made, such as enhancing search engine and image optimization.

TABLE 4. Performance Efficiency Test Results

Tools Used	**Performance Score**	
	BOSESKO Web Application	**BOSESKO Mobile Application**
WebPageTest	76%	74%
Google Lighthouse	98%	76%
PageSpeed Insights	92%	65%
Yellow Lab	90%	89%
Pingdom	88%	-
Average Score	**89%**	**76%**

B. *Usability*

Based on the usability test results outlined in Table 5, it was evident that respondents expressed an acceptable level of agreement regarding the system's adherence to usability sub-characteristics. The survey form contains 36 questions, each sub-characteristic having six (6) questions. Notably, the user interface aesthetics received the highest weighted mean of 4.41, closely followed by appropriateness, recognizability, and learnability, scoring 4.40, followed by user error protection with 4.34, then operability with 4.30, and accessibility with 4.28. The respondents somewhat agreed with the system's overall usability, yielding an overall weighted mean of 4.36. Moreover, this result is considered as acceptable based on the criterion previously defined in Table I.

TABLE 5. Usability Test Results

Sub-Characteristics	**Weighted Mean**	**Verbal Interpretation**
Appropriateness Recognizability	4.40	Acceptable
Learnability	4.40	Acceptable
Operability	4.30	Acceptable
User Error Protection	4.34	Acceptable
User Interface Aesthetics	4.41	Acceptable
Accessibility	4.28	Acceptable
Overall Weighted Mean	**4.36**	**Acceptable**

V. Conclusions

In conclusion, this research discusses the results of AppTest, a comprehensive evaluation of the BosesKo software system, a digital participatory toolkit designed to act as a central hub to solicit insights on existing and upcoming policies in the Philippines. The usability analysis involved 10,500 users, predominantly students from various state universities in the Philippines. Initial findings revealed a commendable user rating, with an average weighted mean of 4.36 across different usability sub-characteristics as evaluated by 346 respondents. The system demonstrated effectiveness in user interface aesthetics, appropriateness, recognizability, learnability, operability, user error protection, and accessibility. These results underscore the potential of BOSESKO to actively engage the public in policy-making, thereby enhancing civic participation.

Further, the performance efficiency assessment, which employed various automated tools, including WebPageTest, Google Lighthouse, PageSpeed Insights, Yellow Lab Tools, and Pingdom, shows that the system exhibited satisfactory scores in performance, accessibility, best practices, and search engine optimization with an average score of 89% and 76% for the web and mobile platform respectively.

However, this research identified areas for improvement, particularly in search engine and image optimization, which will serve as actionable points for further development. Future works may also include evaluating the security and portability of the BosesKo system for a more robust and efficient digital participatory toolkit, thereby fostering increased civic engagement in the policy-making process in the Philippines.

Acknowledgment

The researchers would like to thank the Philippine Commission on Higher Education (CHED) Leading the Advancement of Knowledge in Agriculture and Science (LAKAS) Project No. 2021-007, eParticipation 2.1: Harnessing Natural Language Processing (NLP) for Community Participation, for providing the necessary funds to make this research possible. The researchers are truly grateful for their contribution to the research.

References

[1] N. Nyári, & A. Kerti, “Review of software quality related iso standards,” *Biztonságtudományi Szemle*, vol. 3, no. 2, pp. 61-72, 2021

[2] A. Mustafa et al., “Automated Test Case Generation from Requirements: A Systematic Literature Review,” *Computers, Materials & Continua*, vol. 67, no. 2, pp. 1819–1833, 2021, doi: https://doi.org/10.32604/cmc.2021.014391

[3] L. Alonso-Virgós, J. P. Espada, and R. G. Crespo, “Analyzing compliance and application of usability guidelines and recommendations by web developers,” *Computer Standards & Interfaces*, vol. 64, pp. 117–132, May 2019, doi: https://doi.org/10.1016/j.csi.2019.01.004

[4] H. V. Gamido and M. V. Gamido, “Comparative Review of the Features of Automated Software Testing Tools,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 5, p. 4473, Oct. 2019, doi: https://doi.org/10.11591/ijece.v9i5.pp4473-4478

[5] L. Pinedo et al., “Software quality models: Exploratory review,” *ICST Transactions on Scalable Information Systems*, Sep. 2023, doi: https://doi.org/10.4108/eetsis.3982

[6] T. Galli, F. Chiclana, and F. Siewe, “Software Product Quality Models, Developments, Trends, and Evaluation,” *SN Computer Science*, vol. 1, no. 3, May 2020, doi: https://doi.org/10.1007/s42979-020-00140-z

[7] M. Izzatillah, “Quality measurement of transportation service application Go-Jek using ISO 25010 quality model,” *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 10, no. 1, pp. 233-242, 2019

[8] M.R.A. Assifa, F. Setiadi, and R.G. Utomo, “EVALUATION OF SOFTWARE QUALITY FOR I-OFFICE PLUS APPLICATIONS USING ISO/IEC 25010 AND KANO MODEL,” JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 8, no. 2, pp. 561–571, May 2023, doi: https://doi.org/10.29100/jipi.v8i2.3561

[9] K. Hrabovská, B. Rossi, and T. Pitner, “Software Testing Process Models Benefits & Drawbacks: a Systematic Literature Review,” Jan. 2019. Available: https://arxiv.org/pdf/1901.01450.pdf.

[10] I. Ozkaya, “Can We Really Achieve Software Quality?,” ieeexplore.ieee.org, 2021. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9407296

[11] S. Basaran & R.K.H. Mohammed, “Usability evaluation of open source learning management systems,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, 2020.

[12] R. Gustriansyah, N. Suhandi, J. Alie, F. Antony, & A. Heryati, “Optimization of laboratory application by utilizing the ISO/IEC 25010 model,” *In Iop Conference Series: Materials Science And Engineering, IOP Publishing*, vol. 1088, no. 1, p. 012067, Feb. 2021.

[13] E. Peters and G. K. Aggrey, “An ISO 25010 Based Quality Model for ERP Systems,” *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 2, pp. 578–583, 2020, doi: https://doi.org/10.25046/aj050272

[14] K. Moumane, A. Idri, and F. El Aouni, “ISO/IEC 25010- based Quality evaluation of three mobile applications for reproductive health services in Morocco,” Mar. 2023, doi: https://doi.org/10.21203/rs.3.rs-2720323/v1

[15] N. Angraini and A. Kurniawati, “Comparative Analysis of Fintech Software Quality Against MSMEs Using the ISO 25010:2011 Method,” I*nternational Research Journal of Advanced Engineering and Science*, vol. 6, no. 3, pp. 167–175, 2021, Available: https://irjaes.com/wp-content/uploads/2021/08/IRJAES-V6N3P145Y21.pdf

[16] B. Nugeraha and A. Kurniawati, “Quality Analysis of Access KRL Applications Use Method ISO 25010:2011,” *International Research Journal of Advanced Engineering and Science*, vol. 5, no. 3, pp. 233–240, 2011, Available: https://irjaes.com/wp-content/uploads/2020/10/IRJAES-V5N3P398Y20.pdf

[17] Sunardi, G. F. P. Desak, and Gintoro, “List of Most Usability Evaluation in Mobile Application: A Systematic Literature Review,” *IEEE Xplore*, Aug. 01, 2020. doi: https://doi.org/10.1109/ICIMTech50083.2020.9211160. Available: https://ieeexplore.ieee.org/abstract/document/9211160

[18] A. Yulianty and A. Kurniawati, “Quality Analysis of Bios Portal Website at Banking Companies Using ISO / IEC 25010:2011 Method,” *International Research Journal of Advanced Engineering and Science*, vol. 6, no. 2, pp. 11–16, 2021, Available: https://irjaes.com/wp-content/uploads/2021/03/IRJAES-V6N2P16Y21.pdf

[19] D. Hariyanto, Moch. B. Triyono, and T. Köhler, “Usability evaluation of personalized adaptive e-learning system using USE questionnaire,” *Knowledge Management & E-Learning: An International Journal*, vol. 12, no. 1, pp. 85–105, Mar. 2020, doi: https://doi.org/10.34105/j.kmel.2020.12.005

[20] S. Budi, W. Gata, M. Noor, S. Pangabean, and C. S. Rahayu, “News Portal Website Measurement Analysis Using ISO/IEC 25010 And McCall Methods,” *Journal of Applied Engineering and Technological Science (JAETS)*, vol. 4, no. 1, pp. 273–285, Oct. 2022, doi: https://doi.org/10.37385/jaets.v4i1.1094

[21] D. Mena and M. Santórum, “Maintainability and Portability Evaluation of the React Native Framework Applying the ISO/IEC 25010,” *Advances in intelligent systems and computing*, pp. 429–439, Oct. 2020, doi: https://doi.org/10.1007/978-3-030-59194-6_35

[22] M. S. D. Sutadewi and T. Yusnitasari, “Quality Analysis of PeduliLindungi Application using ISO 25010:2011,” *International Research Journal of Advanced Engineering and Science*, vol. 7, no. 1, pp. 212-216, 2022, Available: https://irjaes.com/wp-content/uploads/2022/03/IRJAES-V7N1P143Y22.pdf.

[23] F. Handayani and M. Mustikasari, “Quality Assurance of Sayurbox Mobile Application Using Model ISO 25010,” *International Research Journal of Advanced Engineering and Science*, vol. 6, no. 3, pp. 176–180, 2021, Available: https://irjaes.com/wp-content/uploads/2021/08/IRJAES-V6N3P148Y21.pdf.

[24] A. A. Pratama and A. B. Mutiara, “Software Quality Analysis for Halodoc Application using ISO 25010:2011,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, 2021, doi: https://doi.org/10.14569/ijacsa.2021.0120844

[25] D. A. Suryadi and E. Sulistiyani, “Evaluation of Information Quality Using ISO/IEC 25010:2011 (Case Research: Menu Harianku Application),” *International Journal of Innovation in Enterprise System*, vol. 6, no. 02, pp. 143–156, Jul. 2022,. Available: https://ijies.sie.telkomuniversity.ac.id/index.php/IJIES/article/view/167

[26] C. Chang and H. Almaghalsah, “Usability evaluation of e-government websites: A case study from Taiwan,” *International Journal of Data and Network Science*, vol. 4, no. 2, pp. 127–138, 2020, Available: http://m.growingscience.com/beta/ijds/3749-usability-evaluation-of-e-government-websites-a-case-study-from-taiwan.html

[27] M. Mubeen et al., “Usability Evaluation of Pandemic Health Care Mobile Applications,” *IOP Conference Series: Earth and Environmental Science*, vol. 704, no. 1, p. 012041, Mar. 2021, doi: https://doi.org/10.1088/1755-1315/704/1/012041

[28] H. Panduwiyasa, Y. Y. Febrian, M. Saputra, and Z. F. Azzahra, “Performance evaluation of ERP based to ISO/IEC 25010:2011 quality model (a case study),” *International Conference on Industrial, Enterprise, and System Engineering*, Jan. 2023, doi: https://doi.org/10.1063/5.0174711

[29] D. Yuniasri, P. Damayanti, and S. Rochimah, “Performance Efficiency Evaluation Frameworks Based on ISO 25010,” *IEEE Xplore*, Aug. 01, 2020. doi: https://doi.org/10.1109/EECCIS49483.2020.9263432. Available: https://ieeexplore.ieee.org/abstract/document/9263432

[30] C. Parrott, “Distributed Load Testing by Modeling and Simulating User Behavior,” *LSU Doctoral Dissertation*s, Dec. 2020, doi: https://doi.org/10.31390/gradschool_dissertations.5436. Available: https://repository.lsu.edu/gradschool_dissertations/5436?utm_source=repository.lsu.edu%2Fgradschool_dissertations%2F5436&utm_medium=PDF&utm_campaign=PDFCoverPages

[31] M. T. Hossain, R. Hassan, M. Amjad, and Md. A. Rahman, “Web Performance Analysis: An Empirical Analysis of E-Commerce Sites in Bangladesh,” *International Journal of Information Engineering and Electronic Business*, vol. 13, no. 4, pp. 47–54, Aug. 2021, doi: https://doi.org/10.5815/ijieeb.2021.04.04

[32] R. M. Abdulla, Hiwa Ali Faraj, K. O. Mohammed, and M. M. Younis, “Usability Evaluation of the Top 10 Universities in Iraq Using Heuristic Methods,” *Maǧallaẗ ʿulūm al-mustanṣiriyyaẗ*, vol. 34, no. 2, pp. 50–59, Jun. 2023, doi: https://doi.org/10.23851/mjs.v34i2.1234

[33] F. Ishengoma, “Exploring Critical Success Factors Towards Adoption of M-Government Services in Tanzania,” *Advances in web technologies and engineering book series*, pp. 225–253, Jan. 2022, doi: https://doi.org/10.4018/978-1-7998-7848-3.ch009

[34] A. C. Barus, E. S. Sinambela, I. Purba, J. Simatupang, M. Marpaung, and N. Pandjaitan, “Performance Testing and Optimization of DiTenun Website,” *Journal of Applied Science, Engineering, Technology, and Education*, vol. 4, no. 1, pp. 45–54, Jun. 2022, doi: https://doi.org/10.35877/454ri.asci841

[35] T. Heričko, B. Šumak, and S. Brdnik, “Towards Representative Web Performance Measurements with Google Lighthouse,” in *Proceedings of the 2021 7th Student Computer Science Research Conference (StuCoSReC)*, Sep. 2021. Accessed: Nov. 29, 2023. [Online]. Available: http://dx.doi.org/10.18690/978-961-286-516-0.9

[36] K. Chan-Jong-Chu *et al.*, “Investigating the Correlation between Performance Scores and Energy Consumption of Mobile Web Apps,” in *Proceedings of the Evaluation and Assessment in Software Engineering*, Apr. 2020. Accessed: Nov. 29, 2023. [Online]. Available: http://dx.doi.org/10.1145/3383219.3383239

[37] A. Mehroof and S. Rai, “Findability and Accessibility of Electronic Thesis and Dissertation Repositories of Newly Established Central Universities in India,” *ir.inflibnet.ac.in*, Nov. 2023, Accessed: Nov. 29, 2023. [Online]. Available: https://ir.inflibnet.ac.in/handle/1944/2425

[38] E. Fahlström and F. Persson, “Higher Education Diploma Software Engineering, Emphasis in Web Programming Sustainable Web Design How Much Can Environmental Friendly Design Principles Improve a website’s Carbon footprint?,” 2023. Available: https://www.diva-portal.org/smash/get/diva2:1775278/FULLTEXT02

[39] B. I. Belinda, A. B. Kayode, N. Solomon, and T. A. F. Bethy, “Analysis of Internal and External Website Usability Factors,” *Communications on Applied Electronics*, vol. 7, no. 35, pp. 9–18, Apr. 2021, doi: 10.5120/cae2021652880.

[40] T. W. Jun, L. Z. Xiang, N. A. Ismail, and W. G. R. Yi, “USABILITY EVALUATION OF SOCIAL MEDIA WEBSITES,” Jan. 2021. Accessed: Nov. 29, 2023. [Online]. Available: https://www.irjmets.com/uploadedfiles/paper/volume3/issue_1_january_2021/5580/1628083226.pdf

[41] F. Noureen *et al.*, “Agility Analysis of Academic Site: Identify Issues of Site by Users ,” *ResearchGate*, Apr. 2020. https://www.researchgate.net/profile/Irum-Hafeez-Sodhar/publication/341494465_Agility_Analysis_of_Academic_Site_Identify_Issues_of_Site_by_Users/links/5ec41e13458515626cb8204f/Agility-Analysis-of-Academic-Site-Identify-Issues-of-Site-by-Users.pdf

Jennifer L. Llovido is a faculty member of the Computer Science and Information Technology Department at Bicol University College of Science, Legazpi City, Philippines, with an academic rank of Associate Professor V. She completed her Doctor in Information Technology (DIT) at the University of the Cordilleras, Baguio City, Philippines. Her published research works are centered on the fields of natural language processing, data mining, and system design and development. She can be reached at jllovido@bicol-u.edu.ph.

Michael Angelo D. Brogada is an Associate Professor at the College of Science of Bicol University-Main Campus, Legazpi City. He is managing a software development company, MAB Business Solutions, which has developed software applications and maintained computer networks and servers for businesses since 2011. He finished his doctorate in Information Technology at the Technological Institute of the Philippines. He passed certifications in IT, such as IBM DB2 Academic Associate and DICT – EDP Specialist in Computer Programming. His research interests include IT Protection and Security, Data Mining, Web Applications, and Cloud Computing. He can be reached at madbrogada@bicol-u.edu.ph.

Lany L. Maceda earned her Doctorate in Information Technology from University of the Cordilleras, Baguio City, Philippines, in 2020. She is a faculty member of the Department of Computer Science and Information Technology, holding an academic rank of Associate Professor V at Bicol University. Moreover, she also serves as the Director of the Research, Development and Management Division at the same institution. She has been actively promoting data-driven policy-making through her research papers published in reputable international journals and conferences with research interests on machine learning particularly on natural language processing and data mining. She can be reached at llmaceda@bicol-u.edu.ph.

Mideth B. Abisado is an Associate Member of the National Research Council of the Philippines and a Board Member of the Computing Society of the Philippines Special Interest Group for Women in Computing. She is the Director of the CCIT Graduate Programs. She completed her Doctor in Information Technology (DIT) at the Technological Institute of the Philippines. Her research focuses on Emphatic Computing, Social Computing, Human-Computer Interaction, and Human Language Technology. She can be reached at mbabisado@national-u.edu.ph.

Research on the Transformation Path of DevOps in the Digital Era

Xiaoling Niu*, Lingling Yang*, Kailing Liu*, Zhaowei Liu*
*The China Academy of Information and Communications Technology, Beijing, China
niuxiaoling@caict.ac.cn, yanglingling1@caict.ac.cn, liukailing@caict.ac.cn, liuzhaowei@caict.ac.cn

Abstract— **DevOps has become one of the most important ways for enterprises to succeed. Enterprises use DevOps to integrate and optimize procedures, such as agile development management, continuous integration, continuous delivery, technical operation, application design, security and risk management (DevSecOps), system and tool construction, monitoring state, and quality assurance. This greatly accelerates the delivery of high-quality software products to users. Combined with agile, lean management, and continuous delivery, this approach has been reconstructing the way to effectively develop cloud service and try out hypothesis safely. Firstly, the flow through the lifecycle, which is divided into plan, design, code, build, test, release, deploy, and operate, should be identified and accelerated by automated test and CI/CD. Then, measurement should be integrated into each part of the flow, and the monitoring system should provide instant feedback. Quality assurance covers the entire lifecycle of development and operation, which is beneficial for identifying defects at an earlier stage and preventing downstream problems.**

Keywords— **DevOps, agile development management, Continuous delivery, Security and Risk Management**

I. INTRODUCTION

With the advent of the digital era, enterprises have gradually discovered the great potential of digitalization. Therefore, smoothing the road of digital transformation has become the focus of enterprise attention.

At present, enterprises have gradually discovered the great potential of digitalization, so how to smooth the road of digital transformation has become the focus of enterprise attention.

The driving factors of enterprise digital transformation are mainly external and internal factors. The external factors are mainly user-driven, technology-driven and competition-driven. User-driven factors mainly reflect users' higher expectations for product capabilities and service efficiency. In recent years, the mature development of new technologies such as cloud computing, big data, blockchain and artificial intelligence has laid a good technical foundation for digital transformation. Competitive drive mainly stems from the competition for products and services within the same industry. The internal factors are mainly a series of internal needs such as strategic planning, business innovation, cost reduction and efficiency improvement, safety and compliance. Strategic planning is reflected in the matching degree and consistency of enterprise strategic transformation; Business innovation is mainly manifested in the ability of the project team at the organizational level in a unified and service-oriented way to achieve business agility and innovation. Cost reduction and efficiency enhancement mainly show that enterprises need to realize the allocation and application of resource value maximization under the condition of market saturation and increasingly fierce competition. Safety compliance is mainly to standardize the safety and compliance risk system of enterprises and escort the stable and long-term development of enterprises. Driven by both internal and external factors, enterprise transformation has become an inevitable trend.

In the digital era, where “software defines the future”, software has become a key factor in driving industrial upgrading. All things are connected by intelligence, and software enables thousands of activities and industries, which puts forward higher requirements for flexibility, collaboration and precision of software services. The IT department of an enterprise, as the executor of digital transformation, faces increasingly high expectations for digital effectiveness in both the market and enterprise management. This trend, signified by the growing demand for digital solutions, has brought more workload and pressure to IT departments. Consequently, the assessment requirements for IT departments are becoming more aligned with those of business departments.

This paper suggests integrating flexible and changeable business needs with technological innovation. The introduction of DevOps will break the functional silos between traditional Development, testing, and operations department. It will optimize the traditional software development lifecycle management process and shorten the product delivery cycle, adapting to the rapid development of customer needs and market changes.

As shown in Fig.1 , the DevOps is commonly regarded as the intersection among the scopes of software Development (Dev), Operation (Ops) and Quality Assurance (QA). DevOps is the culture, process, and technology to foster close collaboration of software development and IT operations[1-2]. It involves processes, methods, and systems that collectively, through research, development, testing, operation, and maintenance, fill the information gap between various departments. This approach reshapes software development, technology operations, and safety compliance processes in terms of organizational culture. It builds automated toolchains to enhance team collaboration and enables more convenient, frequent, and reliable software releases through automated, agile, and integrated delivery processes, as shown in Figure 1.

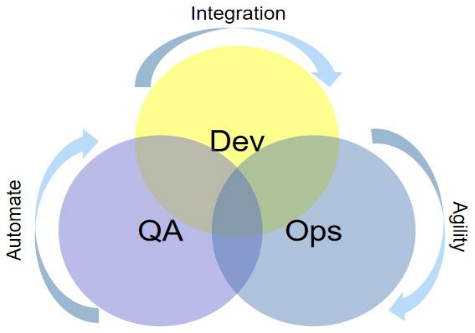


Figure 1. DevOps concept

This paper analyzes the path of enterprise DevOps transformation and industry best practices in the context of digitalization, and provides suggestions on the future trends of enterprise DevOps transformation.

II. Key elements of DevOps transformation in the digital era

The key elements of the enterprise's DevOps transformation cover the full lifecycle of end-to-end software development. The specific implementation is divided into six parts: agile development management, continuous delivery, technology operation, application design, security and risk management, system and tool construction, as follows.

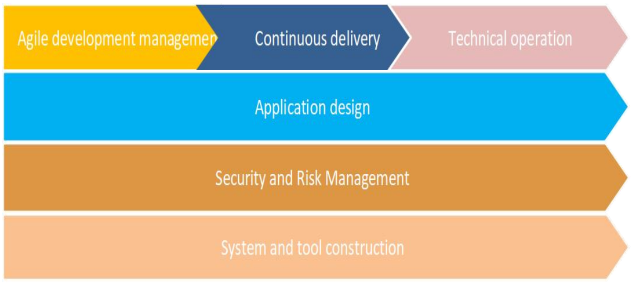

Agile development managemer
Continuous delivery
Technical operation
Application design
Security and Risk Management
System and tool construction

Figure 2. DevOps key elements

1) Agile development management enables enterprises to deliver value through business products and fosters close collaboration among development, testing, and other roles. It adapts to change by employing an evolutionary approach to planning, development, and continuous delivery, allowing for immediate feedback, adjustments, and continuous improvement in a rapidly changing market environment. At present, the main team agile practices adopted by enterprises include Scrum, lean software development, Kanban, extreme programming, and so on[3].
2) Continuous delivery refers to an enterprise's ability to deliver various changes—such as new functions, defect repairs, and configuration changes—safely, quickly, and with high quality to the production environment or users. Companies usually DevOps process can be divided into planning, design, coding, build, test, deployment, distribution and operating stage, using the research, and a half since the research or commercial tool system or platform, implement DevOps process of development and management, continuous integration, testing, continuous deployment, continuing operations, measure feedback closed-loop management[1],[5],[7].
3) Technical operation is a process that builds enterprise technology capacity. It includes monitoring, incident and change management, capacity and cost management, and management of high availability and business continuity, all centered around the business. This ensures stable, safe, and efficient delivery of technical services, aiming to build industry-leading technical operation capabilities. Support the continuous development and strategic success of the enterprise. Different from traditional technology operation and maintenance, technology operation should not only focus on stability, security and reliability, but also focus on user experience, efficiency and benefit.
4) Application design is a crucial horizontal support domain capability for enterprises. By decoupling application and architecture, and considering factors like scalability, testability, and observability, the design architecture can more effectively support the realization of process domain goals, such as agile management, continuous delivery, and technical operation.
5) In the DevOps unified development mode, Security and Risk Management (DevSecOps) integrates security into every aspect of work, rather than it being solely the responsibility of the security team. Embedding security throughout the application development lifecycle, while maintaining controllable security risks, enhances enterprise IT efficiency and achieves the integration of research, development, and operation[4]. We should control the overall risk, the risk of the development process, the risk of the delivery process and the risk of the operation process from four main aspects, by embedding security capabilities into daily norms and tools, with safety training and education means, to reach an internal consensus on security[6].
6) System and tool construction are essential for DevOps implementation. Enterprises need to establish a toolchain that connects the entire end-to-end software delivery lifecycle, encompassing project and development management, application design, continuous delivery, test management, automated testing, and technical operation. The market offers mature open-source or commercial systems and tool platforms, enabling enterprises to choose or develop secondary tools to create a DevOps tool chain tailored to their specific needs[8].

III.Industry Practice Status

Currently, various industries are adopting DevOps transformation. This includes the communication, Internet, banking, securities, insurance, travel, and tourism industries, among others. Enterprises have comprehensively implemented DevOps practice pilot in the organization, and through the whole life cycle of software research and development to improve the overall efficiency and quality. The Table 1 shows the result of conventional method in terms of factor for different industry and Table 2 shows the results of DevOps method for different industry. Compared the results in Table 2 to the Table 1, the advantage of DevOps is obvious.

TABLE 1. RESULT OF CONVENTIONAL METHOD IN TERMS OF FACTOR FOR DIFFERENT INDUSTRY

Industry factors	communications	Internet	Banking	securities	Insurance
Requirement lead time	40 days	14 days	50 days	25 days	35 days
Build time	1 hour	10 min	25 min	25 min	1 hour
Continuous integration response time	6 hours	30 min	5 hours	4 hours	1 time/day
Commissioning frequency	1 time/month	1-2 times/weeks	1-2 times/month	2 times/month	1-2 times/month

TABLE 2. RESULT OF DEVOPS METHOD IN TERMS OF FACTOR FOR DIFFERENT INDUSTRY

Industry factors	communications	Internet	Banking	securities	Insurance
Requirement lead time	20 days	7 days	8 days	10 days	7 days
Build time	30 min	3 min	5 min	3 min	20 min
Continuous integration response time	3 hours	20 min	5 min	5 min	Immediate integration
Commissioning frequency	1 time/month	1-2 times/weeks	4-8 times/month	1-2 times/weeks	3-4 times/month

1) In the communications industry during the 5G era, the combination of cloud-native technology and DevOps will drive various network functions, including control, management, operation, and maintenance in the telecom network. The commercialization of 5G brings significant changes in services and technologies, such as high transmission rates, low latency, wide connectivity, and virtual software services. These advancements mean not only faster speeds but, more importantly, the ability to adapt to diverse service scenarios and quickly respond to market changes.
2) The Internet industry is characterized by rapid business decisions, frequent demand changes, and frequent deployment. Although the software architecture is large and complex, making system maintenance difficult and increasing deployment and operation costs, new technologies are being introduced at a rapid pace. The Internet industry takes the lead in the implementation of DevOps, expanding from team-level pilot DevOps to company-level DevOps implementation practice, creating a one-stop DevOps research and development platform, which runs through the full life cycle of product development, testing and operation.
3) The Banking industry, the financial industry is actively adopting open source software and Internet architecture, gradually transitioning from traditional IT models to fintech. The financial industry has always been in the first echelon of IT technology application development. According to the evaluation data and statistics from the Information and Communication Institute, a number of large banks have successfully implemented DevOps.
4) The securities industry, the role of fintech in the securities industry is gradually emerging. Under the background of new technologies and new forms of business, the use of fintech to enable business development, and the combination of fintech with business income, cost reduction and efficiency increase, has become the common concern of securities industry practitioners. At present, the securities industry is also actively building an internal energy efficiency management platform.
5) The Insurance industry, Information technology drives the IT scale of the financial and insurance industry to increase significantly, and internal changes are taking place in the IT capability and platform structure of enterprises. Secondly, business scenarios are in urgent need of agile IT operation to cope with the flexible and changeable market demand. The insurance industry can achieve cost reduction and efficiency increase through the construction of efficient platform.

IV. FUTURE DEVELOPMENT DIRECTION

DevOps promotes the transformation and upgrading of enterprise IT operation mode, as well as the improvement and optimization of organizational management process. The construction goal of the enterprise has also been continuously expanded from a single pilot project to the ability improvement at the organizational level. The development trend mainly includes the following aspects.

The enterprise is advocating for transformational governance in application architecture. Conway’s law suggests that an organization’s structure mirrors its system design, meaning that the design produced by a firm reflects its internal communication structure. With the transformation of enterprise architecture from single architecture to cloud and container architecture, flexible and decoupled application architecture has become the first choice of enterprises. Such application architecture can achieve smooth collaboration among teams and efficient management of IT teams. The transformation and governance of enterprise application architecture is an urgent task.

The enterprise attaches great importance to the development of an integrated security operation system. With the frequent occurrence of all kinds of security incidents, the loss and negative impact to the enterprise is incalculable, and it has become a huge obstacle on the road to the rapid and healthy development of the enterprise. At the same time, according to National Security Law, and other regulations, clear provisions and requirements are established for information security and control during enterprise digital development. Therefore, enterprises should focus on building DevSecOps capabilities and continuously enhance their security assurance systems. In order to protect the long-term development of enterprise digital transformation.

The multi-scenario application of Intelligent Operation and Maintenance (AIOps) improves enterprise quality and efficiency. According to the analysis of AIOps Status Survey Report, at present, enterprises are facing problems such as expansion of business scale, complex system and explosion of users, which greatly challenge the traditional operation and maintenance mode and increase the difficulty of operation and maintenance. AI, big data and other new-generation innovative technologies can greatly help operation and maintenance personnel improve the quality and efficiency of operation and maintenance work, freeing them to achieve greater breakthroughs and technological innovations.

Measurement of development and operation effectiveness becomes a powerful tool for enterprises to measure the effectiveness of DevOps applications. In recent years, measuring development and operation effectiveness has garnered high-level attention within enterprises. Leading Internet companies like Baidu, Tencent, Alibaba, and JD.com have been exploring engineering practices to enhance development efficiency. This focus on measuring development and operation efficiency aids top management in decision-making and provides insights for transforming research and development processes.

V. Conclusions

Global enterprises have embraced DevOps as a way to improve business processes. As a new governance paradigm in the field of DevOps， not only refers to a technology or solution, but also emphasizes the communication and cooperation between development, testing, operation and maintenance departments and cultural penetration, so as to improve the efficiency, quality and team culture of software delivery. In the future, more enterprises will practice DevOps, and more enterprises will move forward from DevOps comprehensive level to excellent level. It is also expected that DevOps can achieve integrated innovation and development in the future, and continue to promote the high-quality development of software industry.

Acknowledgment

Corresponding author: Lingling Yang.

References

[1] Prashant Agrawal, Neelam Rawat, "Devops, A New Approach To Cloud Development & Testing ",2019 2nd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT),pp.982-985.

[2] Pietrantuono, R., Bertolino, A., De Angelis, G., Miranda, B. and Russo, S.,"Towards continuous software reliability testing in DevOps". In Proceedings of the 14th International Workshop on Automation of Software Test,2019, May, (pp. 21-27). IEEE Press.

[3] M. Younas, D.N. Jawawi, I. Ghani, T. Fries and R. Kazmi, "Agile development in the cloud computing environment: A systematic review", Information and Software Technology, vol. 103, pp. 142-158, 2018.

[4] M. Soni, "End to end automation on cloud with build pipeline: the case for DevOps in insurance industry continuous integration continuous testing and continuous delivery", 2015 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), pp. 85-89, 2015, November.

[5] Ebert, C.; Gallardo, G.; Hernantes, J.; Serrano, N. "DevOps". IEEE Softw. 2016, 33,pp. 94 – 100.

[6] A. A. U. Rahman and L. Williams, "Software security in devops: Synthesizing practitioners ’ perceptions and practices", in 2016 IEEE/ACM International Workshop on Continuous Software Evolution and Delivery (CSED), May 2016, pp. 70 – 76.

[7] ,J. Humble and D. Farley, "Continuous Delivery: Reliable Software Releases through Build, Test, and Deployment Automation", 1 edition. Addison-Wesley Professional, 2010.

[8] Mik Kersten, "A Cambrian Explosion of DevOps Tools ",IEEE Software, March 2018,pp14-17.

Xiaoling Niu is the leader of DevOps Standards Working Group and the editor of DevOps International Standards. He has been engaged in research work related to development operation and maintenance for a long time, including operation and maintenance management system review and other related work. Participated in the compilation of "Cloud computing Service Agreement Reference Framework", "Object storage", "Cloud database", "Development and operation management (DevOps) capability Maturity Model" series standards, "cloud computing operation and maintenance intelligent general evaluation Method" and other more than 20 standards. Participated in the compilation of several white papers and survey reports, including the White Paper on Enterprise IT Operation and Maintenance Development and the Survey Report on the Status of DevOps in China (2019).

LingLing Yang is currently the Director of the Governance and Audit Department of Yunda Institute of China Academy of Information and Communications Technology, and the Secretary-General of the IT Risk Management Working Committee of the Internet Society of China. She is mainly engaged in the research work of digital governance, focusing on the core fields of XOps system construction, information technology risk governance system, and IT value insight. The initiator of the IT New Governance Leadership Summit (GOLF+), led the establishment of the personal information protection compliance audit pilot program, presided over the completion of more than 10 standards in the field of science and technology governance. Prior to joining the firm, he was a partner in Information Technology and Digital Services at Deloitte Touche Tohmatsu and has more than 13 years of IT governance related consulting and auditing experience.

Kailing Liu, engaged in DevOps, DevSecOps and other IT development research and industry standard formulation work for five years. She is one of the core members of the DevOps industry standard and research work group.She has participated in the preparation of more than six industry standards and five research reports.

Zhaowei Liu is graduated from Nanyang Technological University in Singapore with a master's degree. He has been engaged in related research work in development and operation for a long time. He participated in the preparation of the "R&D Operations Integration (DevOps) Capability Maturity Model" series of standards and the "Business R&D Operations Integration (BizDevOps) Model" and other standards. Participated in the preparation of multiple white papers and survey reports, including the "China DevOps Status Survey Report (2021, 2022, 2023)", etc., participated in the evaluation of more than 40 DevOps capability maturity assessment projects, and has rich experience in standard preparation and evaluation testing.

Session 4A: 6G, Mobile Communication 2

Chair: Prof. Juinn-Horng Deng, Yuan Ze University, Taiwan

1 Paper ID: 20240470, 252~262

An Efficient Resource Allocation Algorithm for Traffic of a Content Streaming in Non-standalone OFDM-Based 5G NR

Ms. Narantuya VANDANTSEREN ,

Mongolian Defense University. Mongolia

2 Paper ID: 20240387, 263~270

Design of Communication Countermeasure Simulation Model and Data Interaction Interface for Battlefield Network Based on QualNet

Mr. Wenyi Li, Prof. Peng Gong, Mr. Weidong Wang, Mr. Yu Liu, Mr. Jianfeng Li, Dr. Xiang Gao,

Beijing Institute of Technology. China

3 Paper ID:20240480, 271~274

Location based Data-centric Forwarding for Mobile Ad-hoc Networks

Mr. Hieu Nguyen, Prof. Ilkyeun Ra,

University of Colorado Denver. USA

4 Paper ID: 20240369, 275~280

Optimization of Downlink Power Allocation in NOMA-OTFS based Cross-Domain Vehicular Networks

Mr. Hao Xu, Mr. Zhiquan Bai, Ms. Jinqiu Zhao, Mr. Dejie Ma, Mr. Bangwei He, Mr. KyungSup Kwak,

Shandong University. China

5 Paper ID: 20240040, 281~285

Performance Evaluation of UAV-based NOMA for 5G and Beyond

Ms. MOUNIKA NEELAM, Prof. Anuradha Sundru,

NITW. India

An Efficient Resource Allocation Algorithm for Traffic of Content Streaming in Non-Standalone OFDM Based 5G NR

Narantuya Vandantseren[*], Chuluunbandi Naimannaran[**], Tuyatsetseg Badarch[**], Erdenebayar Lhamjav**, Otgonbayar Bataa[**]

[*]Mongolian National Defense University

[**]School of Information and Telecommunication Technology of MUST, Ulaanbaatar, Mongolia

vnarantuya0123@gmail.com, chuluunbandi@must.edu.mn, tuyatsetseg.b@must.edu.mn, otgonbayar_b@must.edu.mn, erdenebayar.l@must.edu.mn

Abstract— **The 5G NR network provides support for a wide range of service types, including content traffic through large number of wireless connections. To ensure efficient resource allocation for specific service types, such as video streaming, a Traffic Differentiator stage is implemented at the user level. This stage segregates content queues from active users into distinct service queues, enabling tailored resource allocation. In order to investigate the performance of differentiated resource allocation in a content delivery scenario, this research paper focuses on the physical and cross-layer resource allocation of the non-standalone 5G NR network. The paper introduces a Pseudo-Inverse-based Traffic Differentiator algorithm in the TD Scheduler stage, which aims to allocate optimal radio resources to real-time and non-real-time services, while allocating the remaining resources to background services. By segregating users based on their requested services and prioritizing them differently within the service priority-specific queues, the Traffic Differentiator algorithm enhances throughput fairness among all users using Pseudo-Inverse learning. Additionally, the FD Scheduler stage utilizes our proposed optimal Channel Quality Indicator (CQI) selection algorithm to leverage Frequency Domain (FD) Multi-User (MU) diversity for resource allocation.**

Keywords— **Non-standalone 5G NR, Resource Allocation, Physical and Cross layer, Traffic, Content Delivering.**

I. INTRODUCTION

The Third Generation Partnership Project (3GPP) has set the initial specifications for 5G New Radio (NR), which are currently considered the most widely recognized standard for 5G cellular technology. As per 3GPP, the term "5G" encompasses any system that incorporates 5G NR software, a definition that gained significant acceptance in late 2018.

The Non-Standalone (NSA) standards play a significant role in the widespread use of 5G NR. In this approach, the control plane is managed by the existing 4G LTE network, while the user plane incorporates 5G NR. The implementation of this particular standard by 3GPP aligns with the industry's goal of expediting the deployment of faster 5G services, while making use of the already established infrastructure of the existing 4G LTE network.

The latest generation of mobile networks, known as 5G New Radio (NR), has been standardized by building upon the existing concepts and features of 4G/4G+. The objective of this advancement is to bring in a greater level of adaptability, expandability, and effectiveness to the network.

Time division duplex (TDD) plays a crucial role in 5G NR by effectively managing the imbalance between uplink (UL) and downlink (DL) traffic. It achieves this by utilizing real-time traffic estimation and providing greater flexibility in allocating network resources. Just like LTE, the physical layer of 5G NR consists of essential physical channels and signals that are vital for the smooth functioning of the network [1]. This technology specifically caters to use cases such as enhanced Mobile Broadband (eMBB) and Ultra-Reliable and Low Latency Communication (URLLC), emphasizing its commitment to delivering superior broadband experiences and ensuring highly reliable, low-latency communication services.

Unlike LTE [12], NR dual connectivity involves nodes from two different Radio Access Technologies (RATs): the gNB and the eNB. The gNB provides NR access while the eNB provides E-UTRA/NR access. This creates a strong interworking between the two radio technologies and allows for a gradual integration of NR into existing LTE networks. The NR access network can operate in two modes: non-standalone and standalone. Non-standalone operation enables NR to connect within existing LTE networks, which helps expedite the deployment of 5G. On the other hand, standalone operation requires NR to connect to the

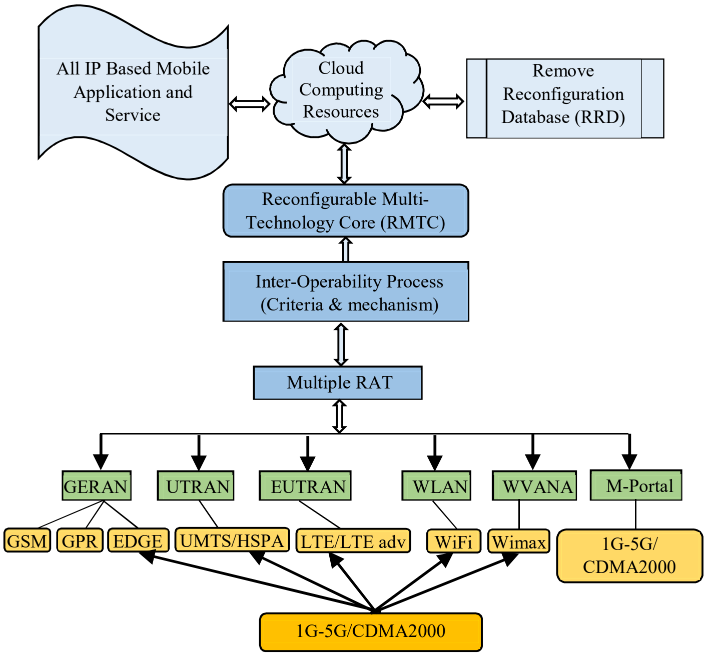

All IP Based Mobile Application and Service
Cloud Computing Resources
Remove Reconfiguration Database (RRD)
Reconfigurable Multi-Technology Core (RMTC)
Inter-Operability Process (Criteria & mechanism)
Multiple RAT
GERAN
UTRAN
EUTRAN
WLAN
WVANA
M-Portal
GSM
GPR
EDGE
UMTS/HSPA
LTE/LTE adv
WiFi
Wimax
1G-5G/ CDMA2000
1G-5G/CDMA2000

Figure 1. 5G NR system connectivity architecture

5G Core (5GC) in addition to LTE's connection to the 5GC. More detailed information about these modes can be found in the subsequent subsection [2]. The focus of this article is on the Efficient Resource Allocation Algorithm created for Non-standalone 5G NR, with a specific emphasis on the physical downlink and uplink control channels and signals.

The system connectivity architecture is depicted in Figure 1, showcasing the extensive connectivity of 5G NR, which aims to enable communication not only between humans but also between machines. It is expected that current technologies like OFDMA will continue to be effective for the next 50 years, and there is no pressing need for modifications, as stated in [3, 4].

Figure 2 depicts the Service-based architecture of the 5G NR non-standalone system, while Table 1 describes its corresponding functions. As the 4G technology features improved as an advance version of 3G in terms of speed, data bandwidth & improvement of used technology. As research work in mobile communication is focusing on 5G technology and researches are progressing towards the World Wide Wireless Web (WWWW), Dynamic Adhoc Wireless Networks (DAWN) & totally real wireless world. We are expecting that 5G will be introduced in communication by 2020 which is basically user oriented. In that user can avail specific feature of 5G such as very high speed & massive data bandwidth at the low cost per bit [3]. With the help of these advanced technologies, LTE networks can now cater to a diverse range of applications and fulfill the data rate requirements set by 3GPP, which is 100 Mbps for DL and 50 Mbps for UL, along with 1.5bps/Hz spectrum efficiency in a 20 MHz bandwidth [4].

TABLE 1. 5G SYSTEM SERVICE-BASED ARCHITECTURE WITH CORE NETWORK FUNCTIONS.

		Main functions
NSSF	Network Slice Selection Function	Selects the Network Slice Instance (NSI) based on information provided during UE attach.
NEF	Network Exposure Function	Facilitates, robust, developer-friendly access the exposed network services.
NRF	Network Repository Function	Facilitates, robust, developer-friendly access the exposed network services.
UDM	Unified Data Management	Authentication Credential Repository, Access Authorization
AUSF	Authentication Server Function	Authentication and Authorization
PCF	Policy Control Function	Ensures policy charging control, authorized QoS.
AMF	Access and Mobility Management Function	NAS Signaling Termination Mobility Management Network Slicing.
SMF	Session Management Function	Selection and control of UP function, UE IP address allocation and management.
UPF	User Plane Function	Packet routing and forwarding QoS handling.
SMF	Session Management Function	Responsible for interacting with the decoupled data plane, creating updating and removing PDU sessions and managing session context with the UPF

The end-to-end network architecture of 5G, as shown in Figure 2, consists of various components. These include the Next Generation RAN (NG-RAN), also known as Cloud RAN, which is responsible for providing wireless connectivity. Additionally, there is Multi-Access Edge Computing (MEC) within the Virtual Evolved Packet Core (vEPC), which enables efficient processing at the network edge.

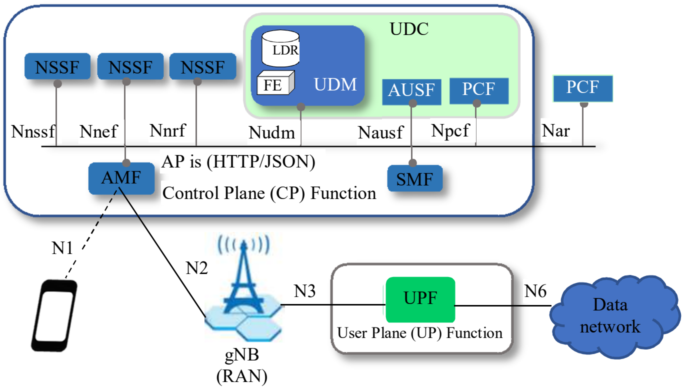


Figure 2. 5G system service-based architecture

To implement the architecture of 5G NR, foundational elements such as NS, NFV, NFV Management and Orchestration (MANO), and SDN are utilized. SDN allows for a logically centralized control and enables the programming of network devices.

The arrangement of the paper is as follows: Section II presents the physical and cross layer resource allocation design of Non standalone 5G NR system. Section III reviews the research related analysis based on the literature of the resource allocation design for the system. Section IV focuses on the proposed Pseudo Inverse scheme based Traffic Differentiator algorithm, its time and frequency scheduler sub-algorithms for content delivering of 5G NR service-oriented deployment. Finally, the research is concluded and future work is discussed.

II. PHYSICAL AND CROSS LAYER RESOURCE ALLOCATION DESIGN FOR NON STANDALONE 5G

A. 5G NR frame structure

Resource Allocation plays a crucial role in 5G NR wireless networks as it serves as the primary foundation. The efficient and equitable implementation of RA schemes heavily relies on the energy consumption and allocation of spectrum. Given the utilization of a new spectrum, the emphasis on spectrum sharing becomes paramount in the context of RA within 5G networks.

The 5G NR technology is capable of supporting both frequency division duplexing (FDD) and time division duplexing (TDD). A 5G NR frame structure is designed for efficient support of users with highly diverse service requirements. While the 5G NR frame structure shares similarities with LTE, it brings about significant improvements and important modifications. Just like LTE, each frame lasts for 10 ms and consists of 10 subframes per frame, with each subframe having a duration of 1 ms. Within a slot, there are 14 OFDM symbols, as shown in Figure 3, which highlights the structure of the 5G NR frame.

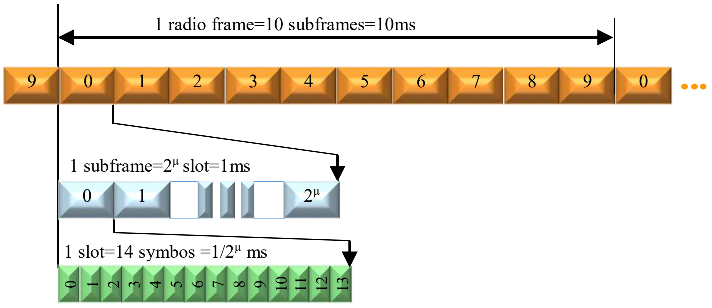


Figure 3. 5G NR Frame structure

In contrast to LTE, 5G NR introduces a significant change in the number of slots per subframe (referred to as 2μ), which determines the duration of each slot and depends on the subcarrier spacing (SCS). Unlike LTE, where the number of slots per subframe was always two, 5G NR allows for variability in this number. More specifically, the number of slots per frame is determined by the subcarrier spacing and can range from 15 to 240 kHz. Table 2 provides a comprehensive overview of the various choices available, along with the corresponding carrier frequencies they are specifically tailored for. Just like LTE, a Resource Element (RE) in this context refers to a single subcarrier within an OFDM symbol. However, unlike LTE, a Resource Block (RB) consists of 12 subcarriers within a single OFDM symbol (Figure 3). Moving on to Table 2, it presents the number of RBs based on the system bandwidth and subcarrier spacing, specifically for frequencies below 6 GHz.

The 5G NR technology employs a versatile subcarrier spacing (SCS) that is derived from the standard 15 KHz used in LTE. This SCS can be adjusted to values of 30, 60, and 120 KHz. In the case of a 15 KHz SCS, each subframe consists of a single slot with a duration of 1 ms. However, for a 30 KHz SCS, each subframe is divided into two slots, each lasting 500 µs [6-8].

TABLE 2. NUMBER OF SLOTS PER SUBFRAME, SLOT DURATION, NUMBER OF SLOTS IN A FRAME AND GUARD PERIOD FOR REFERENCE SCS.

SCS	µ	Number of slots per subframe	Slots duration	Number of slots in a frame	Guard Period
15 Khz	0	1	1 ms	10	Normal
20 Khz	1	2	500 µs	20	Normal
30 Khz	2	4	250 µs	40	Normal/ Extended
60 Khz	3	8	120 µs	80	Normal

Each slot consists of either 14 OFDM symbols or 12 OFDM symbols, depending on whether it follows the normal Guard Period (GP) or the extended GP. Nevertheless, mini slots (2, 4, or 7 symbols) can be assigned for shorter transmissions. Additionally, slots can be combined for longer transmissions.

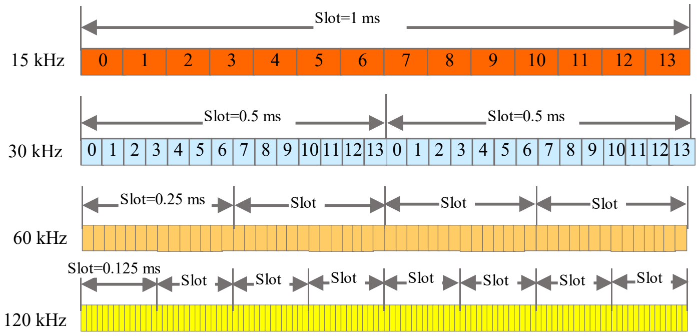


Figure 4. 5G NR scalable slot duration

The 5G NR frame structure is designed to support both Frequency Division Duplex (FDD) for paired spectrum bands and Time Division Duplex (TDD) for unpaired spectrum bands. Unlike LTE, where a subframe is made up of two slots containing 7 OFDM symbols, in 5G NR, a subframe is created by combining slots. Each slot is composed of either 7 or 14 OFDM symbols for

Sub-Carrier Spacing (SCS) of 60 kHz, and 2, 4, 7, or 14 OFDM symbols for SCS of 120 kHz. The 3GPP has standardized 255 symbol combinations, each associated with a slot format identified by a slot format index [9]. These OFDM symbols can be assigned for downlink, uplink, or flexible usage. A downlink symbol is exclusively used for downlink transmission without any simultaneous uplink transmission. Conversely, an uplink symbol is solely utilized for uplink transmission,
avoiding any overlapping transmission in the downlink. Flexible symbols, however, can be adapted for transmissions in either the downlink or uplink direction.

B. 5G NR - OFDM based UL-DL pattern

3GPP has adopted CP-OFDM with a scalable numerology, incorporating subcarrier spacing and cyclic prefix, for both uplink (UL) and downlink (DL) operations, expanding this technology up to at least 52.6 GHz. Building upon LTE, 5G NR enhances the concept of carrier aggregation by introducing the Supplementary Uplink (SUL). Unlike conventional carrier aggregation, where each uplink carrier corresponds to a specific downlink carrier, SUL associates a standard downlink/uplink carrier with an additional supplementary uplink carrier operating at lower frequencies. SUL aims to extend uplink coverage and enhance uplink data rates by capitalizing on reduced path loss in low-frequency bands, particularly in scenarios with limited power availability [10].

The organization of timeslots for uplink and downlink transmission is done through DL-UL patterns. In LTE TDD, there are 7 pre-established patterns for the allocation of UL and DL in a radio frame. While there is no predefined pattern for 5G NR, a flexible pattern can be defined using parameters in TDD UL/DL Common Configuration (tdd-UL-DL-configurationCommon), as illustrated [10].

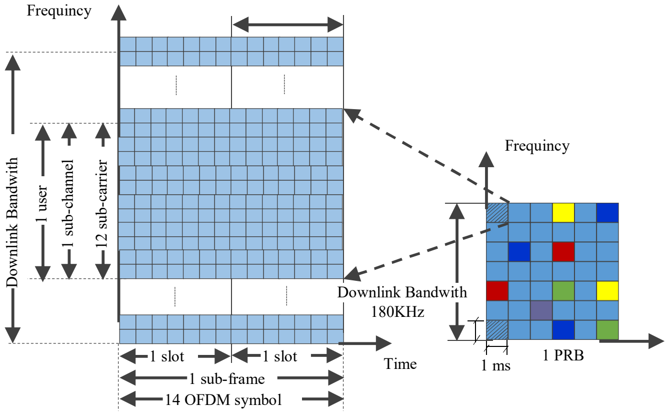


Figure 5. The OFDMA technology in 5G NR

During the air interface phase of 5G NR, mobile connectivity is made possible through the implementation of Orthogonal Frequency Division Multiple Access (OFDMA). OFDMA utilizes Orthogonal Frequency Division Multiplexing (OFDM) for each user and orthogonal subcarriers to guarantee minimal interference among users within a cell or sector. This feature is a significant advantage over previous generations such as 4G LTE [3, 11].

In the realm of 4G LTE, a Physical Resource Block (PRB) is composed of 12 subcarriers and encompasses 7 OFDM symbols within a slot duration of 0.5 ms. Conversely, in the context of 5G NR, a Resource Block (RB) is constituted by 12 subcarriers. Notably, the allocation of resources in 5G diverges from its predecessor, as it involves assigning one or more Contiguous Resource Blocks (CRBs) to a user for a designated count of OFDM symbols.

An example of this would be the transmission of various data types, including speech, text, email, and video, to four users within a cell. To prevent any interference among them, each user is assigned specific CRBs and OFDM symbol ranges. This allocation system is dynamic and ensures that resources are utilized in an optimized manner.

5G NR utilizes two types of OFDM: cyclic prefix OFDM (CP-OFDM) and Discrete Fourier Transform Spread OFDM (DFT-S OFDM). CP-OFDM, similar to the access technology employed in LTE, incorporates a variable subcarrier spacing known as numerology, which facilitates the support of various subcarrier separations. Conversely, DFT-S OFDM, also referred to as SC-OFDM (Single Carrier OFDM), combines the transmission characteristics of a single carrier with OFDM, thereby presenting a transmission scheme akin to OFDMA. This scheme assigns distinct Fourier coefficients to different transmitters based on a complex constellation of symbols derived from the transmitted bits for each user.

The access technology employed in the second version of 5G NR is CP-OFDM (Cyclic Prefix OFDM). Unlike LTE's fixed 15kHz subcarrier spacing, CP-OFDM offers variable subcarrier spacings such as 15kHz, 30kHz, 60kHz, 120kHz, and more. This variability in subcarrier spacing results in changes in the duration of the CP symbol. The cyclic prefix, which is an integral part of OFDM schemes like OFDM, acts as a protective buffer between consecutive symbols, effectively reducing inter-symbol interference [11]. Moreover, 5G takes advantage of OFDM's combination of Quadrature Amplitude Modulation (QAM) and Frequency Division Multiplexing (FDM), enabling higher data rates for wireless networks, even in densely populated wireless network environments [13-15]. In general, a resource allocation of traffic in network area is, therefore, a challenging research area [16]. In LTE system, the algorithm of the received signal with pilot to interference cancellation algorithm for 2x2 MIMO in LTE was proposed to identify pilot-based channel estimation algorithm with low complexity [17]. The 5G NR technology supports both Frequency Division Duplex (FDD) and Time Division Duplex (TDD) schemes, which results in the definition of NR Resource Allocation Types for both downlink and uplink specifications. These types determine how resources are allocated in the frequency domain. More specifically, the allocation methods within the downlink frequency domain are elaborated with a focus on resource allocation techniques within this domain [3].

C. Cross layer design for Traffic Differentiator

The resource allocation approach in NR closely resembles LTE scheduling, but NR demonstrates a higher level of precision, especially in its time domain scheduling within the physical layer. The suggested architecture utilizes a cross-layer concept, incorporating information from different layers to guide scheduling decisions, as illustrated in Figure 6. This approach divides the scheduling process into three stages: the Traffic Differentiator, the TD Scheduler, and the FD Scheduler, which handle the scheduling of Time Domain (TD) and Frequency Domain (FD) operations, respectively.

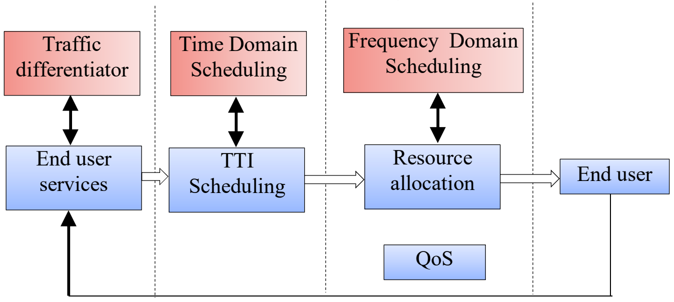


Figure 6. A cross-layer packet scheduling architecture

The Traffic Differentiator module collects information on traffic type from the application layer, Quality of Service (QoS) criteria from the network layer, queue status updates from the Radio Link Control (RLC) layer, and channel status specifics from the physical layer for each user. The TD Scheduler utilizes QoS measurements, such as average Packet Error Rate (PER) from the MAC layer, as part of the input dataset for the Pseudo-Inverse learning process. On the other hand, the FD Scheduler allocates resources across the frequency domain by taking into account user priorities and per-PRB Channel Quality Indicator (CQI) reports for efficient Physical Resource Block (PRB) mapping.

III. ANALYSIS ON RELATED WORKS

A. Analysis of Resource Allocation Algorithms for 5G NR Resource allocation

This section explores the concept of centralization in 5G NR mobile networks, with the goal of achieving adaptable processing and management that can effectively meet the demands of various services. The objective is to strike a balance between the partially decentralized networks of today and a fully centralized cloud radio access network. To address this middle ground, the concept of Radio Access Network as a Service is introduced, which involves partially centralizing radio access network functionalities based on the specific needs of the network [18]. The discussion then may fall into various scheduling schemes that form the foundation of networks [19]. These include Round Robin, BestCQI, Proportional Fair, MLWDF, and EXP-PF, as well as newer schemes developed based on specific metrics outlined in Table 4. Round Robin, for example, is a simple method that evenly distributes resources among all users in the current cell [21], [22]. Its simplicity has made it widely used [20]. On the other hand, the Best CQI scheme prioritizes users with the best Channel Quality Indicator (CQI) [23]. Proportional Fair (PF) divides achieved data rates by past average rates, ensuring fairness by giving priority to users with lower previous rates [24], [25]. Variations of MLWDF, such as EXP-MLWDF, combine MLWDF with an exponential term to favor users with poor channel conditions [26, 27].

In addition to these schemes, there have been other developments in the field. For instance, Charles Katila's Neighbors Aware Proportional Fair (N-PF) method takes into account the presence of IoT devices adjacent to each scheduled user in heterogeneous systems [28]. Furthermore, there is a scheduler specifically designed for eMBB downlink, aiming to optimize resources and prioritize cell-edge users [29, 30]. Another proposed scheduler focuses on prioritizing URLLC flows based on the signal-to-noise ratio multiplied by a normalized queue state parameter, which helps prevent buffer congestion [31].

In the context of 3GPP LTE-Advanced, specifically in Release 10 and 11, it was acknowledged through feasibility studies that there existed sufficient capacity for further enhancements in performance. These enhancements were particularly targeted towards the cell edges and involved coordinated transmission among multiple transmitters located across different cell sites [33]. The conventional PS algorithms, such as MAX C/I and Proportional Fairness (PF) algorithms [1], solely focus on enhancing resource utilization based on the channel conditions of users. However, they do not take into account the Quality of Service (QoS) requirements, such as the delay requirements for real-time (RT) traffic or the minimum throughput requirements for non-real-time (NRT) traffic [34]. To address this limitation, several QoS-aware PS algorithms have been developed based on the Modified Largest Waited Delay First (M-LWDF) approach [35]. The M-LWDF algorithm aims to improve the packet delay and Packet Delivery Ratio (PDR) for RT traffic, as well as meet the minimum throughput requirement for NRT traffic over extended periods of time.

The M-LWDF algorithm is utilized in Frequency Division Multiplexing (FDM) systems to optimize the allocation of sub-carriers in OFDMA-based networks [37]. Within the Sum Waiting Time-Based Scheduling (SWBS) algorithm [6], M-LWDF is modified by updating the queue statuses after each allocation of sub-carriers. SWBS prioritizes Real-Time (RT) and Non-Real-Time (NRT) traffic types based on the product of cumulative waiting times of packets and individual user channel conditions. The objective of this algorithm is to improve the Quality of Service (QoS) for RT and NRT traffic in OFDMA-based systems, showcasing enhanced QoS in comparison to M-LWDF. However, despite the improvement in QoS for RT and NRT traffic, these algorithms [34-39] do not simultaneously enhance the overall system throughput and fairness among users.

B. Analysis of Content Delivering

There are currently two primary mechanisms for sharing spectrum in use: distributed and centralized. In light of the findings obtained, a total of ten recommendations have been put forward to enhance the allocation of resources within 5G networks [40]. The non-standalone (NSA) mode in 5G NR involves deploying a system that relies on the control plane of an existing 4G LTE network for control functions, while dedicating 5G NR exclusively to the user plane [41][42]. Various algorithms for resource allocation cater to service control, such as cooperative game theory, a virtual token mechanism, and notable ones like EXP-RULE and Modified-Largest Weighted Delay First (M-LWDF) algorithms that prioritize real-time flows, particularly in the downlink system. To meet the demands of maintaining high data quality without loss or delay, several packet scheduling algorithms have emerged, including Maximum-Largest Weighted Delay First (M-LWDF), Proportional Fair (PF), and Exponential Rule (EXP-RULE) [43-45]. The best CQI scheduler assigns resource blocks (RBs) to the user with the best radio link conditions or channel quality for a particular RB at every TTI. Each cellular user sends a CQI to the eNodeB to perform the scheduling. The eNodeB transmits a reference signal (i.e., downlink pilot) in the downlink channel to the cellular user [46].

The design of 5G NR integrates LTE elements with a new radio access technology that does not support backward compatibility with LTE. It is worth noting that 5G NR functions over a broad spectrum, ranging from less than 1 GHz to 100 GHz. In the domain of 5G, PCF, which stands for Policy Control Function, operates as a Network Function (NF) that assumes the responsibility of shaping policy control and charging regulations for 5G services and applications. Through service-based interfaces, PCF interacts with various network functions such as AMF, SMF, UPF, UDM, AUSF, and AF (Table 1). Its primary role involves authorizing Quality of Service (QoS) and assigning QoS flow identifiers (QFI) for individual data flows. In addition to these functions, PCF enables network slicing, roaming, and mobility management. Moreover, it collects subscriber metrics and provides real-time management of subscribers, applications, and network resources in accordance with the business rules defined by the service provider.

The scheduling process in 5G NR, which is a part of cellular communication including LTE, is mainly controlled by the network, while the UE adheres to the instructions given. Although the scheduling mechanism in NR is similar to LTE, it has a more refined granularity, particularly in time domain scheduling at the physical layer. This increased granularity enables more accurate scheduling and management, particularly in the time domain aspects of the physical layer in NR when compared to LTE.

After conducting a thorough examination of the network architectures of 4G and 5G, a set of eight criteria has been identified for careful consideration during the resource allocation process in 5G networks [42]. These criteria, as presented in Table 3, encompass Delay, Fairness, Packet Loss Ratio, Spectral Efficiency, Throughput, Electrical Energy Usage, Processor Cycle, and Memory.

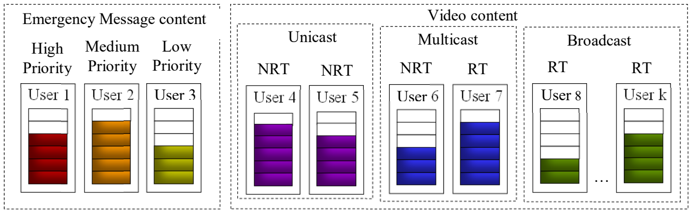
TABLE 3. CRITERIA FOR RESOURCE ALLOCATION

Criteria	Description
Delay	Delay is defined as the latency measured when packets are transmitted from a source to a destination, and it is measured in milliseconds
Fairness	Fairness is the percentage of QoS requirements met during resource allocation.
Packet Loss Ratio	Packet loss ratio is the ratio of packets received by the receiver to the number of packets sent by a sender.
Spectral Efficiency	Spectral efficiency is the rate at which information is transmitted on a provided bandwidth and it is measured in bits/sec/Hz.
Throughput	Throughput is the number of bits in a flow processed per unit of time in a network and it is measured in Kbps.
Electrical Energy Usage	Electrical energy is the electricity used by electronic devices such as routers, switches, servers, and equipment to perform their work. It is usually measured in kWh. Different devices have different needs in terms of electricity according to their functionalities within the 5G end-to-end infrastructure, so having a differentiated approach to electrical energy resource allocation is needed.
Processor Cycle	The processor cycle is the time taken to run an elementary instruction by a processor in a machine or computer. SDN and NFV have increased the need for customized high-end processing on every device along with the 5G end-to-end infrastructure so that assigning processor cores and processor cycles no longer require a one size fits all approach.
Memory	Memory is the amount of main memory or Random-Access Memory (RAM) that is available for all devices in the 5G end-to-end architecture to store running programs. Virtual machines hosting VNFs now have main memory requirements tailor-made for their functionalities.

The development of the 5G NR aims to achieve high throughput, low latency, scalability, ubiquitous connectivity, and energy-efficient solutions. The adaptability and scalability of 5G NR technology within the physical layer design provide benefits, enabling it to support various use cases, such as using orthogonal frequency division multiple access (OFDMA) at the medium access control (MAC) layer with different numerologies. The effectiveness of OFDMA depends on how the access point (AP) allocates channel resources among stations (STAs), with commonly used schedulers including round-robin (RR), maximum rate (MR), and proportional fair (PF). By improving resource allocation, the envisioned 5G NR non standalone version based services can be effectively delivered.

IV. Proposed Algorithms of content delivering for 5G NR service-oriented deployment

The mobile network's amalgamation of various service requests can be likened to a mixed-traffic flow that is directed towards the eNB. This flow showcases a range of Quality of Service (QoS) requirements. To segregate users based on their unique service requests, the Traffic Differentiator funnels them into service-specific queues that are aligned with varying scheduling priorities. Each queue employs a tailored queue sorting algorithm that caters to the specific QoS demands of the respective service type. This study presents the classification of mixed traffic into six distinct queues, with an RT queue comprising emergency message content, followed by RT and NRT queues designated for streaming video services. These queues are structured in descending priority order, with the RT queue positioned at the top, followed by the NRT queue.



Figue 8. Traffic differentiator Stage

Emergency message content is given the highest priority status, while streaming video services are considered to be high-throughput services that can be delivered in real-time or non-real-time. The scheduling information for emergency content is transmitted before the actual data, carrying critical scheduling grants or allocated Physical Resource Blocks (PRBs) that are used for data transmission to users. In the case of non-real-time streaming video, it is crucial to maintain a sustained minimum throughput guarantee in order to ensure high-quality video streaming over extended periods of time. The transmission of emergency information to all scheduled users follows a First Come First Serve (FCFS) approach. Furthermore, two innovative Service Priority Specific queue Sorting Algorithms (SPSSA) have been introduced—one for real-time emergency messages and another for real-time and non-real-time streaming services [47]. The prioritization of users who request background services is determined based on the prioity coefficient, which aims to strike a balance between fairness and system throughput.

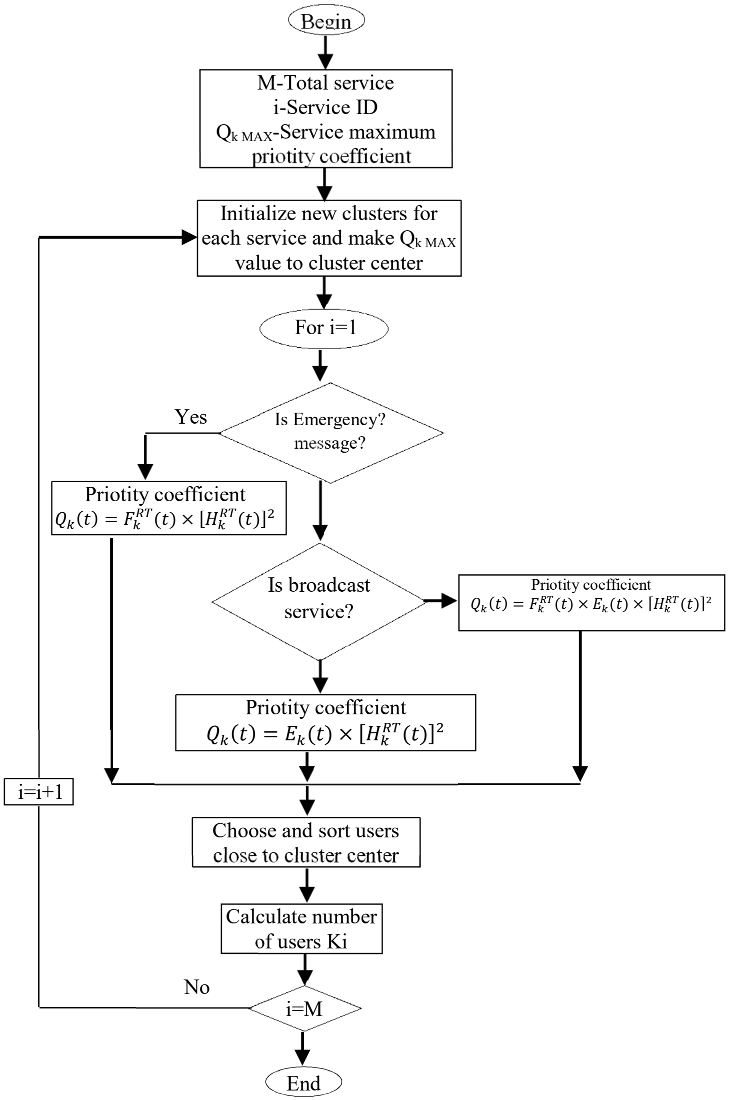

Begin
M-Total service
i-Service ID
$Q_{k\,MAX}$-Service maximum priotity coefficient
Initialize new clusters for each service and make $Q_{k\,MAX}$ value to cluster center
For i=1
Yes
Is Emergency? message?
Priotity coefficient
$Q_k(t) = F_k^{RT}(t) \times [H_k^{RT}(t)]^2$
Is broadcast service?
Priotity coefficient
$Q_k(t) = F_k^{RT}(t) \times E_k(t) \times [H_k^{RT}(t)]^2$
Priotity coefficient
$Q_k(t) = E_k(t) \times [H_k^{RT}(t)]^2$
i=i+1
Choose and sort users close to cluster center
Calculate number of users Ki
No
i=M
End

Figure 9. Traffic differentiator Stage uses Dynamic C-mean clustering algorithm

Figure 9 illustrates the complex operation of the Traffic Differentiator, demonstrating the categorization of mixed traffic into separate segments: RT emergency users, NRT streaming users, and the allocation of their control information into a dedicated queue. This procedure employs a Dynamic c-mean clustering algorithm to effectively organize the priorities of chosen RT users. The arrangement is determined by their individual $Q_k(t)$ priority coefficients, with the objective of enhancing and optimizing the priority metric to enhance system performance. Non standalone 5G NR uses QCI to categorize different types of data traffic based on their requirements and characteristics.

A. Time Domain Scheduler

The TD Scheduler plays a crucial role in determining the users that should be scheduled in the ongoing Transmission Time Interval (TTI), taking into account both the available radio resources and the various user priorities. The TD Scheduler stage introduces a proposed Time Domain Scheduling Algorithm based on Pseudo-Inverse. This algorithm incorporates the Pseudo-Inverse process, which enables adaptive resource allocation. By applying the Pseudo-Inverse (Figure 10) structure, the algorithm identifies the optimal number of Real-Time (RT) and Non-Real-Time (NRT) users to be scheduled in the current TTI. This approach guarantees a fair distribution of resources across different service types, ultimately improving the overall system-level performance.

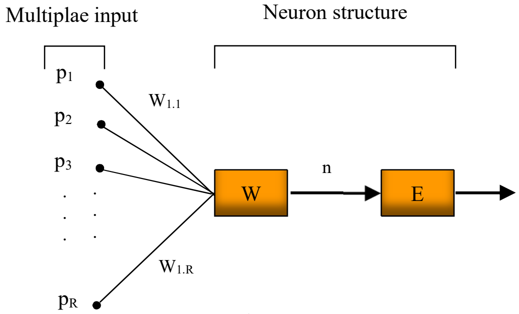


Figure 10. Pseudo-Inverse structure

In the process of pseudo-inverse, the input values are referred to as p_i, whereas the corresponding output values are designated as r. Both the input and output values are expressed as row vectors. By utilizing a methodology similar to that employed in neural networks, we determine the quantity of chosen requests from the complete set of user requests for different services during a given time period (Figure 11). This selection procedure assists in determining the ideal number of requests to be accommodated within a specific timeframe.

$$p_c = \begin{bmatrix} K_1 \\ K_2 \\ \dots \\ K_S \end{bmatrix} \quad s = 1, \dots, S \tag{1}$$

$$P = [p_1 \quad p_2 \quad p_c \quad \dots \quad p_C] \qquad c = 1, \dots, C \tag{2}$$

$$K_{DemandUser\ s} = \sum K \tag{3}$$

P- A total service cluster, p_c- Random service matrix, C- Total service numbers, S- Total service classify

Recall that the task of the linear associator was to produce number of users for an input of s^th service. The primary goal of the linear associator was to produce the tally of users linked to a given service input to furnish the quantity of users connected to a particular service input.

$$t_s = Wp_s \quad (s = 1,2,3) \tag{4}$$

Let's rewrite Eq. (4) in matrix form:

$$T = WP \tag{5}$$

Here, t_s- s^th service output, W - Pseudo weight matrix, T- Selected users from each service.

A coefficient (E) determines whether the desired output value is 100% verified by the processing of multiple input values.

$$E = T - WP \tag{6}$$

The pseudo-inverse weighting matrix has defined output values depending on input multiple input values. A size of weighting matrix equal size of input matrix. Obtain Weight matrix:

$$W = TP^+ \tag{7}$$

$$P^+ = (P^T * P)^{-1} * P^T \tag{8}$$

P^+- Moore Penrose pseudo inverse, P^T- Reverse row and column matrix, P^{-1}- Inverse matrix.

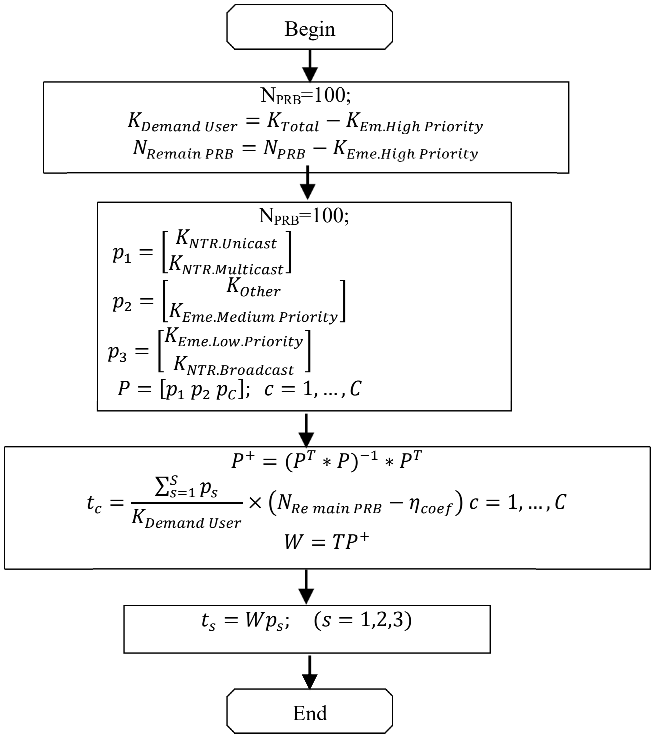

Begin
$N_{PRB}=100;$
$K_{Demand\ User} = K_{Total} - K_{Em.High\ Priority}$
$N_{Remain\ PRB} = N_{PRB} - K_{Eme.High\ Priority}$
$N_{PRB}=100;$
$p_1 = \begin{bmatrix} K_{NTR.Unicast} \\ K_{NTR.Multicast} \end{bmatrix}$
$p_2 = \begin{bmatrix} K_{Other} \\ K_{Eme.Medium\ Priority} \end{bmatrix}$
$p_3 = \begin{bmatrix} K_{Eme.Low.Priority} \\ K_{NTR.Broadcast} \end{bmatrix}$
$P = [p_1\, p_2\, p_C];\;\; c = 1, \dots, C$
$P^+ = (P^T * P)^{-1} * P^T$
$t_c = \frac{\sum_{s=1}^{S} p_s}{K_{Demand\ User}} \times (N_{Re\ main\ PRB} - \eta_{coef})\ c = 1, \dots, C$
$W = TP^+$
$t_s = Wp_s; \quad (s = 1,2,3)$
End

Figure 11. Pseudo-Inverse based Time Domain Scheduling algorithm

B. Frequency Domain Scheduler

The role of the FD Scheduler has a significant value as it determines the specific allocation of Physical Resource Blocks (PRBs) to the users selected by the TD Scheduler. In order to optimize the assignment of PRBs to each user, multi-user diversity is leveraged in the frequency domain. Once the resource allocation decision is made, the Quality of Service (QoS) measurement unit comes into play and calculates key metrics such as the average throughput of Non-Real-Time (NRT) users, average delay, and average Packet Error Rate (PER) of Real-Time (RT) users. This crucial information is then transmitted to the TD Scheduler stage, where it contributes to the decision-making process in the subsequent Transmission Time Interval (TTI). It is worth noting that the availability of accurate Channel Quality Indicator (CQI) reports from all users in each TTI is assumed, as commonly acknowledged in existing literature on channel-aware scheduling.

$$PER_{average} = \frac{1}{K}\sum_{k=1}^{K} PER_k \tag{9}$$

$$PER_k(t) = 1 - \prod_{J=1}^{J}(1 - BLER_j) \tag{10}$$

$PER_{average}$- average PER, $BLER_j$- Block error ratio for each users

The available resource blocks are allocated to user through an iterative process. At each iteration only one RB is allocated to the user which maximizes the following proposed priority function.

$$RF(k) = \frac{RSRP_i(k) \times BLER_i(k) \times P_i(k)}{RSSI_t \times PER_i \times p_i} \tag{11}$$

$RSRP_i(k)$-k-th user power level. $BLER(k)$ – k-th user block error ratio, $P_i(k)$- k-th user service level coefficient, $RSSI_i(k)$-total power level of free band width, PER_i - packet error ratio

The total transmit power of an OFDM symbol in the t^{th} TTI can be expressed as:

$$RSSI_t = \sum_{r=1}^{R}(a_{k.r}, RSRP_i(k)), \ \forall r \in S_r^t \tag{12}$$

```
Algorithm: Optimal CQI selection algorithm

№  Input: $K, S_r, S_t, S_p,$
   Output: $a_{k.r}, b_{k.r}$

Initialization
Set $a_{k.r} = 0, b_{k.r} = 0, t = 1, U = \{1,2,\dots,K\}$
Calculate $RSSI_t$ in Eq(2)
Calculate $PRF(k), \forall k, \forall i \in U$
While $S_r^t \neq 0$
If $U = 0$ Set $S_t = S_t - t$ and GOTO Step($h$)
for $k \in U$
Find $k^* = argmin_k \frac{|S_r^k|}{PRF(k)}$
Find $t^* = argmin_t |S_r^t \cap S_r^k|, \forall \in S_t$
Find $r^* = argmax_r S_{k^*,r}, \forall r \in S_r^{t^*} \cap S_r^{k^*}$
Set $a_{k^*,r^*} = 1, \ b_{k^*,r^*} = CQI_{r^*}(k^*)$
Set $S_r^t = S_r^t - r^*, S_{r^*}^{k^*} = S_{r^*}^{k^*} + r^*$
Add n to user subset $U_i = \{U_1, U_2, \dots U_k\}$
if $t > |S_t|$ Set $t = 1$ else Set $t = t + 1$
```

The frequency of reporting CQI is typically set to once every few milliseconds. It is recommended in [46] that, in static channel conditions, the UE report CQI such that it achieves a block error rate (BLER) 10% when scheduled data corresponding to the median reported CQI. For example, in HSDPA, the UE monitors the quality of the downlink wireless channel and periodically reports this information to the base station (referred to here as NodeB) on the uplink [48]. The channel quality indicator (CQI) feedback mechanism provides various signaling options to enable the radio link adaptation between a user equipment (UE) and a base-station.

V. CONCLUSION

In the era of a content delivering growing traffic, the broadcast and multicast technologies will be able to enhance the efficiency of network resource utilization. Non-standalone 5G New Radio (NR) access technology in the physical layer design accommodate to be flexible and scalable that can support many use cases such as orthogonal frequency division multiple access (OFDMA) at its medium access control (MAC) layer and numerology. This research supports a fair distribution of resources across different service types, ultimately improving the overall system-level performance for non-standalone 5G NR system. The proposed algorithms of a content delivering for 5G NR service-oriented deployment focuses on the classification of mixed traffic into distinct queues, with an RT queue comprising emergency message content, followed by RT and NRT queues designated for streaming video services.

Traffic Differentiator Traffic FD Scheduler supports CQI is to determine the specific allocation of Physical Resource Blocks (PRBs) to the users selected by the TD Scheduler. This study analyzed the downlink access with varying resource allocations the number of user equipment, and numerology with additional proposed schedulers. Furthermore, the further studies focus on the ultimate results of the proposed algorithms compared to the other related algorithms. In addition, our future research will intend to get the effective result of to use AI in the TD scheduler to intelligently pick users from the queues to improve the QoS of different traffic types and overall system performance.

REFERENCE

[1] C.-X. Wang et al., ‘‘Cellular architecture and key technologies for 5G wireless communication networks,’’ IEEE Comm. Mag., vol. 52, no. 2, pp. 122–130, Feb. 2014.

[2] Rinaldi, F., Raschellà, A. & Pizzi, S. 5G NR system design: a concise survey of key features and capabilities. Wireless Netw 27, 5173–5188 (2021). https://doi.org/10.1007/s11276-021-02811

[3] S. Kumar, T. Agrawal, and P. Singh, “A Future Communication Technology : 5G,” International Journal of Future Generation Communication and Networking, vol. 9, no. 1, pp. 303-310, 2016, doi: 10.14257/ijfgcn.2016.9.1.26.

[4] Gupta A, Jha RK. A survey of 5G network: Architecture and emerging technologies. IEEE Access. 2015;3:1206-1232. DOI: 10.1109/ACCESS.2015.2461602

[5] B. Sokappadu, A. Hardin, A. Mungur, S. Armoogum, "Software Defined Networks: Issues and Challenges," Conference on Next Generation Computing Applications (NextComp), Mauritius, 2019.

[6] Technical Specification: NR Physical Layer Procedures for Control. TS 38.213 V16.2.0. 2020

[7] Kamal MA, Raza HW, Alam MM, Su'ud MM, Sajak ABAB. Resource Allocation Schemes for 5G Network: A Systematic Review. Sensors (Basel). 2021 Oct 2;21(19):6588. doi: 10.3390/s21196588. PMID: 34640908; PMCID: PMC8512213.

[8] Degambur, L. -., Mungur, A., Armoogum, S., & Pudaruth, S. (2021).Resource allocation in 4G and 5G networks: A review. International Journal of Communication Networks and Information Security, 13(3),401-408.

[9] 3GPP (2020). TS 38.213; NR; Physical layer procedures for control; Rel.16; 3GPP: Sophia Antipolis Valbonne, France.

[10] Rinaldi, F., Raschellà, A. & Pizzi, S. 5G NR system design: a concise survey of key features and capabilities. Wireless Netw 27, 5173–5188 (2021). https://doi.org/10.1007/s11276-021-02811

[11] https://telcomaglobal.com/]

[12] 3GPP (2014). TR 36.842; Study on small cell enhancements for E-UTRA and E-UTRAN; Higher layer aspects; Rel.12. 3GPP: Sophia Antipolis Valbonne, France.

[13] R. N. Mitra, P. Agrawal, “5G Mobile Technology: A Survey”, ICT Express, Vol. 1, No. 3, pp. 132-137, 2015.

[14] L. B. Le, V. Lau, E. Jorswieck, N-D. Dao, A. Haghighat, D. I. Kim, T. Le-Ngoc, “Enabling 5G mobile wireless technologies,” EURASIP Journal on Wireless Communications and Networking, Article 218, 2015.

[15] B. Bangerter, S. Talwar, R. Arefi, K. Stewart, “Network and Devices for the 5G Era,” IEEE Communications Magazine, Vol. 52, No. 2, pp. 90-96, 2014

[16] T. Badarch and O. Bataa, "Big Data Area: A Novel Network Performance Analysis Technique Based on Bayesian Traffic Classification Algorithm," 2017 International Conference on Green Informatics (ICGI), Fuzhou, China, 2017, pp. 238-245, doi: 10.1109/ICGI.2017.46.

[17] O. Bataa, E. Lamjav, U. Purevdorj, Y. i. Kim and K. Gonchigsumlaa, "ICS Algorithm Implementation in 3GPP/LTE," 2014 7th International Conference on Ubi-Media Computing and Workshops, Ulaanbaatar, Mongolia, 2014, pp. 17-21, doi: 10.1109/U-MEDIA.2014.41.

[18] M. Peng, Y. Li, Z. Zhao, and C. Wang, ‘‘System architecture and key technologies for 5G heterogeneous cloud radio access networks,’’ IEEE Netw., vol. 29, no. 2, pp. 6–14, Mar./Apr. 2015. Available: http://ieeexplore.ieee.org/document/7986957/

[19] T. Badarch and O. Bataa, "An adaptive scheduling scheme to efficient emergency call holding times in public safety network," Ifost, Ulaanbaatar, Mongolia, 2013, pp. 210-213, doi: 10.1109/IFOST.2013.6616888.

[20] W. Mansouri, K. B. Ali, F. Zarai, and M. S. Obaidat, ‘‘Radio resource management for heterogeneous wireless networks: Schemes and simulation analysis,’’ in Modeling and Simulation of Computer Networks and Systems. Amsterdam, The Netherlands: Elsevier, 2015, pp. 767–792. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780128008874000274

[21] S. Shakkottai and R. Srikant, ‘‘Scheduling real-time traffic with deadlines over a wireless channel,’’ in Proc. 2nd ACM Int. Workshop Wireless Mobile Multimedia (WOWMOM). New York, NY, USA: ACM Press, 1999, pp. 35–42. [Online]. Available: http://portal.acm.org/citation.cfm?doid=313256.313273

[22] X. Yuan and Z. Duan, ‘‘Fair Round-Robin: A low complexity packet scheduler with proportional and worst-case fairness,’’ IEEE Trans. Comput., vol. 58, no. 3, pp. 365–379, Mar. 2009. [Online]. Available: http://ieeexplore.ieee.org/document/4626953/

[23] S. Schwarz, C. Mehlführer, and M. Rupp, ‘‘Low complexity approximate maximum throughput scheduling for LTE,’’ in Proc. Conf. Rec. 44th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR), Nov. 2010, pp. 1563–1569. [Online]. Available: http://ieeexplore.ieee.org/document/5757800/

[24] G. Femenias, F. Riera-Palou, X. Mestre, and J. J. Olmos, ‘‘Downlink scheduling and resource allocation for 5G MIMO-multicarrier: OFDM vs FBMC/OQAM,’’ IEEE Access, vol. 5, pp. 13770–13786, 2017. [Online].

[25] S. O. Aramide, B. Barakat, Y. Wang, S. Keates, and K. Arshad, ‘‘Generalized proportional fair (GPF) scheduler for LTE-A,’’ in Proc. 9th Comput. Sci. Electron. Eng. (CEEC), Sep. 2017, pp. 128–132. [Online]. Available: http://ieeexplore.ieee.org/document/8101612/

[26] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, ‘‘Providing quality of service over a shared wireless link,’’ IEEE Commun. Mag., vol. 39, no. 2, pp. 150–154, Feb. 2001. [Online]. Available: http://ieeexplore.ieee.org/document/900644/

[27] M. Mahfoudi, M. E. Bekkali, A. Najd, M. E. Ghazi, and S. Mazer, ‘‘A new downlink scheduling algorithm proposed for real time traffic in LTE system,’’ Int. J. Electron. Telecommun., vol. 61, no. 4, pp. 409–414, Dec. 2015. [Online]. Available: http://journals.pan.pl/dlibra/publication/101933/edition/87947/content

[28] C. J. Katila, C. Buratti, M. D. Abrignani, and R. Verdone, ‘‘Neighborsaware proportional fair scheduling for future wireless networks with mixed MAC protocols,’’ EURASIP J. Wireless Commun. Netw., vol. 017, no. 1, p. 93, Dec. 2017. [Online]. Available: https://jwcneurasipjournals.springeropen.com/articles/10.1186/s13638-017-0875-6

[29] M. I. Saglam and M. Kartal, ‘‘5G enhanced mobile broadband downlink scheduler,’’ in Proc. 11th Int. Conf. Electr. Electron. Eng. (ELECO), Nov. 2019, pp. 687–692. [Online]. Available: https://ieeexplore.ieee.org/document/8990378/

[30] W. S. Afifi, A. A. El-Moursy, M. Saad, S. M. Nassar, and H. M. El-Hennawy, ‘‘A novel scheduling technique for improving cell-edge performance in 4G/5G systems,’’ Ain Shams Eng. J., vol. 12, no. 1, pp. 487–495, Mar. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2090447920301878

[31] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, ‘‘A RAN resource slicing mechanism for multiplexing of eMBB and URLLC services in OFDMA based 5G wireless networks,’’ IEEE Access, vol. 8, pp. 45674–45688, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9020161/

[32] A. Karimi, K. I. Pedersen, and P. Mogensen, ‘‘Low-complexity centralized multi-cell radio resource allocation for 5G URLLC,’’ in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), May 2020, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9120469

[33] W. Nam, D. Bai, J. Lee, and I. Kang, ‘‘Advanced interference management for 5G cellular networks,’’ IEEE Commun. Mag., vol. 52, no. 5, pp. 52–60, May 2014.

[34] B. Chisung, and C. Dong, “Fairness-Aware Adaptive Resource Allocation Scheme in Multihop OFDMA System”, “Communication letters, IEEE, vol.11,pp.134- 136,Feb.2007

[35] M. Andrews, K. Kumaran, K. Ramanan, A Stolyar and P. Whiting,”Providing Quality Over a Shared Wireless Link” IEEE communications magazine, February 2001.

[36] S. Shakkottai, A.L. Stolyar,”Scheduling for Multiple Flows Sharing a Time-Varying Channel: The Exponential Rule” Bell Labs, Lucent Technologies, NJ 07974.

[37] P. Parag, S. Bhashyam, and R. Aravind, “A subcarrier allocation algorithm for OFDMA using buffer and channel state information”, Vehicular Technology Conference, 2005.VTC-2005-Fall.2005 IEEE 62nd, 2005, pp.622-625.

[38] Jun Shen, Na Yi, An Liu and Haige Xiang, “Opportunistic scheduling for heterogeneous services in downlink OFDMA system,” School of EECS, Peking University, Beijing, P.R.China, IEEE computer Society 2009,pp.260- 264.

[39] T. Badarch and O. Bataa, "Incoming traffic modeling of heterogeneous Public Safety Network," 2013 15th Asia-Pacific Network Operations and Management Symposium (APNOMS), Hiroshima, Japan, 2013, pp. 1-6

[40] Degambur, Lavanya-Nehan & Mungur, Avinash & Armoogum, Sheeba & Pudaruth, Sameerchand. (2021). Resource Allocation in 4G and 5G Networks: A Review. International Journal of Communication Networks and Information Security (IJCNIS). 13. 10.54039/ijcnis.v13i3.5116.

[41] 5G NR Deployment Scenarios or modes-NSA,SA,Homogeneous,Heterogeneous". rfwireless-world.com.

[42] Junko Yoshida. "What's Behind 'Non-Standalone' 5G?". Eetimes.com. Retrieved 2018-11-13.

[43] Ameigeiras, P., Wigard, J., & Mogensen, P. (Sep. 2004). Performance of the m-lwdf scheduling algorithm for streaming services in hsdpa. In IEEE Transactions on vehicular technology conference vol. 2, pp. 999–1003 Los Angeles, USA.

[44] Choi, J.-G., & Bahk, S. (2007). Cell-throughput analysis of the proportional fair scheduler in the single-cell environment. IEEE Transactions on Vehicular Technology, 56(2), 766–778.

[45] Sadiq, B., Madan, R., & Sampath, A. (2009). Downlink scheduling for multiclass traffic in LTE. EURASIP Journal on Wireless Communications and Networking, 2009, 1–18

[46] Habaebi MH, Chebil J, Al-Sakkaf AG, Dahawi TH. Comparison between scheduling techniques in long term evolution. IIUMEJ. 2013;14(1)

[47] R. Kausar, Y. Chen and K. K. Chai, "Service Specific Queue Sorting and Scheduling Algorithm for OFDMA-Based LTE-Advanced Networks," 2011 International Conference on Broadband and Wireless Computing, Communication and Applications, Barcelona, Spain, 2011, pp. 116-121, doi: 10.1109/BWCCA.2011.22.

[48] T. Cui, F. Lu, V. Sethuraman, A. Goteti, S. P. N. Rao and P. Subrahmanya, "Throughput Optimization in High Speed Downlink Packet Access (HSDPA)," in IEEE Transactions on Wireless Communications, vol. 10, no. 2, pp. 474-483, February 2011, doi: 10.1109/TWC.2010.120610.091294.

NARANTUYA Vandantseren in 2002 graduated from the Mongolian University of Science and Technology majoring in telecommunication. Bachelor degree thesis: Investigation Asynchronous Transfer Mode Technology, Master degree thesis (M.Sc.) in 2004: 3G mobile communication technology research. Doctoral student of Mongolian University of Science and Technology. Works in Mongolian National Defense University. Research topic: The Development of Defense communication service based on Mobile broadband convergence network.

CHULUUNBANDI Naimannaran, in 1994 graduated from the Mongolian University of Science and Technology majoring in information engineering and electronics. Master degree thesis (M.Sc.) in 2004: Optimization of Rural Radio communications. PhD degree thesis in 2012: Graph theory based performance evaluation for Rural cellular network services. Associate Professor, Consulting engineer of Mongolia. Research direction: Mobile Broadband technologies (WiMAX, WiBro, Mobile IPTV, 4G LTE, 5G technology etc.).

TUYATSETSEG Badarch, in 1996 graduated from the MUST majoring in Telecommunications engineer. Master degree thesis (M.Sc.) in 2013 in Northeastern University, USA: The Bayesian mixture model for traffic classification in Computer Networks. PhD degree in 2015: Performance Analysis of Integrated information network. MBA from Kansas Park University in 2017; Associate Professor at the Department of Computer Science at SICT of MUST. Research interest: Bayesian models for Traffic classification, clustering model in big data, AI.

ERDENEBAYAR Lamjav, in 2010 graduated from the Mongolian University of Science and Technology, Bachelor degree thesis: — The Study of Artificial Line. Master degree thesis (M.Sc.) in 2011: The study of link level simulation for Wimax. PhD degree thesis in 2018: The Study of Resource allocation in LTE networks. He is an Associate Professor at Department of Communication at SICT of MUST. Research direction: Mobile Broadband, high speed integrated services technologies (WiMAX, WiBro, Mobile IPTV, 4G LTE, 5G technology etc.).

OTGONBAYAR Bataa, in 1978 graduated from Polytechnic Institute of Mongolia majoring in Radio communication engineer. Bachelor degree thesis: Master degree thesis (M.Sc.) in 1995: Some issues of speech synthesis. PhD degree thesis in 1996: Study of Mongolian speech synthesis and applying it in telecom techniques, in 2003, post Ph.D. program thesis: Optimal version of OFDM system frequency and timing offset. Professor, Consulting engineer of Mongolia. Research topic: Mobile Broadband, high speed integrated services technologies.

Design of Communication Countermeasure Simulation Model and Data Interaction Interface for Battlefield Network Based on QualNet

Wenyi LI*, Peng GONG*, Weidong WANG*, Yu LIU*, Jianfeng LI*, Xiang GAO*

*National Key Laboratory of Mechatronic Engineering and Control, School of Mechatronical Engineering, Beijing Institute of Technology, Beijing, China

18215621889@163.com, penggong@bit.edu.cn, 3220185030@bit.edu.cn, 67577335@qq.com, lidanhai@sina.com, bitxianggao@bit.edu.cn

***Abstract*—The performance analysis of battlefield communication network has been more and more complex and difficult with its increasing scale, heterogeneity and geographical distribution of nodes. Computer simulation technology is considered as a potential technology to efficiently and accurately solve this problem. This paper focuses on the simulation requirements of anti-interference performance of battlefield communication networks in complex electromagnetic environments, and designs reconnaissance interference and frequency hopping models based on QualNet simulation software. The model introduces scout and jammer nodes in the communication network, which can conduct reconnaissance and directional interference on communication nodes in the network. Other nodes can set frequency hopping parameters to achieve anti-interference. In addition, a data interaction interface for the distributed simulation system is designed based on the DDS specification, and a structure definition file is designed according to the data interaction requirements to achieve dynamic control of the QualNet simulation model by external control modules. Finally, this article tested the functionality of the communication countermeasure model and conducted a delay test on the data interaction interface. The experimental results verify the functionality of the designed model and the high real time of the interface, which is of great significance to the anti-interference performance assessment of the battlefield communication network.**

***Keywords*—Battlefield network; QualNet; Communication countermeasure; DDS; Data interaction interface**

I. INTRODUCTION

Since the 21st century, network simulation has emerged as the primary methodology for investigating and analyzing the network technology [1-2]. Using simulation software to simulate battlefield networks has also become a hot topic [3-4]. The battlefield network has the characteristics of larger scale, more types of nodes, and wider distribution of nodes. Compared with discrete event network simulators such as NS-2 and OPNET, QualNet has strong advantages in large-scale and high-precision network simulation [5-7].

QualNet adopts parallel design, and its simulation capability can reach tens of thousands of nodes. The algorithm design is excellent, and its simulation speed for large-scale networks is dozens of times that of other simulators. It adopts modular design and has strong scalability. The user interface is flexible and it supports users for secondary development. It supports real-time communication with the real network, and the simulation accuracy is close to the real network [8-11]. Therefore, using QualNet for battlefield network simulation can effectively improve the speed and accuracy of simulation.

Battlefield networks usually face complex electromagnetic environments. The research on interference methods for battlefield networks is endless. Literature [12] proposed a fuse interference decision-making method based on Q-learning algorithm. Literature [13] proposes an interference suppression method for LTE signal based passive bistatic radar. It can be seen that the simulation of battlefield networks must also consider a comprehensive evaluation of the anti-interference performance of the entire network in a complex electromagnetic environment. Therefore, it is necessary to introduce the reconnaissance interference model and anti-interference model into network simulation to simulate and evaluate the anti-interference performance of the network [14].

A complete communication countermeasure network simulation requires the construction of a distributed simulation platform, which requires a series of tasks such as simulation scenario planning, simulation process deduction, and simulation data analysis and evaluation [15]. The functions of each part of the platform are separated, and the simulation model is controlled through the data interaction interface. Currently, common methods for implementing data interaction in distributed simulation systems include HLA and DDS technologies.

HLA(High Level Architecture) is an event-centric, universal, reconfigurable software architecture that supports heterogeneous device access. It is usually used to develop and execute very large distributed simulation applications. However, its real-time performance is poor, so it is unable to perform real-time dynamic control of the simulation model [16-17].

DDS is data-centric. In a distributed system, it abstracts and encapsulates the resources provided by the operating system and provides a variety of advanced services and functions for applications, such as communication or data sharing [18]. The data sharing implemented by DDS can be understood as an abstract "global data space". Applications can access this "global data space" in the same way as if it accessed the local storage space [19], even if the development language or the type of operating system are probably different. DDS also provides a very flexible QoS policy to meet users' different needs for data sharing methods, such as reliability, fault handling, etc [20-21].

In summary, in response to the characteristics of large scale, fast speed, and high accuracy of battlefield network simulation, as well as the need for network anti-interference performance evaluation, this article introduces communication countermeasure models into network simulation based on QualNet's secondary development, and designs data interaction interfaces for distributed simulation systems based on DDS specifications.

II. Communication Countermeasure Model Design

A. Overall Architecture

The overall architecture of the communication countermeasure model considered in this article is shown in Figure 1. The introduction of each module is as follows:

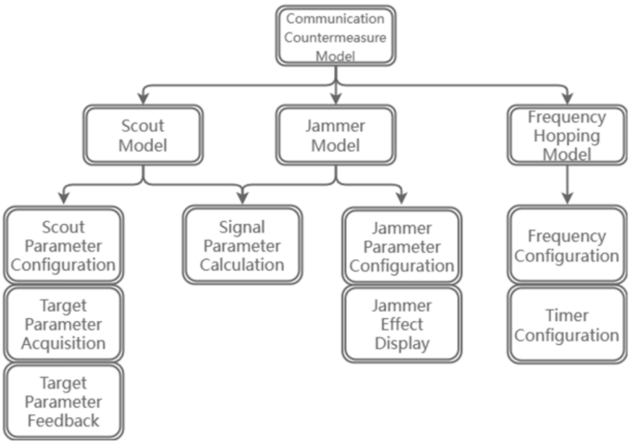

Communication Countermeasure Model
Scout Model
Jammer Model
Frequency Hopping Model
Scout Parameter Configuration
Signal Parameter Calculation
Jammer Parameter Configuration
Frequency Configuration
Target Parameter Acquisition
Jammer Effect Display
Timer Configuration
Target Parameter Feedback

Figure 1. Architecture of communication countermeasure model

Scout model: It's an application layer model, which can conduct reconnaissance by configuring parameters such as direction, interval, and working frequency of scout nodes. Besides, it can also display parameters such as the position, distance, working frequency, and signal power of target nodes in the network. The parameters of scout node can be dynamically configured.

Jammer model: It's an application layer model, which can interfere with the communication between target nodes by configuring parameters such as direction, frequency band, and interference power of jammer nodes based on target information in reconnaissance intelligence. The parameters of jammer node can also be dynamically configured.

Frequency hopping model: It's a physical layer model, which enables frequency hopping communication between nodes by configuring the same frequency hopping pattern for nodes in the same subnet, thereby achieving anti-interference of the communication network.

B. Scout Model Workflow

The workflow of the reconnaissance interference model is shown in Figure 2.

Before the simulation starts, we need to set a node in the network as a scout node according to the simulation requirements, and configure parameters such as the working mode, interval, and working frequency.

After the simulation starts, the scout node will traverse the other nodes in the network and judges the communication status of the target node. If the target node is within the reconnaissance range and is in the signal transmitting state, the scout node will calculate the transmission parameters of the signal on the path from the target node to itself.

When the signal reaches the scout node, the signal power need to be compared with the receiver threshold. If the signal power is greater than the receiver threshold, it is considered to be detectable. At this time, the parameters of the target node such as location, distance and communication frequency are extracted and encapsulated, and sent to the front end through the data interaction interface. After that, the scout node will enter the next cycle. While the scout node is working, the working parameters of it can be dynamically adjusted through the data interaction interface.

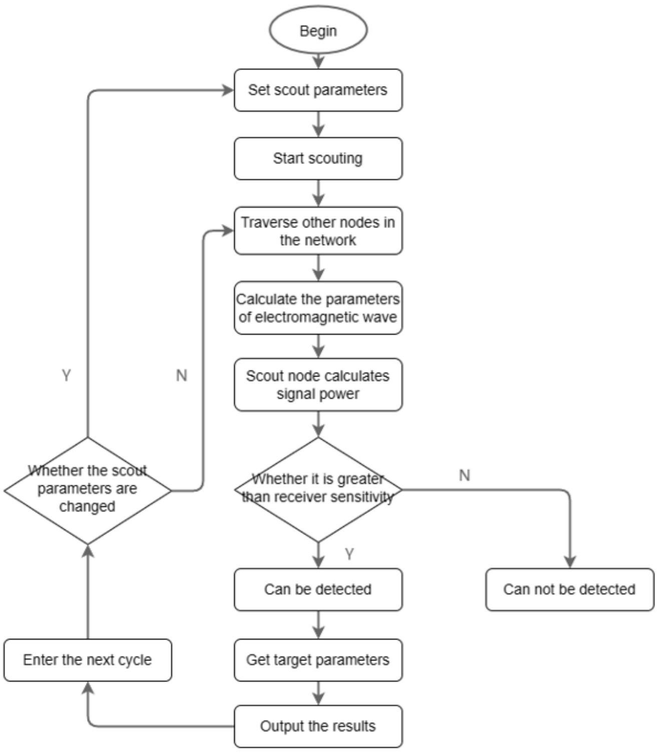

Begin
Set scout parameters
Start scouting
Traverse other nodes in the network
Calculate the parameters of electromagnetic wave
Scout node calculates signal power
Whether it is greater than receiver sensitivity
N
Y
Can be detected
Can not be detected
Get target parameters
Output the results
Enter the next cycle
Whether the scout parameters are changed
Y
N

Figure 2. Workflow of scout model

C. Jammer Model Workflow

The commonly used interference methods include energy type interference and information type interference, and their workflow is shown in Figure 3.

Based on the reconnaissance intelligence obtained by the scout node, we need to configure parameters such as the working mode, direction, power, and frequency band of the jammer nodes.

When a node in the network receives a data packet, it will traverse other nodes in the network, and find the jammer node to obtain the parameters.

Afterwards, the node will enter the interference calculation module and determine whether itself is affected by the interference according to the direction, frequency, power, and time.

If it is in a state of interference, different treatments will be carried out according to different interference methods. If it is energy type interference, the interference power will be loaded onto the noise power. If it is information type interference, the content of data packet will be randomly tampered.

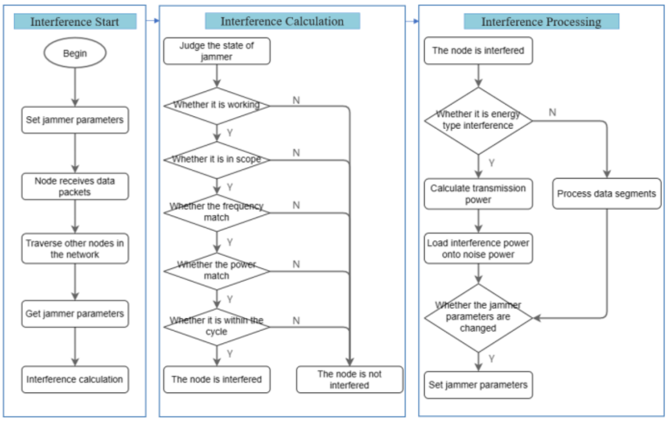

Interference Start
Begin
Set jammer parameters
Node receives data packets
Traverse other nodes in the network
Get jammer parameters
Interference calculation
Interference Calculation
Judge the state of jammer
Whether it is working
Whether it is in scope
Whether the frequency match
Whether the power match
Whether it is within the cycle
N
Y
The node is interfered
The node is not interfered
Interference Processing
The node is interfered
Whether it is energy type interference
N
Y
Calculate transmission power
Process data segments
Load interference power onto noise power
Whether the jammer parameters are changed
Y
Set jammer parameters

Figure 3. Workflow of jammer model

D. Frequency Hopping Model Workflow

The workflow of the frequency hopping model is shown in Figure 4.

Before starting the simulation, we need to set parameters such as start frequency, end frequency, time interval, and random seed in the configuration file of frequency hopping model. When the simulation begins, if frequency hopping is not enabled, the communication frequency will be fixed. If frequency hopping is enabled, the start frequency, end frequency, and time interval in the configuration file will be read, and the frequency hopping pattern will be generated based on the seed. Afterwards, the timer is started to change the network communication frequency. The frequency hopping patterns of different subnets can be different.

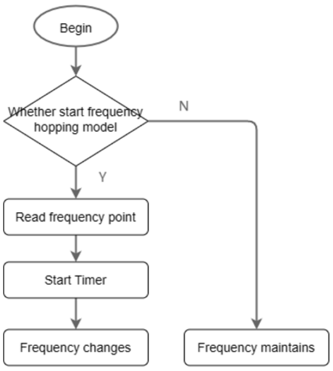


Figure 4. Workflow of frequency hopping model

III. Data Interaction Design

A. Data Interaction Interface Overall Architecture

The data interaction of DDS is based on publish-subscribe, and its principle is shown in Figure 5.

At the beginning of communication, a data interaction space called Domain is allocated. Each entity participating in communication is within the Domain, called Domain Participant. Domain Participants in different Domains cannot perform data interaction. The data interacted exists within the Domain as a unified Topic.

The Domain Participant at the data sender side needs to register a Publisher to send data. Publisher registers the data as a Topic and publishes it to Domain through the DataWriter.

All Topics are in the global data space, and the receivers obtain the data from the global data space according to their own needs. The Domain Participant at the data receiver side needs to register a Subscriber to read data. The Subscriber obtains the Topic from the Domain through the DataReader, thereby achieving the purpose of data interaction. In addition, each Domain Participant needs to be equipped with a Listener to listen for new Topic published in the Domain.

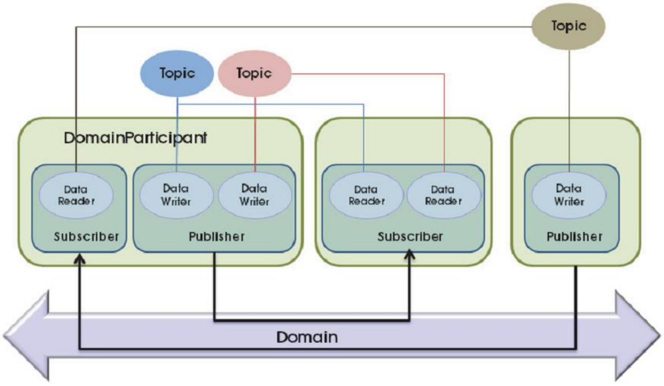


Figure 5. Principle of DDS data interaction

The architecture of the data interaction interface based on DDS is shown in Figure 6. This interface is responsible for achieving data transmission between communication countermeasure model and external simulation control software.

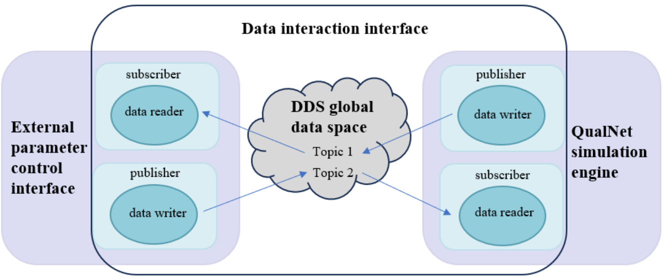


Figure 6. Architecture of data interaction interface

The data interaction interface consists of three parts:

QualNet simulation engine data interface: This interface follows the specifications for the secondary development of external interfaces in QualNet. It listens to the requests and responses from external simulation control modules, and publishes the real-time simulation situation information.

External simulation control module data interface: This interface is called by external simulation control modules to issue simulation control instructions and listen for response information from the QualNet simulation engine.

DDS global data space: The publish-subscribe model establishes a virtual shared global data space for all distributed simulation modules. The Publisher of each module sends the data to the global data space, and the Subscriber obtains the corresponding data from the global data space according to their own needs.

B. Core Data Structure of Data Interaction Interface

According to the requirements of communication countermeasure network simulation testing, the data interacted between the QualNet simulation engine and external simulation control module includes model parameter settings, model parameter responses, and scout result display. The interaction relationship is shown in Table 1.

TABLE 1. DATA INTERACTION RELATIONSHIP

Scene	Data	Sender	Receiver
model parameter settings	model parameters	External control module	QualNet simulation software
model parameter responses	model parameters	QualNet simulation software	External control module
reconnaissance result display	Scout results	QualNet simulation software	External control module

Based on the data interaction relationship between each module, a data structure definition file in Topic was designed, and its core data structure is shown in Tables 2, 3, and 4. The ScoutParamsState structure in Table 2 and the JammerParamsState structure in Table 3 describe the parameter configuration of scout and jammer nodes, while the ScoutTargetState structure in Table 4 describes the target parameter information obtained by scout nodes.

TABLE 2. SCOUTPARAMSSTATE STRUCTURE

Variable Name	Variable Type	Variable Description
active	bool	Work or not
scoutMode	int	Scout work mode
interval_ms	float	Scout work cycle
gain_dB	float	Scout antenna gain
startFrequency_mHz	float	Scout start work frequency
endFrequency_mHz	float	Scout end work frequency
scoutHAngleStart	double	Start angle of horizontal
scoutHAngleEnd	double	End angle of horizontal
scoutVAngleStart	double	Start angle of vertical
scoutVAngleEnd	double	End angle of vertical
threshold_dBm	float	Receiver sensitivity

TABLE 3. JAMMERPARAMSSTATE STRUCTURE

Variable Name	Variable Type	Variable Description
active	bool	Work or not
jammerHAngleStart	double	Start angle of horizontal
jammerHAngleEnd	double	Scout end work frequency
jammerVAngleStart	double	Start angle of horizontal
jammerVAngleEnd	double	Start angle of horizontal
stratFrequency	double	Jammer start work frequency
endFrequency	double	Jammer end work frequency
txPower_dBm	float	Jammer transmitting power
gain_dB	float	Jammer antenna gain
startTime	short	Interference start time
endTime	short	Interference end time
interval	short	Jammer work cycle
jammerMode	short	Jammer work mode
smartJammerMode	short	Jammer work mode of information type

TABLE 4. SCOUTTARGETSTATE STRUCTURE

Variable Name	Variable Type	Variable Description
targetID	int	Target node ID
lon	float	longitude of target node
lat	float	Latitude of target node
height	float	Height of target node
bandwidth	short	Target working bandwidth
signalFrequency	float	Target signal frequency
scoutID	int	Scout node ID
signalPower	float	Target signal power
targetAngle	float	Signal arrival angle
distance	float	Target distance
time	short	Discovery time of target

C. Data Interaction Interface Workflow

The workflow of the data interaction interface is shown in Figure 7.

Interface initialization: While initializing the interface, we need to create a data space Domain and simultaneously create Domain Participants on both receiving and sending ends. According to the DDS specification, we should encapsulate the data that needs to be interacted into a unified data structure and create corresponding Topics. The sender creates a Publisher and corresponding DataWriters for different Topics. The receiver creates Subscribers and corresponding DataReaders, and simultaneously creates Listeners to listen to data in the Domain.

Data sending: After the simulation starts, each module encapsulates the data to be sent into a specific structure, and then calls the corresponding data sending function, and uses DataWriter to send the data to the Domain.

Data receiving: The Listener will keep listening to the global data space. while new data is arriving, it will judge the type of data, and call the corresponding DataReader of Subscriber to receive the data. At last, it will call the corresponding data processing function to parse and process the data.

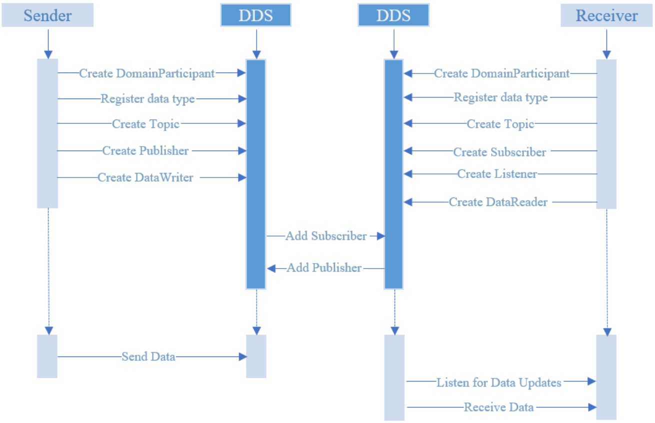


Figure 7. Workflow of data interaction interface

IV. Experiments And Results

A. *Simulation Environment and Scenarios*

According to actual requirements, the experimental testing software and hardware environment is shown in Figure 8. The simulation scenario is generated by the simulation scenario generation host and sent to communication network simulation host which runs QualNet simulator. The communication countermeasure model is added to the scenario and can be dynamically controlled through the model dynamic control host. The real-time simulation information will be shown on the global situation display host. Communication between hosts is achieved through DDS components. The specific configuration is shown in Table 5.

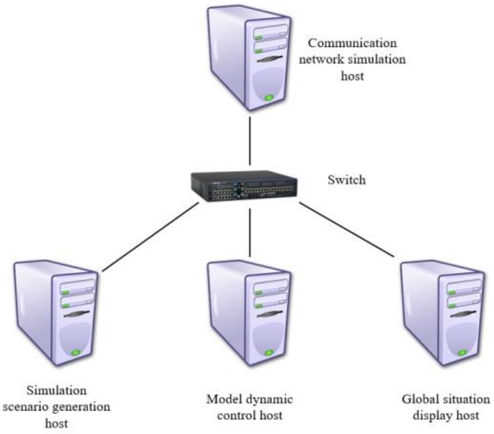

Communication network simulation host
Switch
Simulation scenario generation host
Model dynamic control host
Global situation display host

Figure 8. Experiment environment

TABLE 5. Hosts Configuration

Server	Running Software	Memory	Processor	Operating System
Communication network simulation host	QualNet 5.1	4GB	Intel core i5 CPU M520@2.40GHz*4	Ubuntu 12.04 LTS 64bit
Simulation scenario generation host	Simulation scenario generation software	4GB	Intel(R) Core(TM) i3-2310M CPU @ 2.10GHz	Windows 10 Ultimate 64bit
Model dynamic control host	reconnaissance interference dynamic interaction software	4GB	Intel(R) Core(TM) i3-2310M CPU @ 2.10GHz	Windows 10 Ultimate 64bit
Global situation display host	global situation display software	4GB	Intel(R) Core(TM) i3-2310M CPU @ 2.10GHz	Windows 10 Ultimate 64bit

We created a simulation test scenario in the simulation scenario generation software as shown in Figure 9. The scenario contains 11 nodes. Nodes 1-10 are communication nodes for 10 fighters, while node 11 is a scout and jammer integrated node.

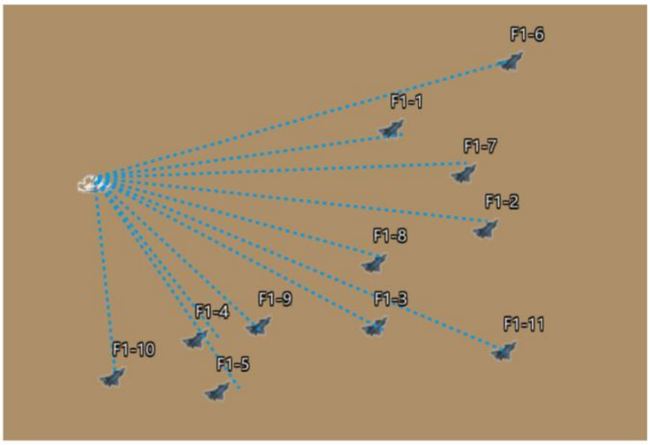
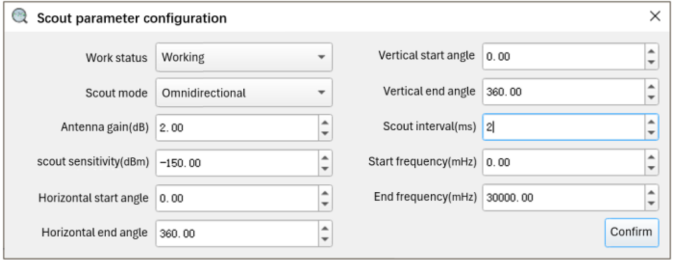


Figure 9. Network simulation scenario

B. *Communication Countermeasure Model Test*

When starting the simulation, the scenario script file will be sent to the communication network simulation host. 45 seconds after the simulation starts, the scout node is turned on and the parameters are configured as shown in Figure 10.

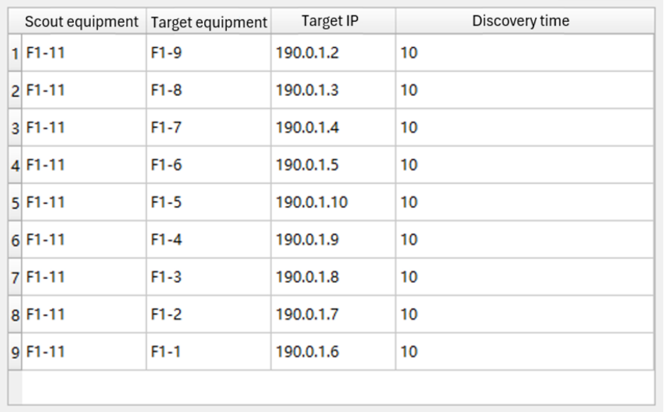


Scout parameter configuration

Work status: Working
Scout mode: Omnidirectional
Antenna gain(dB): 2.00
scout sensitivity(dBm): -150.00
Horizontal start angle: 0.00
Horizontal end angle: 360.00
Vertical start angle: 0.00
Vertical end angle: 360.00
Scout interval(ms): 2
Start frequency(mHz): 0.00
End frequency(mHz): 30000.00

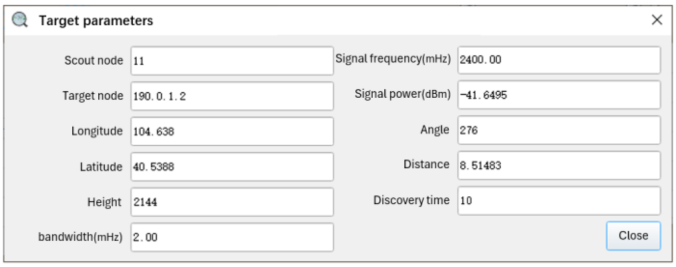
Figure 10. Scout parameters configuration

When the scout sensitivity is "-150dBm" and the antenna gain is "2dBm", the obtained target information is shown in Figure 11 and Figure 12.



	Scout equipment	Target equipment	Target IP	Discovery time
1	F1-11	F1-9	190.0.1.2	10
2	F1-11	F1-8	190.0.1.3	10
3	F1-11	F1-7	190.0.1.4	10
4	F1-11	F1-6	190.0.1.5	10
5	F1-11	F1-5	190.0.1.10	10
6	F1-11	F1-4	190.0.1.9	10
7	F1-11	F1-3	190.0.1.8	10
8	F1-11	F1-2	190.0.1.7	10
9	F1-11	F1-1	190.0.1.6	10

Figure 11. Target intelligence

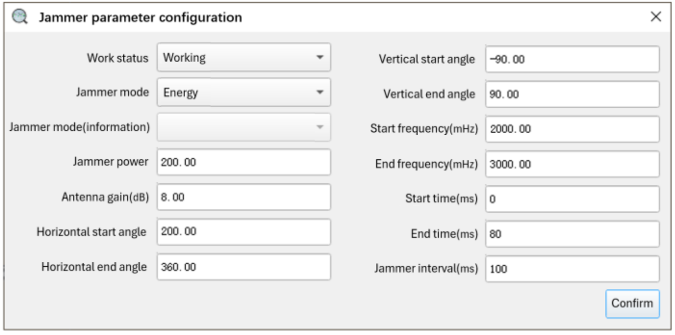


Target parameters

Scout node	11	Signal frequency(mHz)	2400. 00
Target node	190. 0. 1. 2	Signal power(dBm)	-41. 6495
Longitude	104. 638	Angle	276
Latitude	40. 5388	Distance	8. 51483
Height	2144	Discovery time	10
bandwidth(mHz)	2. 00		

Figure 12. Specific target intelligence

The parameters of the jammer node is configured according to the reconnaissance intelligence. The method is energy type, with power set to "200 dBm", antenna gain set to "8 dBm", horizontal beam direction set to "200° ~360° ", vertical beam direction set to "-90° ~90° ", start time set to "0 ms", end time set to "80 ms", and jamming period set to "100 ms". The configuration is shown in Figure 13.



Jammer parameter configuration

Work status	Working	Vertical start angle	-90. 00
Jammer mode	Energy	Vertical end angle	90. 00
Jammer mode(information)		Start frequency(mHz)	2000. 00
Jammer power	200. 00	End frequency(mHz)	3000. 00
Antenna gain(dB)	8. 00	Start time(ms)	0
Horizontal start angle	200. 00	End time(ms)	80
Horizontal end angle	360. 00	Jammer interval(ms)	100

Figure 13. Jammer parameters configuration

The number of received data packets of nodes 8, 9, and 10 is printed and shown in Figure 14. The number varies depending on the degree of interference.

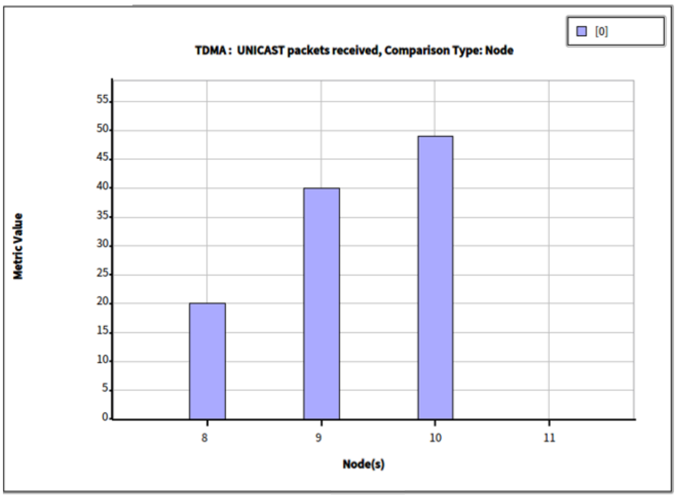
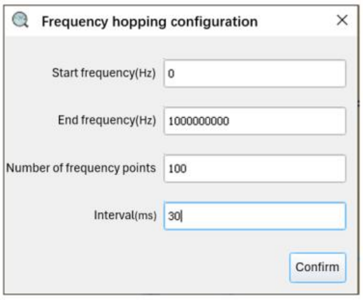


Figure 14. Number of received packets under interference

We Kept the jammer configuration unchanged, restarted the simulation and set the frequency hopping parameters in the configuration file. While the frequency hopping is enabling, the result is shown in Figure 15.

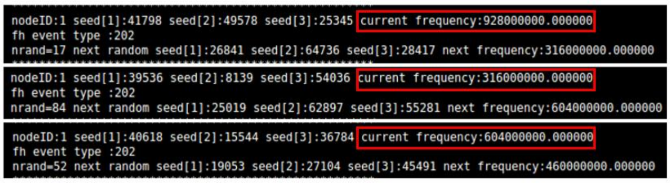


Frequency hopping configuration

Start frequency(Hz)	0
End frequency(Hz)	1000000000
Number of frequency points	100
Interval(ms)	30

Figure 15. Frequency hopping configuration

According to the printed information, it can be seen that the network communication frequency changes over time, as shown in Figure 16.



```
nodeID:1 seed[1]:41798 seed[2]:49578 seed[3]:25345 current frequency:928000000.000000
fh event type :202
nrand=17 next random seed[1]:26841 seed[2]:64736 seed[3]:28417 next frequency:316000000.000000
```

```
nodeID:1 seed[1]:39536 seed[2]:8139 seed[3]:54036 current frequency:316000000.000000
fh event type :202
nrand=84 next random seed[1]:25019 seed[2]:62897 seed[3]:55281 next frequency:604000000.000000
```

```
nodeID:1 seed[1]:40618 seed[2]:15544 seed[3]:36784 current frequency:604000000.000000
fh event type :202
nrand=52 next random seed[1]:19053 seed[2]:27104 seed[3]:45491 next frequency:460000000.000000
```

Figure 16. The changing frequency

The number of received data packets of nodes 8, 9, and 10 is printed again. It shows a significant improvement in packet reception, as shown in Figure 17.

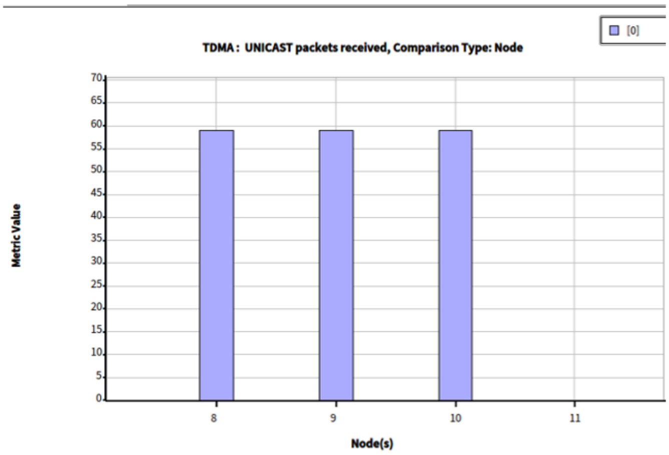


Figure 17. Number of received packets in frequency hipping mode

The experiment results show that the reconnaissance interference model can achieve reconnaissance and directional interference on the target node normally. After enabling frequency hopping, the anti-interference performance of the communication node is significantly improved.

C. Delay Test

According to the design principles, we tested the connection delay between the external control module and the QualNet network simulator to ensure that the delay generated by the data interaction interface during the simulation testing process is within an acceptable range. This delay includes the processing time of interactive software and the transmission time of data between the software interface and the simulator. To evaluate the latency of the proposed data interaction interface, 100 tests were conducted. As shown in Figure 18, the interface delay is on the millisecond level, which can be ignored compared to the transmission delay of data in the virtual simulation network. It can be seen that this interface can ensure real-time data interaction in simulation.

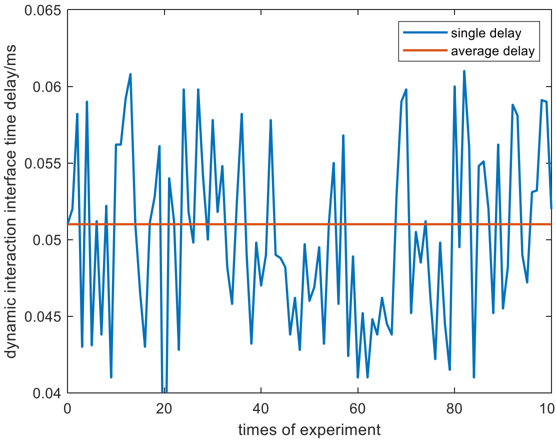


Figure 18. Interface delay

V. Conclusion

This paper introduces reconnaissance interference model and frequency hopping model in the simulation of QualNet communication network, designs the data interaction interface of the model, and tests the model's functionality and real-time performance of the interface. The test results have proven the effectiveness of the model, and provides effective model and technical support for evaluating the anti-interference performance of communication networks.

Acknowledgment

This research was funded by the National Natural Science Foundation of China Youth Fund (Grant No. 62203048) and the National Natural Science foundation of China (Grant No. 62073039). Corresponding author: Xiang Gao.

Reference

[1] Qingrui Guo, Wenbin Guo, Xuerang Guo, Qiang Zhang. Research on Power Communication Network Simulation for Energy Internet. *IEEE Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*. 2022; v 2022-June: 606-609.

[2] Xin Fang, Xiaodong Yuan, Yi Pan, Mingming Shi, Jinggang Yang, Tiankui Sun. Analysis of Communication Simulation Model of Urban Comprehensive Energy Network Based on the Internet of Things. *Journal of Physics: Conference Series*. 2023; v 2480, n 1.

[3] Xu L, Yao WX, Zhang YJ, Lu ZM, Li HL. Battlefield Network Topology Inference and Feature Analysis Based on OPNET. *Journal of Network Intelligence*. 2023;658-675.

[4] Shuai W, Han Z, Yongli Y. S. Simulation and Performance Analysis of Tactical Battlefield Communication Network. *Proceedings of 2018 IEEE 3rd International Conference on Cloud Computing and Internet of Things, CCIOT 2018*. 2018;472-478.

[5] Suliman Islam Abdelmuti, Abdalla Ghassan Mohammed Taha. Comparative Study of wireless LAN using OPNET and NS-2. *Proceedings of 2016 Conference of Basic Sciences and Engineering Studies, SGCAC 2016*. 2016; 196-200.

[6] Kim,Sungsoo, Hood, Cynthia. Impact of simulation tool on TCP performance results: A case study with ns-2 and opnet. *Simulation Series*. 2007;173-179

[7] Nawaz Haque, Ali Husnain Mansoor, Massan Shafiq-ur-Rehman. Comparative analysis of simulator tools for unmanned aerial vehicle communication networks. *International Journal of Modelling, Identification and Control*. 2021; v 39, n 4: 293-302

[8] Alsafwani Nadher, Ali Musab A. M., Tahir Nooritawati Md. Evaluation of the Mobile Ad Hoc Network (MANET) for Wormhole Attacks using Qualnet Simulator. *2021 IEEE 11th International Conference on System Engineering and Technology, ICSET 2021 – Proceedings*. 2021;46-49

[9] Huibo Li, Peng Gong, Yu Liu, Guolin Li, Dan Shan. PDSI-based Static Route Online Configuration Method for QualNet Simulator. *International Conference on Advanced Communication Technology, ICACT*. 2019; v 2019-February:325-329.

[10] Arora, N. Performance Analysis of AODV, DSR and ZRP in MANETs using QualNet Simulator. *Journal of Engineering Science and Technology Review*. 2013; v 6, n 1:21-24.

[11] Zhenjing Zhang, Zhigang Jin, Huan Chen, Yantai Shu, Chuandong Zhao. Design and implementation of a delay-tolerant network emulator based in QualNet simulator. *Proceedings - 5th International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2009*.

[12] Dingkun Huang, Xiaopeng Yan, Jian Dai, Xinwei Wang, Yangtian Liu. Cognitive interference decision method for air defense missile fuze based on reinforcement learning. *Defence Technology*. 2023 Article in Press.

[13] Xiao Zhang, Xiaofeng Xi. LTE signal based passive bistatic radar co-channel interference suppression method. *Proceedings of SPIE - The International Society for Optical Engineering*. 2023; v 12703.

[14] Tianyang Pang, Yonggui Li, Yingtao Niu, Chen Han, Zhi Xia. Classification and Development of Communication Electronic Interference. *Communication Technology*. 2018; 51(10):2271-2278.

[15] Yong Chen, Xinyu Yao, Yulin Pan, Xiaofeng Tang. Design and application of multi-level distributed real-time simulation platform. *Proceedings - 2011 International Conference on Intelligence Science and Information Engineering, ISIE 2011*.2011;221-225.

[16] Brito Alisson V. , Costa Luis Feliphe S., Bucher Harald, Sander Oliver, Becker Juergen, Oliveira Helder, Melcher Elmar U.K. A distributed simulation platform using HLA for complex embedded systems design. *Proceedings - 2015 IEEE/ACM 19th International Symposium on Distributed Simulation and Real Time Applications, DS-RT 2015*.2016; 195-202.

[17] Rusheng Ju, Lijuan Huang, Jian Huang, Kedi Huang. Real-time evaluation of HLA-based simulation systems. *ICCMS 2010 - 2010 International Conference on Computer Modeling and Simulation*. 2010; v 2:421-424.

[18] Zhenyu Ma, Qiang Luo, Junmin Zhao, Yuyang Huang. Design and Implementation of Intelligent Cluster Simulation System Based on DDS. *Lecture Notes in Electrical Engineering*. 2023; v 1010 LNEE:2088-2096.

[19] Kim D, Oh HS, Hwang SW. A DDS-based distributed simulation for anti-Air missile systems. *SIMULTECH 2016 - Proceedings of the 6th International Conference on Simulation and Modeling Methodologies, Technologies and Applications*. 2016;270-276.

[20] Changqing Li, Chi Zhou, Shibing Zhu. Design of QoS protocol matching method in the DDS specification. *Applied Mechanics and Materials*. 2014; v 577:898-902.

[21] Hakiri Akram, Berthou Pascal, Slim Abdellatif Slim, Diaz Michel, Gayraud Thierry. Supporting end-to-end internet QoS for DDS-based large-scale distributed simulation. *SIGSIM-PADS 2013 - Proceedings of the 2013 ACM SIGSIM Principles of Advanced Discrete Simulation*. 2013;397-402.

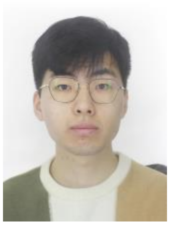
Wenyi Li received the BS degree in Mechatronical Engineering from Beijing Institute of Technology in 2021, and now he is a MS candidate in School of Mechantronical Engineering, Beijing Institute of Technology. His research interests include wireless network simulation and emulation, wireless communication and so on.

Peng Gong received the BS degree in Mechatronical Engineering from Beijing Institute of Technology, Beijing, China, in 2004, and the MS and Ph.D. degrees from the Inha University, Korea, in 2006 and 2010, respectively. In July 2010, he joined the School of Mechatronical Engineering, Beijing Institute of Technology, China. His research interests include link/system level performance evaluation and radio resource management in wireless systems, information security, and the next generation wireless systems such as 3GPP LTE, UWB, MIMO, Cognitive radio and so on.

Weidong Wang received his B.S degree in electronic and communication engineering from Harbin Institute of Technology, Harbin, China in 2001. And he received the M.S degree in Information Technology & Telecommunications from Inha University, Incheon, South Korea in 2005. He worked for Huawei Technologies, ZTE Corporation and Samsung electronics. He is currently the GM of Wuxi Junction Information Technology Incorporation Company.

Yu Liu received the BS degree from National University of Defense Technology in June 2006 and the MS degree from Military Economy College of the CPCA in April 2009. He is currently working toward the Ph.D. degree with the School of Mechantronical Engineering, Beijing Institute of Technology. His research direction is Communication engineering.

Jianfeng Li received the BS degree in Computer Science and Technology from Equipment Command and Technology Academy, in 2001, the MS degrees in Military Equipment Science from Equipment Command and Technology Academy,, in 2004. His research interests include equipment operation test, in-service assessment, test data management and so on.



Xiang Gao received the B.S. degree, the MS degree and the Ph.D. degree in school of Mechatronical Engineering from Beijing Institute in 2014, 2016, and 2021, respectively. Now he is a postdoctor in Beijing Institute of Technology. His research interests include network simulation and emulation, NOMA, VLC, D2D communications, IoT, cryptography and so on.

Location based Data-centric Forwarding for Mobile Ad-hoc Networks

Hieu Nguyen, Ilkyeun Ra
Department of Computer Science & Eng.
University of Colorado Denver
Denver, Colorado, USA
hieu.nguyen@ucdenver.edu, ilkyeun.ra@ucdenver.edu

***Abstract*—With the ever-increasing usage and deployment of mobile devices, mobile ad-hoc networks have become more and more prevalent in replacing centralized networks. However, high mobility of nodes can lead to challenging performance issues, such as high packet loss, frequent path failures causing high route-reinitiating overheads, and significant data retrieval delay. Named Data Networking (NDN) offers an alternative solution to these problems. But, existing approaches still have many issues and can be further improved. This paper proposes a novel location-based approach that aims to address existing challenges of using NDN in mobile ad hoc networks and improve the performance of existing works. The simulation results presented in the paper prove that our approach is very much feasible.**

***Index Terms*—forwarding strategy, MANET, named data networking, network simulations**

I. Introduction

Named Data Networking (NDN) [1] is an information-centric Internet architecture designed to tackle various unsolved challenges in the existing TCP/IP Internet. Many challenges can be found during network deployment for ad-hoc wireless networks, such as involving various communication protocols, insufficient infrastructure routing support, and deteriorated performance due to an unstable network environment.

While existing NDN forwarding techniques can partially address these issues, there is still a lot of room for improvement. Broadcasting, while being an effective mechanism for NDN in an ad-hoc wireless network environment, can cause strain on a limited-bandwidth network. Adaptive methods may help with limiting bandwidth usage caused by broadcasting. However, they can incur significant overhead due to complicated set-up and route reconfiguration, especially for volatile and mobile networks.

This work aims to improve the performance of current NDN based data forwarding protocols in a mobile ad-hoc environment by using neighboring nodes' location information to determine the most reliable path to forward Interest packet. To demonstrate this, we conducted comprehensive simulations to compare the packet delivery performance of our proposed approach with the existing broadcast-based NDN forwarding protocol.

The rest of this paper is organized as following. Section II presents other related researches. The design of the proposed system is detailed in section III. In section IV, our approach is evaluated and compared with existing solutions. Lastly, section V summarizes our contributions and outlines future works.

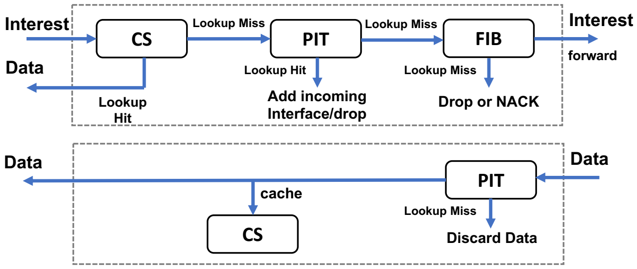


Fig. 1: Normal Forwarding pipeline in a NDN node

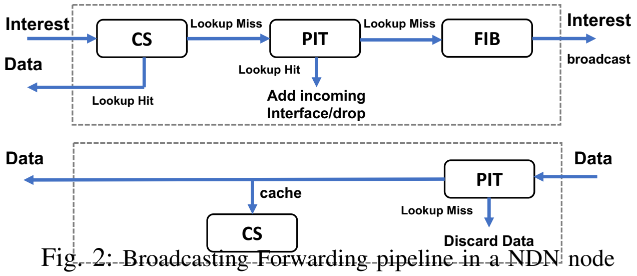


Fig. 2: Broadcasting Forwarding pipeline in a NDN node

II. Related Work

Originally , NDN utilizes the reverse paths between Consumer and Producer to deliver Data [2]. However, in a mobile ad-hoc environment, these reverse paths can be easily broken down when consumer nodes start moving away from their original position . While NDN has a natural advantage in dealing with this issue via in-network caching (to shorten the distance between consumers and data) [3] [4], it raises other issues as well such as which data should be cached/dropped, considerable data retrieval delays due to having the Interests going all the way back to the data source or reestablishing a new Data path if the requested Data is not cached anywhere [5]. Also, this does not fully solve the mobility issue.

The most common NDN forwarding strategy for mobile ad-hoc network is broadcasting /flooding . An NDN broadcast-based forwarding strategy follows the general NDN architecture [1]. An NDN consumer sends an Interest with a name when it wants to retrieve data. Upon receiving the Interest, a node will look at its Content Store (CS) and replies with data on a name match. Otherwise , the Interest is passed to the Pending Interest Table (PIT). If there is an entry with the same name in the PIT, the Interest is discarded, and the entry is updated to include the new request. If no entry in the PIT matches the Interest's name, a new entry is created with the incoming interface and passes the Interest to its Forwarding Information Base (FIB). Next, the node performs the longest common prefix match to existing FIB prefix entries. With the general NDN architecture, the node forwards the Interest through the associated interface(s) to the next hop node(s) if a match is found. If a match is not found, it either drops the Interest or sends a NACK (negative acknowledgment) back. This will inform the other nodes that a route through this node is unavailable. However, this is not true for a broadcast-based (or flooding) NDN strategy. The FIB hit/miss status is ignored in this case; all Interests are broadcast after passing through the PIT. The rationale behind this is that it can potentially increase data retrieval possibility through multiple possible paths between a consumer and data node(s). This scenario also has no NACK, as every packet transmission is broadcast. When a node receives data, it will first check the PIT. If a match is found, the data is passed to CS; otherwise, the data is dropped. After updating CS, the data is forwarded to the associated node based on the existing PIT entry. The PIT entry is then clear afterward. The main disadvantage of broadcasting/flooding in NDN over mobile ad-hoc network is that it can lead to a high possibility of collision and channel contention in the network due to high number of Interest transmissions (and retransmissions caused by nodes' mobility).

To minimize the negative effect of broadcasting Shi *et al.* proposed NDN self-learning [6]. Similar to IP-AODV [7], a consumer will first broadcast discovery Interests until information about the next hop toward data is learned. When it receives the data which travels as unicast, the consumer will switch to unicast Interests to a next-hop in the determined path (via the discovery Interests). This approach, however will not work very well for mobile ad-hoc networks as only the consumer node makes discovery decisions. Also, it can suffer from many application-level timeouts and is not very reactive to mobility within the network. [8] - [9] also discuss NDN routing and forwarding protocols for mobile ad-hoc networks. The performance of these approaches, however, can be further improved.

Another approach to reduce the number of Interest transmission/retransmission caused by broadcasting/flooding is via controlling the forwarding rate of each participating nodes. NAIF (Neighborhood-Aware Interest Forwarding) [10] uses data retrieval statistics to decide whether to lower or increase the Interest forwarding rate. However, this can incur high overheads and will not react well to sudden changes in network. [11] uses a counter-based suppression mechanism to reduce the forwarding rate. Before forwarding Data/Interest, a random time interval is introduced. During this time period, if the node notices that the same packet is being transmitted by another node then the packet will be dropped.

[12] uses geographic positions of producer nodes and the remaining energy of neighboring nodes to determine which node to forward Interest packets to. The probability of a neighboring node being chosen to forward Interest packets to is calculated by balancing its distance to the producer node and its residual energy. Because of the focus on limiting energy consumption, this does not guarantee optimal route to producers. Also, this approach does not consider the mobility of the nodes in the network.

III. The Proposed Framework

This section will list the challenges of designing an NDN framework for mobile ad-hoc networks and detail how we plan to address them with our proposed system.

A. Challenges

The environment of edge networks is usually very unstable, which could lead to packets lost, disruption in the communication link, and dynamically changed topology. To address this, existing approaches [6] [13] choose to broadcast Interest as the core mechanism as it is very simple and effective to discover content with broadcasting in NDN. NDN is able to detect and drop looped Interests at the forwarding plane - an NDN node can safely forward Interests to any face in any topology without worrying about creating loops. However, Interest broadcasting takes up a significant amount of network and device resources. Another issue with this is that different types of communication protocols may have different route-finding performances via broadcasting. For example, Ethernet and UDP communication interfaces perform significantly worse than a unicast interface in a Wi-fi network.

Lastly, many devices involved in edge networks have mobility and limited resources, so it is difficult for them to assume the roles of forwarders and have the capability to perform in-network caching.

B. System Architecture

To address the impact of nodes' mobility in mobile ad-hoc networks, we take into account the position of the neighboring nodes when choosing which hop to forward Interests to. We assume that nodes are equipped with a global position system (GPS) and connect to each other using an omnidirectional antenna. Each node will regularly send discovery Interests with its current position to its neighbor nodes. Based on this information, mobile nodes can compute the speed and distance with respect to the neighbor nodes and will use this information to determine which path to forward the Interests.

1) Nodes Distance Calculation: Mobile nodes' location, speed, and direction are changed dynamically in MANETs. We calculate the Euclidean distance between two neighbor nodes to determine whether the nodes are moving toward each other or away from each other. This information will then be used to rate the next hop to forward the Interests to.

Suppose there are two mobile nodes n_1 and n_2 with a transmission range of r. Nodes n_1 and n_2 are moving at the speed of V_{n_1} and V_{n_2}, respectively.

The distance between n_1 and n_2 - D_{n_1,n_2} at a time t can be calculated as:

$$D_{n_1,n_2}(t) = \sqrt{(X_{n_2(t)} - X_{n_1(t)})^2 + (Y_{n_2(t)} - Y_{n_1(t)})^2}, \quad (1)$$

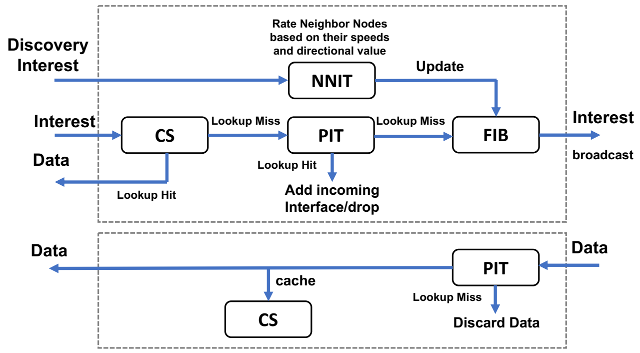


Fig. 3: NDN pipeline for our proposed system with the added Neighbor Nodes Information Table

where $(X_{n_1(t)}, Y_{n_1(t)})$ and $(X_{n_2(t)}, Y_{n_2(t)})$ are the locations of n_1 and n_2 at a time t, respectively.

At $(t + \Delta t)$ time, the distance between n_1 and n_2, $D_{n_1,n_2}(t + \Delta t)$ can be calculated as:

$$D_{n_1,n_2}(t+\Delta t) \\ = \sqrt{(X_{n_2(t+\Delta t)} - X_{n_1(t+\Delta t)})^2 + (Y_{n_2(t+\Delta t)} - Y_{n_1(t+\Delta t)})^2}, \tag{2}$$

If we define mobility as the average change in distance between all nodes over that period of time, then it can be expressed as a function of speed and movement pattern. The speed of a node at a time t can be calculated as follows:

$$V(t, t+\Delta t) = \frac{|(X_2 - X_1) + (Y_2 - Y_1)|}{(t + \Delta t) - t} \tag{3}$$

By these formulas, each node can compute its speed and distance from its neighbors at any given time.

After determining the distance between n_1 and its neighbors at t and $(t + \Delta t)$, we can determine whether n_1 is heading toward or away from its neighbors. If $D_{n_1,n_2}(t)$ is larger than $D_{n_1,n_2}(t+\Delta t)$, then nodes n_1 and n_2 are closer to each other within the time interval between t and $(t + \Delta t)$. Hence, the two nodes are heading toward each other for this interval. On the other hand, if $D_{n_1,n_2}(t)$ is less than $D_{n_1,n_2}(t + \Delta t)$, the nodes are moving away from each other and the nodes have a high probability of being disconnected.

2) Neighbor Nodes Information Table - NNIT: When a node receives discovery Interests, it will store the sending node's information in a separate table. This table contains two fields: the neighbor nodes' speed and the neighbor nodes' direction. The direction value is calculated using equation 2 and 3. If the nodes are heading toward each other and the distance between the two nodes is constant, the direction value is set to one. Otherwise, the direction value is set to 0.

Each of the neighbor nodes in the table will then be rated based on these two values. Suppose a neighbor node has a direction value of 1 and a lower speed. In that case, it will be ranked higher as a connection with a node that moves slower and heads toward or in the same direction will have a higher probability of lasting longer.

IV. Evaluation

This section analyzes the performance of our approach in comparison with NDN flooding. We use packet delivery ratio, end-to-end delay, and response time as performance metrics.

1) *Packet Delivery Ratio*: The ratio of the number of data packets successfully received at the consumer application with respect to the number of unique Interests sent.
2) *End-to-end Delay*: The average round-trip time measured at the consumer applications.
3) *Response Time*: The average time it takes for a consumer application to receive the first data packet with respect to the Interest sent.

A. Experimental Set-up

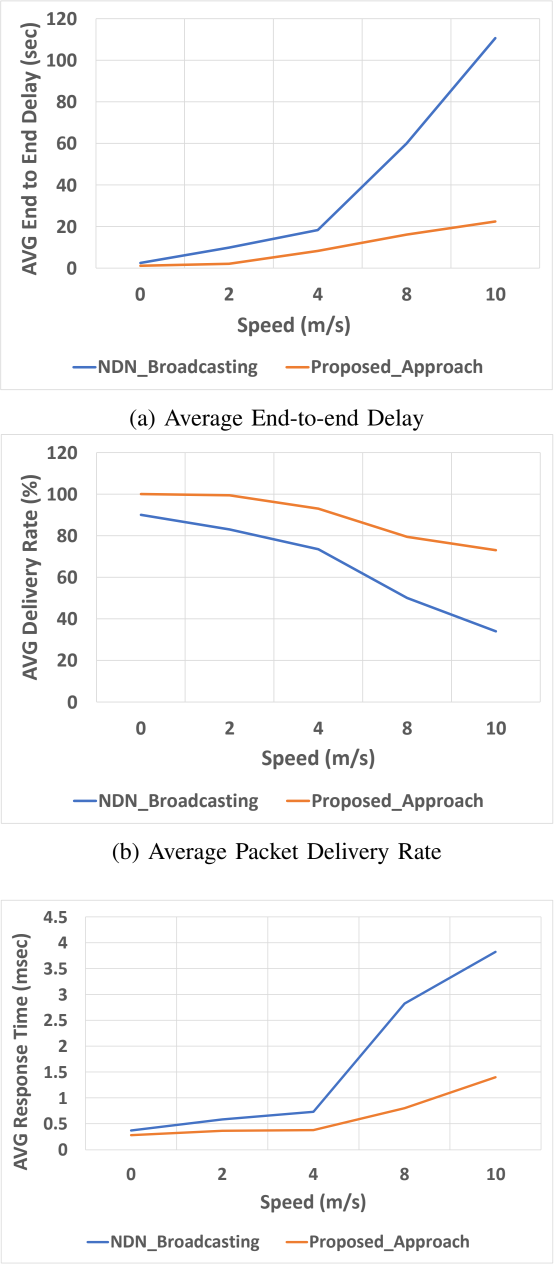
To verify the theoretical results, we devise a series of simulations using ndnSim simulator [14]. For our simulations, we place the nodes randomly within an area of 200m x 250m for 20 nodes network. Each node has a fixed transmission range of 50 meters in diameter. Each node has only one wireless interface communicating over a single channel of IEEE 802.11a at 24Mbps. The system consists of 4 consumers and 2 producers, with the other nodes serving as forwarders.

Each consumer will request 50 different data packets of 512 bytes, each using a constant bit-rate application. The consumers will start randomly during the 5 seconds period after the simulation starts. We apply the random waypoint mobility model for the nodes' mobility. The speed of the nodes is set to be $0m/s$ (stationary) first, then changed to be $2m/s$, $4m/s$, $8m/s$, and $10m/s$ for each scenario.

B. Results

Figure 4 shows the result of experiments for our proposed system and NDN broadcasting in a 20 nodes network. At lower speeds (stationary to $4m/s$), our approach performs better than NDN broadcasting across all metrics. However, it is not by a more significant margin, especially for the end-to-end delay and response time. The result of the packet delivery rate, however, clearly shows the disadvantage of the broadcasting approach. For the scenarios in which the nodes are either stationary or moving at lower speed, if the nodes are close together, broadcasting/flooding interest packets can lead to high frequency of packet collision in the network which in turn will result in lower delivery rate. When the nodes start moving faster, if the nodes are moving away from each other, it can help with the delivery rate of the broadcasting strategy as there are more spaces between each node (which can lead to less packet collision). However, this will come at the cost of introducing more traffics to the network.

When we change the mobile nodes' moving speed from $4m/s$ to $8m/s$, NDN broadcasting suffers a sharp increase in both response time and end-to-end delay while the change of speed has a much lesser effect on our approach's performance. The increase in response time and end-to-end delay in our approach is due to the slight overheads incurred by re-calculating the neighbor information when neighboring nodes start moving away from their original positions.



(a) Average End-to-end Delay

(b) Average Packet Delivery Rate

(c) Average Response Time

Fig. 4: Effect of mobility in a 20 nodes network for the proposed work and NDN broadcasting

V. Conclusion

This paper presented a novel location-based NDN forwarding scheme for mobile ad-hoc networks. As shown in the paper, we proved that our approach could reduce network latency while improving data retrieval comparing to existing works. Due to the limited scope of the experiments, we cannot verify the scalability of the framework. As part of our future work, we plan to address this by evaluating our approach with real-life and larger-scale network.

References

[1] L. Zhang, A. Afanasyev, J. Burke, V. Jacobson, K. Claffy, P. Crowley, C. Papadopoulos, L. Wang, and B. Zhang, "Named data networking," *ACM Computer Communication Reviews*, Jun. 2014.

[2] C. Yi, A. Afanasyev, L. Wang, B. Zhang, and L. Zhang, "Adaptive forwarding in named data networking," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 3, pp. 62–67, 2012.

[3] Y.Zhang, A. Afanasyev, J. Burke, and L. Zhang, "A survey of mobility support in named data networking," in *Proc. IEEE Conf. Comput. Commun. Workshops*, (San Francisco, CA, USA), pp. 83–88, 2016.

[4] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. I. Braynard, "Networking named content," in *Proc. 5th Int. Conf. Emerg. Netw. Experiments Technol.*, pp. 1–12, 2009.

[5] B. Feng, H. Zhou, and Q. Xu, "Mobility support in named data networking: A survey," *J. EURASIP J. Wireless Commun. Netw.*, vol. 2016, no. 220, 2016.

[6] J. Shi, E. Newberry, and B. Zhang, "On broadcast-based self-learning in named data networking," in *2017 IFIP Networking Conference (IFIP Networking) and Workshops*, pp. 1–9, June 2017.

[7] C. E. Perkins and E. M. Royer, "Adhoc on-demand distance vector routing," in *Proceedings of the Second IEEE Workshop on Mobile Computer Systems and Applications*, (Washington, DC, USA), pp. 90–, 1999.

[8] B. Etefia and L. Zhang, "Named data networking for military communication systems," in *2012 IEEE Aerospace Conference*, pp. 1–7, 2012.

[9] H. Khelifi, S. Luo, B. Nour, H. Moungla, Y. Faheem, R. Hussain, and A. Ksentini, "Named data networking in vehicular ad hoc networks: State-of-the-art and challenges," *IEEE Communications Surveys Tutorials*, 2019.

[10] Y. Yu, R. Dilmaghani, S. Calo, M. Y. Sanadidi, and M. Gerla, "Interest propagation in named data manets," in *2013 International Conference on Computing, Networking and Communications (ICNC)*, pp. 1118–1122, 2013.

[11] M. Amadeo, A. Molinaro, and G. Ruggeri, "E-chanet: Routing, forwarding and transport in information-centric multihop wireless networks," *Computer Communications*, vol. 36, no. 7, pp. 792–803, 2013.

[12] A. Aboud and H. Touati, "Geographic interest forwarding in ndn-based wireless sensor networks," in *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, pp. 1–8, 2016.

[13] M. A. Rahman and B. Zhang, "On data-centric forwarding in mobile ad-hoc networks: Baseline design and simulation analysis," in *2021 International Conference on Computer Communications and Networks (ICCCN)*, 2021.

[14] S. Mastorakis, A. Afanasyev, and L. Zhang, "On the evolution of ndnsim: an open source simulator for ndn experimentation," *ACM Computer Communication Review*, Jul. 2017.

HIEU NGUYEN received a BS degree in telecommunication engineering from Hanoi University of Science and Technology, Vietnam in 2013. He is currently working toward a Ph.D. degree at the DECENT Lab, Department of Computer Science and Engineering, University of Colorado Denver, Colorado, USA. His main research interests include computer network and high-speed communication system utilizing SDN, NFV and NDN technologies.

ILKYEUN RA received a Ph.D. degree in Computer and Information Science from Syracuse University, USA, MS degree in Computer Science from University of Colorado Boulder, Colorado, USA, and BS degree and MS degree in Computer Science from Sogang University, Seoul, Korea. Currently, he is an associate professor in the Department of Computer Science and Engineering at the University of Colorado Denver. His main research interests include computer network, cloud computing, and high-performance computing.

Optimization of Downlink Power Allocation in NOMA-OTFS based Cross-Domain Vehicular Networks

Hao Xu[1], Zhiquan Bai[1*], Jinqiu Zhao[1], Dejie Ma[1], Bangwei He[1] and KyungSup Kwak[2]
[1]Shandong Provincial Key Lab. of Wireless Communication Technologies,
School of Information Science and Engineering, Shandong University, Qingdao 266237, China
[2]Department of Information and Communication Engineering, INHA University, Incheon 22212, Korea
xhxhn999@163.com, zqbai@sdu.edu.cn*, 202020373@mail.sdu.edu.cn, madj0212@163.com, hbw017@ mail.sdu.edu.cn, kskwak@inha.ac.kr

***Abstract*—Orthogonal time frequency space (OTFS) and non-orthogonal multiple access (NOMA) are pivotal for enhancing the transmission performance of vehicular communications. This paper delves into the downlink power allocation of a NOMA-OTFS system with a frequency domain linear equalizer (FD-LE), where a high-speed user in delay-Doppler domain and multiple low-speed time-frequency domain NOMA users coexist. Considering the fairness of NOMA users, we optimize the minimum rate of the low-speed users, constrained by the quality of service (QoS) of the high-speed user. To address this problem, we propose an iterative power allocation optimization (IP-AO) strategy and obtain an accurate optimal solution based on the auxiliary variables by transforming the original non-convex problem into a convex one. Moreover, we derive a closed-form solution for the optimal power allocation (OPA). Simulation results validate the superiority of our schemes over traditional power allocation methods in maximizing the minimum user rate in the NOMA-OTFS vehicular system.**

***Keywords*— OTFS, NOMA, Power allocation, IP-AO, OPA.**

I. INTRODUCTION

In the 5G landscape, vehicle-to-everything (V2X) has elevated the intelligent transportation system (ITS) with optimized data rate and reduced latency [1], paving the way for autonomous driving. In the V2X scenarios, diverse mobile users converge on a unified wireless infrastructure. To meet the stringent communication demands of V2X, many studies intensively focus on refining communication protocols and advanced multi-access techniques [2].

Orthogonal time frequency space (OTFS) as an innovative modulation technique, was specifically designed to improve the reliability and robustness of wireless communications characterized by high mobility and delay spread [3][4]. It operates by mapping input data in the delay-Doppler (DD) domain into a two-dimensional time-frequency (TF) plane. The mapping approach enables OTFS to effectively mitigate the time-varying Doppler effects caused by the high mobility and multipath propagation [5][6]. Meanwhile, massive multiple-input and multiple-output OTFS (MIMO-OTFS) systems have been studied to achieve high spectral and energy efficiency with low complexity downlink multi-user precoding [7]. Non-orthogonal multiple access (NOMA) is a unique multi-access, enabling concurrent communication of multiple users on the same TF resource [8]. Compared with the traditional orthogonal multiple access (OMA), NOMA employs power-domain multiplexing and facilitates user differentiation on a singular resource block [9]. This capability makes NOMA achieve superior spectral efficiency and system capacity in dense user scenarios like V2X. In [10], a novel NOMA approach was presented for heterogeneous mobile communications by integrating OFDM and OTFS for users in the TF and DD domains. [11] employed the sparse code multiple access (SCMA) for low-mobility users to improve the bit error rate. H. Cheng *et al.* introduced an optimal power allocation strategy to enhance the system throughput with inherent fairness [12]. A power allocation strategy was presented in [13] to maximize the system's energy efficiency. In [14], the beamforming design for the OTFS-NOMA system was explored for significant performance gains, considering the imprecise channel state information. However, current literature inadequately addresses fair rate allocation in NOMA and its integration with OTFS under high-speed scenarios, with prevailing power allocation strategies remaining highly complex.

In this paper, we present a NOMA-OTFS scheme with a high-speed DD domain user and multiple low-speed users in the TF domain. A frequency domain linear equalizer (FD-LE) is employed to resist the inter-symbol interference in the DD plane. We first derive the signal-to-interference-noise ratio (SINR) of different system users and formulate an optimization problem to maximize the minimum bit rate of NOMA users by ensuring the quality of service (QoS) of the high-speed user and the total power for rate fairness among NOMA users. Given the non-convex nature of the original problem, we propose an iterative power allocation optimization (IP-AO) strategy and utilize auxiliary variables to transform it into a convex problem. We further present a closed-form solution for the optimal power allocation (OPA) with lower complexity. With simulation results, we show the superiority of the proposed power allocation strategies over the traditional methods in maximizing the minimum NOMA user rate of the NOMA-OTFS system.

II. SYSTEM MODEL

A. NOMA-OTFS Resource Allocation

This work was supported in part by the Key Project of National Quality Infrastructure System of National Key Research and Development Plan under Grant 2022YFF0604900, in part by the Innovation and Development Joint Foundation of Shandong Provincial Natural Science Foundation under Grant ZR2021LZH003, in part by the National Research Foundation of Korea-Grant Funded by the Korean Government (MSIT-NRF-2020R1A2B5B02002478). *Corresponding author, and in part by the National Defense Preliminary Research Project under Grant 90903030202. (Corresponding author: Zhiquan Bai.)

We focus on the downlink of the NOMA-OTFS vehicular communication system in this paper, where a base station (BS) serves a DD domain high-speed user with OTFS modulation and multiple TF domain low-speed users in a manner similar to traditional OFDM as depicted in Fig. 1.

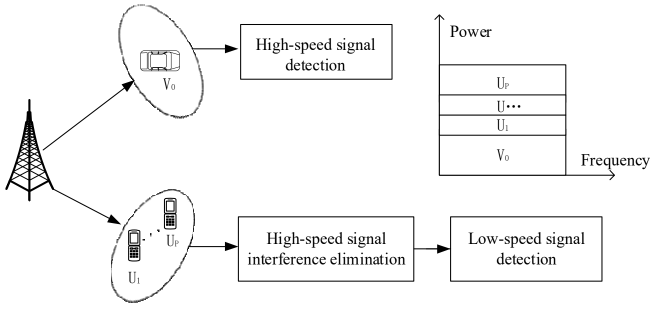

V0
High-speed signal detection
Power
UP
U···
U1
V0
Frequency
UP
U1
High-speed signal interference elimination
Low-speed signal detection

Fig 1. NOMA-OTFS V2X Downlink Model

The presented NOMA-OTFS system can flexibly utilize both the TF and DD domains. Every OTFS frame has a time duration of NT (s) and a bandwidth of $M\Delta f$ (Hz). To ensure symbol orthogonality, T and Δf are chosen to be larger than the maximum delay extension and Doppler shift, respectively. Assuming a single BS communicates with (P+1) users, including the high-speed user V_0 and the low-speed users U_i, $i \in \{1,2,\ldots,P\}$, we can get a NOMA user group. Considering the multipath transmission, the DD domain channel response for each user is represented as

$$h_i(\tau,\nu)=\sum_{t=1}^{P_i} h_{i,t}\delta\left(\tau-\tau_{i,t}\right)\delta\left(\nu-\nu_{i,t}\right), \tag{1}$$

where P_i denotes the number of multipaths for user $i, i\in\{0,1,\ldots,P\}$. The channel gain, delay, and Doppler shift of the *t-th* path for user i are denoted by $h_{i,t}$, $\tau_{i,t}$, and $\nu_{i,t}$, respectively. All the channels are independent and identically distributed (i.i.d). For the *i-th* user, the delay and Doppler taps corresponding to the *t-th* path are written as $\tau_{i,t}=l_{\tau_{i,t}}\big/M\Delta f$ and $\nu_{i,t}=k_{\nu_{i,t}}\big/NT$, respectively. $l_{\tau_{i,t}}$ and $k_{\nu_{i,t}}$ represent the delay and Doppler index of the *i-th* user's *t-th* path, respectively.

B. Cross Domain Signal Transmission

The BS transmits MN symbols to V_0 operated in the DD domain and these signals are represented by $x_0^{HS}[k,l]$ with $0\le k\le N-1$ and $0\le l\le M-1$. By inverse symplectic finite Fourier transform (ISFFT), the signal of V_0 in the DD domain can be transformed into a TF domain signal $X_0^{HS}[n,m]$ as

$$X_0^{HS}[n,m]=\sum_{k=0}^{M-1}\sum_{l=0}^{N-1}x_0^{H}[k,l]e^{j2\pi\left(\frac{km}{M}-\frac{nl}{N}\right)}\quad, \tag{2}$$

where $0\le n\le N-1, 0\le m\le M-1$.

Due to the bandwidth constraints, K (K⩽M) out of P low-speed users work take NOMA scheme and the signals of the NOMA users are processed directly on the TF domain. Each NOMA user occupies a sub-bandwidth channel and receives N information-bearing symbols. The TF domain signal for the *i-th* NOMA user is denoted as $X_i^L[n.m]$ with the following mapping scheme

$$X_i^L[n,m]=\begin{cases}x_i^L(n) & \text{if } m=i-1\\ 0 & \text{otherwise}\end{cases}. \tag{3}$$

In NOMA transmission, users share the same spectrum. The received signal for each user in the TF domain is

$$Y_i[n,m]=H_i[n,m]X_a[n,m]+Z_i[n,m]\ . \tag{4}$$

where $X_a[n,m]$ is after multi-user signals superposition and $V_i[n,m]$ is additive Gaussian white noise (AWGN), $H_i[n,m]$ is the channel gain of the *i-th* user. Here, we use the ideal pulse shaping waveform that satisfies bio-orthogonality [5].

The BS sends the superimposed signal $X_a[n,m]$ after the Heisenberg transformation. After transmission, the TF domain received signal of the *i-th* user is expressed as

$$\begin{aligned}Y_i^L[n,m]=&\underbrace{\sqrt{\beta_0}H_i[n,m]\left(\sum_{k=0}^{M-1}\sum_{l=0}^{N-1}x_0^{HS}[k,l]e^{j2\pi\left(\frac{km}{M}-\frac{nl}{N}\right)}\right)}_{\text{High speed user interference}}\\&+\underbrace{\sqrt{\beta_i}H_i[n,m]X_i^L[n,m]}_{\text{Effective signal}}+\underbrace{H_i[n,m]\sum_{i'=1,i'\neq i}^{K}\sqrt{\beta_{i'}}X_{i'}^L[n,m]}_{\text{Interference from other low-speed users}}+\underbrace{Z_i[n,m]}_{\text{noise}},\end{aligned} \tag{5}$$

where β_0, β_i is the power allocation of V_0 and U_i, respectively.

The signal of V_0 has significant intensity in the DD domain, which should be demodulated first, and the NOMA users' signals can be treated as the noise through serial interference cancellation (SIC). The received signal of user V_0 in the DD domain is written as

$$\begin{aligned}y_0^{HS}[k,l]&=\mathrm{SFFT}^{-1}\left(Y_0^{HS}[n,m]\right)\\&=\frac{1}{NM}\sum_{i=1}^{K}\sqrt{\beta_i}\sum_{k'=0}^{N-1}\sum_{l'=0}^{M-1}x_i^L[k',l']h_{w,0}\left(\frac{k-k'}{NT},\frac{l-l'}{M\Delta f}\right)\\&+\frac{1}{NM}\sqrt{\beta_0}\sum_{k'=0}^{N-1}\sum_{l'=0}^{M-1}x_0^{HS}[k',l']h_{w,0}\left(\frac{k-k'}{NT},\frac{l-l'}{M\Delta f}\right)+z_0[k,l],\end{aligned} \tag{6}$$

where $x_i^L[k',l'], 1\le i\le M$ is the delay-Doppler representation of $X_i^L[n,m]$ after symplectic finite Fourier transform (SFFT). The representation of channel $h_{w,0}(\nu,\tau)$ for V_0 is shown in [5].

The received signal of the high-speed user is expressed as

$$y_0^{HS}[k,l]=\sum_{i=0}^{K}\sqrt{\beta}_i\sum_{t=0}^{P_0}h_{0,t}x_i\left[\left(k-k_{\nu_{0,t}}\right)_N,\left(l-l_{\tau_{0,t}}\right)_M\right]+z_0[k,l], \tag{7}$$

where $(\cdot)_X$ is mod X operation. We define $\mathbf{y}_{0,k}^{HS}=\left[y_0^{HS}[k,0]\cdots y_0^{HS}[k,M-1]\right]^T$ and $\mathbf{y}_0^{HS}=\left[\mathbf{y}_{0,0}^T\cdots\mathbf{y}_{0,N-1}^T\right]^T$.

Similarly, $\mathbf{x}_i$ and $\mathbf{z}_i$, $0\le i\le K$, are the matrix version of $x_i[k,l]$ and $z_0[k,l]$. The received signal vector of V_0 is

$$\mathbf{y}_0^{HS}=\sqrt{\beta_0}\mathbf{H}_0^{\mathrm{dd}}\mathbf{x}_0^{HS}+\sum_{i=1}^{K}\sqrt{\beta_i}\mathbf{H}_0^{\mathrm{dd}}\mathbf{x}_i^L+\mathbf{z}_0\ , \tag{8}$$

where $\mathbf{H}_0^{\mathrm{dd}}$ is the channel gain vector expression as

$$\mathbf{H}_0^{\mathrm{dd}} = \begin{bmatrix} \mathbf{G}_{0,0} & \mathbf{G}_{0,N-1} & \cdots & \mathbf{G}_{0,1} \\ \mathbf{G}_{0,1} & \mathbf{G}_{0,0} & \ddots & \mathbf{G}_{0,2} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{G}_{0,N-1} & \mathbf{G}_{0,N-2} & \cdots & \mathbf{G}_{0,0} \end{bmatrix}_{MN\times MN} . \quad (9)$$

Then, the FD-LE is taken to mitigates inter-symbol interference. Its implementation includes the following two steps:

Step1: Detect high-speed user's signal and substitute the detection matrix $\mathbf{F}_N \otimes \mathbf{F}_M^H$ to (9) as

$$\hat{\mathbf{y}}_0^{HS} = \mathbf{E}_0\left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)\left(\sqrt{\beta_0}\mathbf{x}_0^{HS} + \sum_{i=1}^{K}\sqrt{\beta_i}\mathbf{x}_i^L\right) + \hat{\mathbf{z}}_0, \quad (10)$$

with $\hat{\mathbf{y}}_0^{HS} = \left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)\mathbf{y}_0^{HS}$ and $\hat{\mathbf{z}}_0 = \left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)\mathbf{z}_0$. $\mathbf{E}_0$ is a diagonal matrix which $(kM+l+1)-th$ main diagonal element is

$$E_0^{k,l} = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} a_{0,n}^{m,1} e^{j2\pi(\frac{lm}{M}-\frac{kn}{N})}, \quad (11)$$

where $a_{0,n}^{m,1}$ is the element of the $(nM+m+1)-th$ row and the first column of $\mathbf{H}_0^{\mathbf{dd}}$.

By multiplying $\hat{\mathbf{y}}_0^{HS}$ by $\left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{E}_0^{-1}$, we can get

$$\hat{\mathbf{y}}_0^{HS} = \sqrt{\beta_0}\mathbf{x}_0^{HS} + \sum_{i=1}^{K}\sqrt{\beta_i}\left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)\hat{\mathbf{x}}_i^L + \left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{E}_0^{-1}\hat{\mathbf{z}}_0, \quad (12)$$

with $\hat{\mathbf{y}}_0^{HS} = \left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{E}_0^{-1}\hat{\mathbf{y}}_0^{HS}$ and $\hat{\mathbf{x}}_i^L = \left(\mathbf{F}_N \otimes \mathbf{F}_M^H\right)^{-1}\mathbf{x}_i^L$. It is assumed that all user signals carry an equal amount of information and the noise power has been normalized. The transmit signal-to-noise ratio (SNR) can be defined as $\varphi = E\left\{\left|x_0^{HS}[k,l]\right|^2\right\} = E\left\{\left|x_i^L(n)\right|^2\right\}$.

Thus, the SINR of V_0 is calculated as

$$\mathrm{SINR}_{0,kl}^{HS} = \frac{\beta_0\varphi}{\varphi\sum_{i=1}^{K}\beta_i + \frac{1}{MN}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1}\left|E_0^{k,l}\right|^{-2}} . \quad (13)$$

Step2: Detect the NOMA users' signals. we first detect the signal of V_0 in the DD domain. After demodulation, the NOMA users' signals are processed in the TF domain.

$$\mathbf{y}_i^L = \sqrt{\beta_0}\mathbf{H}_i^{\mathrm{TF}}\mathbf{x}_0^{HS} + \underbrace{\sum_{q=1}^{K}\sqrt{\beta_q}\mathbf{H}_i^{\mathrm{TF}}\mathbf{x}_q^L + \mathbf{z}_i}_{\text{Interference and noise}} , \quad (14)$$

where $\mathbf{H}_i^{\mathrm{TF}}$, $1 \le i \le K$, is the TF domain block diagonal channel of U_i and its diagonal block is $\mathbf{G}_{i,0}$.

Assuming that the NOMA users' channels are time-invariant, we can simplify the channel of the *i-th* user as

$$h_i(\tau) = \sum_{a=0}^{P_i} h_{i,a}\delta\left(\tau - \tau_{i,a}\right) . \quad (15)$$

Similar to the detection of the high-speed user's signal in Step 1, the signal of NOMA users after adding the detection matrix $\mathbf{F}_M^H$ is [10]

$$\tilde{\mathbf{y}}_i^L = \square\mathbf{E}_i\mathbf{F}_M^H\left(\sqrt{\beta_0}\mathbf{x}_0^{HS} + \sum_{q=1}^{K}\sqrt{\beta_q}\mathbf{x}_q^L\right) + \tilde{\mathbf{z}}_i , \quad (16)$$

where $\tilde{\mathbf{y}}_i^L = \mathbf{F}_M^H\mathbf{y}_i^L$, $\tilde{\mathbf{z}}_i = \mathbf{F}_M^H\mathbf{z}_i$, $1 \le i \le K$. $\square\mathbf{E}_i$ is the diagonal matrix with dimension $M \times M$, whose $(l+1)-th$ main diagonal element is $\tilde{E}_i^l = \sum_{m=0}^{M-1} b_{i,0}^{m,1} e^{j2\pi\frac{lm}{M}}$ and $b_{i,0}^{m,1}$ is the element at the $(l+1)$*-th* row and the first column of $\mathbf{G}_{i,0}$.

Multiply $(\mathbf{F}_M^H)^{-1}\square\mathbf{E}_i^{-1}$ matrix before $\tilde{\mathbf{y}}_i^L$ to obtain the received SINR of the high-speed user at U_i as

$$\mathrm{SINR}_0^i = \frac{\varphi\beta_0}{\varphi\sum_{i=1}^{K}\beta_i + \frac{1}{M}\sum_{l=0}^{M-1}\left|\tilde{E}_i^l\right|^{-2}} . \quad (17)$$

For SIC in the following, the NOMA users must successfully decode the signal of V_0. We have $\log_2(1+\mathrm{SINR}_0^i) \ge R_0$, where R_0 is the threshold at V_0. Then, SIC can be performed by the NOMA users. With the assumption that the signal of V_0 is successfully demodulated and can be removed ideally, the NOMA signal of U_i in the TF domain is obtained as

$$\begin{aligned} Y_i^L[n,m] &= \sum_{q=1}^{K}\sqrt{\beta_q}H_i[n,m]X_q^L[n,m] + Z_i[n,m] \\ &= \sqrt{\beta_i}H_i[n,m]x_{m+1}^L(n) + Z_i[n,m] \end{aligned} , \quad (18)$$

where the final step follows the mapping in (3). Since the signal of U_i is only related to $Y_i^L[n,i-1]$, the SINR of the NOMA user U_i can be detected by a single-tap equalizer as

$$\mathrm{SINR}_i = \varphi\beta_i\left|\tilde{E}_i^{i-1}\right|^2 . \quad (19)$$

III. NOMA-OTFS System Power Allocation Strategy

A. Iterative Power Allocation Optimization Strategy

In the downlink NOMA-OTFS vehicular system, considering the cross-domain transmission, power allocation can be based on three objectives: (1) Ensuring the high-speed user meet its QoS and target rate R_0; (2) Guaranteeing the NOMA users can successfully decode the signal of the high-speed user; (3) Ensuring successful signal detection of the NOMA users.

In this subsection, we propose a convex optimization-based hierarchical power allocation strategy to ensure the fairness in NOMA users and maximize minimum data rates. The optimization problem can be formulated as

$$\begin{aligned} \max_{\boldsymbol{\beta}}.\min. \quad & \{\log_2\left(1+\mathrm{SINR}_i\right),\ \ i=1,\ldots,K\} \\ \text{s.t.} \quad & \left\{\log_2\left(1+\mathrm{SINR}_0^i\right)..R_0, i=0,\ldots,K\right\} \\ & \beta_0 + \sum_{i=1}^{K}\beta_i = P_T \\ & \beta_i \ge 0 \end{aligned} ,(20)$$

where SINR_0^i denotes the received SINR at U_i from V_0 and $\boldsymbol{\beta} = [\beta_1 \dots \beta_K]$. The objective function in (20) aims to maximize the minimum data rates of $U_i, i = 1, \dots, K$. The constraint guarantees the target data rate of V_0 while ensuring that the NOMA users can successfully decode the signal of V_0 with a given total system power limit. Since the optimization problem is non-convex, auxiliary variable is introduced to transfer the original problem into a convex one as

$$\begin{aligned} \max_{\boldsymbol{\beta}, T}. \quad & T \\ \text{s.t.} \quad & \left\{ \mathrm{SINR}_0^i \geq R_0,\ i = 1, \dots, K \right\} \\ & T \leq SINR_i \\ & \boldsymbol{\beta} \geq 0 \\ & \beta_0 + \sum_{i=1}^{K} \beta_i = P_T \end{aligned} \quad . \qquad (21)$$

The above optimization problem is equivalent to (20). In order to optimize the power allocation for each NOMA user, we present the IP-AO algorithm in Table I.

TABLE I. IP-AO ALGORITHM

```
Algorithm 1: IP-AO
1: Input: $\mathbf{H}_0^{\mathrm{dd}}, \mathbf{H}_i^{\mathrm{TF}}, \varphi$ ;
2: Initialization: $R_0 = 0.1\,\mathrm{bps/Hz}$, $P_T = 10\,\mathrm{dB}$, $M = 3$ ;
3: Introduce auxiliary variable $T$ to transform (20) into (21);
4: Solve (21) by CVX to obtain $\boldsymbol{\beta}$ ;
5: return $\boldsymbol{\beta}$ and the rate of each NOMA user.
```

Then, we easily find the maximum value of the minimum rate for all NOMA users through Algorithm 1.

B. Prior Determination based Optimal Power Allocation

The key strategy of this paper is termed as Max-Min power allocation, targeting the maximization of the minimum rate for the NOMA users to enhance the fairness among them. Based on this strategy, the following assumptions can be made:

Assumption 1: For the Max-Min power allocation, we aim to get equal bit rate for all NOMA users.

The validity of Assumption 1 stems from the fairness goal of the Max-Min power allocation approach, and it may achieve a closed-form solution of the optimization problem with the following steps.

Step1: Initialize total power P_T and the channel state information for V_0 and U_i as $\mathbf{H}_0^{\mathrm{dd}}$ and $\mathbf{H}_i^{\mathrm{TF}}, i = 1, \dots, K$.

Step2: To calculate the allocated power for the high-speed user while meeting the QoS requirements, the convenience of SIC, and the total power constraints, we can perform the following procedure.

1) To meet the target rate of V_0 and ensure that the SIC can be properly executed, we can get

$$\beta_0' = \max\left(\frac{R_0\left(\varphi P_T + \frac{1}{MN} \sum_{k=0}^{N-1} \sum_{l=0}^{M-1} \left| E_0^{k,l} \right|^{-2} \right)}{\varphi(1 + R_0)}, \frac{R_0\left(\varphi P_T + \frac{1}{M} \sum_{l=0}^{M-1} \left| \tilde{E}_i^{l} \right|^{-2} \right)}{\varphi(1 + R_0)} \right), \qquad (22)$$

where β_0' is the power allocated to high-speed user V_0, $E_0^{k,l}$ and $\tilde{E}_i^l$ represent the channel information of V_0 at itself and at the $i-th$ NOMA user U_i.

2) Given the total power constraint and the non-negative power for the NOMA users, the final power allocation for the high-speed user is determined as

$$\beta_0 = \max\left(P_T, \beta_0' \right) \qquad (23)$$

Step 3: After determining the power for V_0, the remaining power can be allocated to NOMA users. The allocated power for the *i-th* user becomes

$$\beta_i = \begin{cases} \max\left(P_T - \beta_0 , 0 \right), & i = 1 \\ \max\left((P_T - \beta_0) F_i \Big/ \sum_{i=1}^{K} F_i , 0 \right), & i \geq 2 \end{cases} , \qquad (24)$$

with $F_i = \left(\prod_{i=1}^{M} \left| \tilde{E}_i^{i-1} \right|^2 \right) \Big/ \left| \tilde{E}_i^{i-1} \right|^2$.

Although Assumption 1 provides a simple framework to analyze the optimization problem, the bit rates of all NOMA users may not be exactly the same due to a variety of factors in real vehicular communication, e.g. distance from the user to the BS etc. According to the current analysis, this assumption provides a reasonable starting point and makes a more feasible derivation of the closed-form solution.

IV. NUMERICAL RESULTS

TABLE II. SIMULATION PARAMETERS

Parameters	Values
Subcarrier (M), Time slot (N)	4, 8
Carrier frequency	4 GHz
subcarrier interval (Δf)	10 kHz
Number of NOMA users	3 or 4
Channel estimation	Ideal
Channel model	Rayleigh
Equalization	FD-LE

In this section, Monte Carlo simulation is performed for the proposed power allocation scheme in the NOMA-OTFS vehicular system. The main system parameters are set as Table Ⅱ . In particular, the delay Doppler indexes of the four multipath channel taps are set to be (0,1,3,5) and (0,1,2,3) for V_0. The maximum Doppler shift of the channel is 1875 Hz that corresponds to a maximum speed of 140.625 km/h. Different NOMA users exist in the system with maximum transmission delays 12.5 μs, 25 μs, and 37.5 μs. Furthermore, for all user channels, we assume the multipath fading with $\sum_{t=0}^{P_i} E\left\{ \left| h_{i,t} \right|^2 \right\} = 1$ and $\left| h_{i,t} \right|^2 \sim CN\left(0, \frac{1}{P_i} \right)$.

As shown in Fig. 2, both of the proposed power allocation schemes show significant advantages in rate distribution compared with the equal power allocation (EPA) strategy. Both schemes can ensure the rate fairness among NOMA users under different transmit SNRs, while the EPA method leads to obvious rate difference. However, power allocation schemes based on user fairness are realized at the expense of system throughput or efficiency.

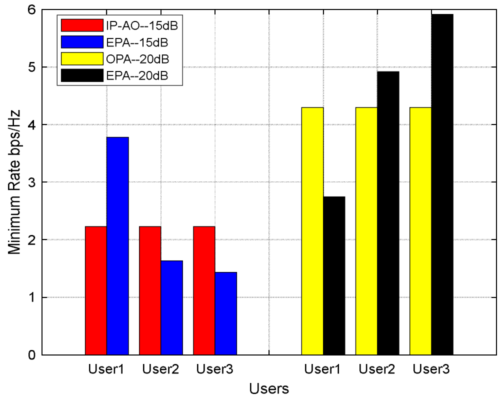


Fig 2. IP-AO and OPA strategy vs. EPA scheme

We compare the NOMA users' minimum bit rates under different power allocation policies and NOMA user numbers in Fig. 3, the proposed IP-AO and OPA outperform both EPA and channel state information-based power allocation (CSI-PA) strategy [9] (the allocated power of each user is proportional to the channel gain). The OPA scheme achieves the maximum system bit rate with minimum complexity. Meanwhile, CSI-PA method exacerbates user unfairness and results in suboptimal performance. Specially, the fixed power allocation yields decreased minimum rates with more users, which is due to the total power constraints.

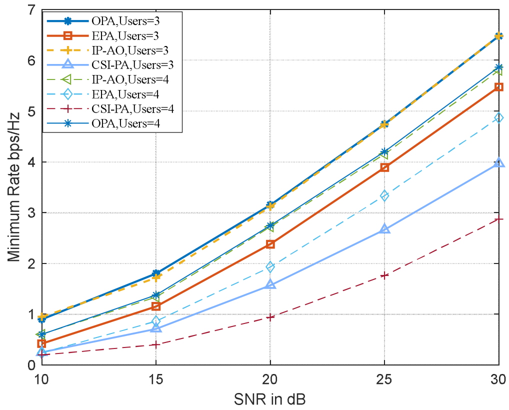


Fig 3. Minimum bit rate of NOMA users with different power allocation schemes and user numbers with 4QAM and R_0=1 bps/Hz.

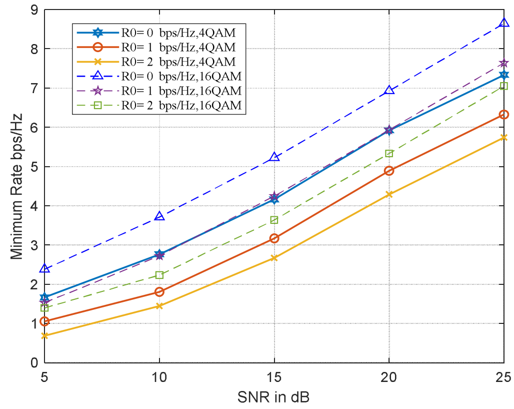


Fig 4. Minimum bit rate of NOMA users with different modulation schemes and R_0.

Fig. 4 depicts the minimum bit rates of NOMA users for OPA policy with different modulation modes and QoS constraints of V_0 considering three NOMA users, we use the appropriate channel coding scheme respectively. It is seen that the case of 16QAM outperforms the case of 4QAM since it transmits more bits per symbol. As the QoS constraints of V_0 intensify (continuously increasing with R_0), the minimum rate for NOMA users diminishes due to increased power allocation for QoS of V_0. In the low SNR regions, this depletes the power of the NOMA users, which further impacts the accuracy and performance of SIC. When without rate constraints for V_0, the high-speed user will not influence power allocation, the NOMA users' rates peaked.

V. Conclution

In this paper, we explore the power allocation in the downlink NOMA-OTFS vehicular communication system and propose two strategies, IP-AO and OPA, to ensure the transmission fairness of NOMA users and guarantee the QoS of high-speed user. Both of the two power allocation strategies are validated through simulation compared with the traditional power allocation strategies, which shows the advantages in the minimum bit rate of NOMA users. The OPA scheme takes both the optimal performance and low complexity into account. Our future work can focus on the impact of fractional Doppler channel conditions on the performance of the NOMA-OTFS system.

References

[1] S. Chen, et al., "Vehicle-to-everything (V2X) services supported by LTE-based systems and 5G." *IEEE Commun Mag.*,1.2 (2017): 70-76.

[2] Z. Li, K. Wang, T. Yu and K. Sakaguchi, "Het-SDVN: SDN-Based Radio Resource Management of Heterogeneous V2X for Cooperative Perception," in *IEEE Access*, vol. 11, pp. 76255-76268, 2023.

[3] Z. Wei, et al.,"Orthogonal time-frequency space modulation: A promis-ing next-generation waveform," *IEEE Wireless Commun.*, vol. 28, no. 4, pp. 136-144, Aug. 2021.

[4] Hadani, Ronny, et al., "Orthogonal time frequency space modulation." in *Proc. IEEE WCNC*, 2017, pp. 1-6.

[5] P. Raviteja, K. T. Phan, Q. Jin, Y. Hong and E. Viterbo, "Low-complexity iterative detection for orthogonal time frequency space modulation," in *Proc. IEEE WCNC*, 2018, pp. 1-6.

[6] Arman Farhang, et al.,"Orthogonal Time Frequency Space Modulation: Principles and Implementation," in Radio Access Network Slicing and Virtualization for 5G Vertical Industries , IEEE, 2021, pp.103-120.

[7] B. C. Pandey, et al., "Low Complexity Precoding and Detection in Multi-User Massive MIMO OTFS Downlink," *IEEE Trans. Veh*, vol. 70, no. 5, pp. 4389-4405.

[8] Z. Yuan, et.al., "Multi-User Shared Access for Internet of Things," in *Proc. IEEE 83rd VTC Spring*, 2016, pp. 1-5.

[9] Yuxiang Fan, "Downlink power allocation strategy for OTFS-NOMA-based V2X systems," *in* Proc. SPIE 12594, Second International Conference on EIECC 2022.

[10] Z. Ding, et,al., "OTFS-NOMA: An Efficient Approach for Exploiting Heterogenous User Mobility Profiles," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7950-7965, Nov. 2019.

[11] T. Sefako and T. Walingo, "Application of Biological Resource Allocation Techniques to SCMA NOMA Networks," *IEEE AFRICON*, 2019, pp. 1-7.

[12] H. V. Cheng, E. Björnson and E. G. Larsson, "Performance Analysis of NOMA in Training-Based Multiuser MIMO Systems," *IEEE Wireless Commun.*, vol. 17, no. 1, pp. 372-385, Jan. 2018.

[13] S. Bai, et.al., "The Allocation Strategy Optimization of Mobile Energy Storages for Power System Restoration," in *Proc. IEEE I&CPS Asia*, Shanghai, China, 2022, pp. 1742-1746.

[14] Z. Ding, "Robust Beamforming Design for OTFS-NOMA," IEEE Open J. Commun. Soc., vol. 1, pp. 33-40, 2020.

Hao Xu was born in Heze, Shandong Province, China in Dec 2001. He studied at Shandong Agricultural University from 2018 to 2022 and obtained a bachelor's degree in communication engineering. Now he is studying for a master's degree in electronic information engineering at Shandong University. His specific research fields include optimal design on orthogonal time frequency space modulation and signal detection based on nonlinear equalization.

Zhiquan Bai received the M.Eng. degree in communication and information system from Shandong University, Jinan, China, in 2003, and the Ph.D. degree (Hons.) from INHA University, Incheon, South Korea, in 2007, under the Grant of Korean Government IT Scholarship. He held a postdoctoral position with INHA University, and was a Visiting Professor with The University of British Columbia, Canada. He is currently a Professor with the School of Information Science and Engineering, Shandong University. His research interests include cooperative technology and spatial modulation, orthogonal time frequency space modulation, MIMO technology, resource allocation and optimization, and deep-learning based 5G wireless communications. He is a member of the editorial board of Journal of Systems Engineering and Electronics and also an associate editor of the International Journal of Communication Systems.

Jinqiu Zhao received B.E. degree from Shandong Normal University, Jinan, China, in 2020. She is currently pursuing her Ph.D. degree in the School of Information Science and Engineering, Shandong University, Qingdao, China. Her main research interests include reconfigurable intelligent surface and machine learning.

Dejie Ma is currently pursuing the M.S. degree in Electronic Information at the School of Information Science and Engineering, Shandong University, Qingdao, China. His research interests include reconfigurable intelligent surface, integrated sensing and communication and signal processing.

Bangwei He (Member of IEEE) was born in Yantai, Shandong Province, China in May 1999. He studied at Shandong University from 2017 to 2021 and obtained a bachelor's degree in communication engineering. Now he is studying for a master's degree in electronic information engineering at Shandong University. His specific research fields include channel estimation based on orthogonal time frequency space modulation, face and voiceprint recognition based on deep learning.

Kyung Sup Kwak received his BS degree from the Inha University, Inchon, Korea,in 1977 and his MS degree from the University of Southern California in 1981and his PhD degree from the University of California at San Diego in 1988, under the Inha University Fellowship and the Korea Electric Association Abroad Scholarship Grants, respectively.From 1988 to 1989, he was with Hughes Network Systems, San Diego, California. From 1989 to 1990, he was with the IBM Network Analysis Center, North Carolina. Since then, he has been with the School of Information and Communication Engineering, Inha University, Korea, as a professor. He is the director of UWB Wireless Communications Research Center (UWB-ITRC).Since 1994, he served as a member of the board of directors and the vice president and the president of Korean Institute of Communication Sciences (KICS) in 2006 and the president of Korea Institute of Intelligent Transport Systems (KITS) in 2009. He received many research awards, such as the award of research achievements in UWB radio from the Ministry of Information and Communication and Prime Ministry of Korea in 2005 and 2006, respectively. In 2008, he is elected as Inha Fellow Professor (IFP). In 2010, he received the Korean President official commendation for his contribution to ICT innovation and industrial promotion.He published more than 100 SCI journal papers, 300 conference/domestic papers, obtained 20 registered patents and 35 pending patents, and proposed 21 technical proposals on IEEE 802.15 (WPAN) PHY/MAC. He is one of the members of the IEEE, IEICE, KICS, and KIEE. His research interests include multiple access communication systems, cognitive radio, UWB radio systems and WBAN, WPAN, and sensor networks.

Performance Evaluation of UAV-based NOMA for 5G and Beyond

Mounika Neelam*, Anuradha Sundru**

*Department of ECE, National Institute of Technology, Warangal, India

** Department of ECE, National Institute of Technology, Warangal, India

nm22ecr2r10@student.nitw.ac.in, anuradha@nitw.ac.in

***Abstract*— The use of Unmanned Aerial Vehicles (UAVs) as flying Base Stations (BSs) is an efficient way to enhance wireless communication throughput and coverage, especially in Line-of-Sight (LoS) networks. Both military and civilian applications have shown interest in studying communication aided by UAVs. Utilizing UAVs as BSs to expand the range of current cellular networks is widely discussed, and Non-Orthogonal Multiple Access (NOMA) is a potential approach for UAV communications. This paper investigates the use of NOMA for upcoming 5G radio access and beyond, focusing on a multiuser communication system in which a UAV-BS with a single antenna communicates with ground users via NOMA. Simulation results show that NOMA outperforms Orthogonal Multiple Access (OMA) in terms of rates, Energy Efficiency (EE), Spectral Efficiency (SE), and Signal to Noise Ratio (SNR). The main objective is to develop a MATLAB platform for NOMA-related systems at the system-level analysis. NOMA's SE performance is significantly higher than other potential 6G options. The paper provides a NOMA system for two or more users, and the system-level analyses include scenarios with three cells, seven cells, and 19 cells.**

Keywords—6G, EE, NOMA, OMA, UAV, SE, SNR.

I. Introduction

UAVs have an inclusive range of uses in the civilian and commercial sectors in addition to carrying out combat operations, including cargo delivery, traffic control, aerial imaging, and other tasks. UAVs are now used in communication systems for wireless broadcasting, data collection, and relaying. Due to their flexibility and mobility, UAVs are used as aerial BSs which can offer ubiquity coverage and meet customers' diverse requirements. UAVs can be deployed more flexibly than standard terrestrial communication devices, which are often fixed once deployed.

The use of UAVs and advanced communication technologies have garnered much interest from academic and business circles. In the past few years, the application of UAVs in civil and military sectors, such as cargo delivery, search and rescue, and precision agriculture, has rapidly increased. Moreover, the use of UAVs for communication purposes is expected to provide faster deployment options, increased mobility, and reliable line-of-sight connections in comparison to traditional ground-based communication methods. Therefore, UAVs are poised to play a significant role in the forthcoming communication networks.

However, due to the increasing density of equipment and frequency reuse, the amount of resources available for UAVs is diminishing rapidly. To address this issue, NOMA has been recognized as a promising technique that allows UAVs to reuse the spectrum assigned to terrestrial BSs or users. By implementing Successive Interference Cancellation (SIC) at receivers and Superposition Coding (SC) at transmitters, NOMA can enhance network performance, support more users, and improve SE, EE, and SNR. Several studies have explored the integration of NOMA into UAV communication systems. This paper focuses on a new NOMA-based framework for UAV networks and investigates uplink NOMA transmission in UAV communication networks that are connected to cellular networks.

This paper explores a method for improving spectrum utilization by using a UAV system with NOMA technology to connect two ground users. The challenges faced by wireless networks operating on 5G and beyond 5G (B5G) are significant, with high SE requirements and connection rates causing damage to BSs and making reliable service provision difficult, particularly in disaster zones where IoT devices are located. UAVs offer flexibility and mobility to quickly provide wireless coverage in such areas, and NOMA is a strong contender for enhancing SE and enabling IoT device connection by sharing physical resources such as time, frequency, and code among multiple users, in contrast to the traditional OMA method.

An emergency communications system for a NOMA-UAV network is established in this paper. During exceptional events, such as natural disasters, clogged roadways, concerts, athletic events, war situations, etc., UAVs can serve as flying BSs to supplement traditional communication networks and manage traffic. By acting as temporary hotspots, UAVs can establish links between a safe location and areas affected by the tragedy. UAV-BSs will provide LoS air-to-ground connectivity to ground users, enabling various wireless communications. Researchers have recently shown significant interest in UAV-enabled communication, and numerous studies have been conducted on the subject.

The optimal choice for UAV communication is a UAV with a single antenna due to the poor transmission reliability of downlink communication with UAVs. However, a single UAV cannot offer the same transmission bandwidth to multiple users concurrently. Typically, each user is assigned their own bandwidth channel, and the available rate substantially depends on the number of users sharing the bandwidth. By splitting users in the power domain, NOMA can effectively serve numerous clients with non-orthogonal resources simultaneously. NOMA can enhance the obtainable rate for remote users by enabling close users to receive the information intended for faraway users who face weaker received signals.

II. System Model

Let us assume that a single-antenna UAV, like the one shown in Fig. 1, is used to provide coverage for a certain outdoor venue (such as a stadium, traffic jam, concert, etc.). Assuming there are M ground users in the area m ∈ {1, . . ., M/2}, we will refer to those as "near users" or "cell-centered users" because they are situated closer to the UAV (in terms of Euclidean distance) than the other users. The remaining M/2 users m ∈ {M/2+1,...,M} are known as "far users" or "cell-edge users" since they are situated comparatively further away from one another. Each close user can be paired with a far user by the UAV using NOMA.

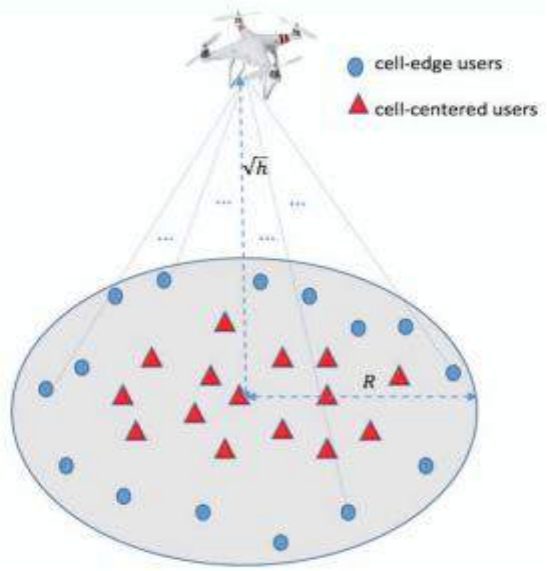


Fig. 1. UAV-NOMA system

We propose a system with a UAV that can speak to two NOMA users, U1 and U2, through the system. Such a UAV maintains a circular trajectory with a constant velocity Uv, a circular trajectory of radius U_r, and an altitude U_h while flying. The location of the UAV is represented as, $UAV(U_r\cos\varphi, U_r\sin\varphi, U_h)$where φ is the angle of the UAV position in the UAV circle. As a result, the following equation can be used to determine the Euclidean distance between users U1 and U2 and the UAV.

$$\overline{U_{d1}} = \sqrt{U_h^2 + U_r^2 + U_L^2 - 2U_rU_L\cos\phi} \tag{1}$$

$$\overline{U_{d2}} = \sqrt{U_h^2 + U_r^2 + U_L^2 + 2U_rU_L\cos\phi} \tag{2}$$

Finally, $P_{LOS}(\theta_m)$ and $P_{NLOS}(\theta_m)$ =1- $P_{LOS}(\theta_m)$ denote the probability of LOS and NLOS respectively. These final two numbers are derived using the formula below.

$$\mathrm{P}_{LOS}\left(\theta_m\right) = \frac{1}{1+pe^{-q(\theta_m - p)}} \quad ,m \in \{1,2\} \tag{3}$$

Where $\theta_m = \arcsin\left(\frac{H}{d_m}\right)$ is elevation angle of the UAV with respect to each user, and p and q are constant values based on the surroundings. Denote $U_{u,m}$ as the serving indicator. $U_{u,m} = 1$indicates that the UAV 'u' is serving the user 'm', $U_{u,m} = 0$if otherwise. So, here is the transmitting signal from UAV 'u' to user 'm' can be expressed as,

$$i^u(t) = \sum_{m=1}^{M} U_{u,m}(t)\sqrt{P_m^u(t)}\, i_m^u(t) \tag{4}$$

Where $i_m^u(t)$ is the transmitting signal of UAV 'u', $P_m^u(t)$represents the assigned power of user 'm', may be used to express the superposition. Equation (4) has the result that the signals received at the user 'm' are

$$z_m^u(t) = g_m^u(t)i^u(t) + I_{\mathrm{inter}.m}^u(t) + I_{\mathrm{intra}.m}^u(t) + \sigma_m^u(t) \tag{5}$$

Where $\sigma_m^u(t)$ denotes the AWGN, $I_{\mathrm{intra}.m}^u(t)$denotes the intra-cluster interference, and $I_{\mathrm{inter}.m}^u(t)$ denotes the cumulative inter-cluster interference to user 'm' from all other UAVs other than UAV 'u'.

The composition of $I_{\mathrm{inter}.m}^u(t)$ can be expressed as

$$I_{\mathrm{inter}.m}^u(t) = \sum_{r=1,r\neq u}^{U} g_m^u(t)\sqrt{P^r(t)}\, i^r(t) \tag{6}$$

Where $P^r(t)$represents the total power consumed by the UAV $r \neq u$, which is represented by,

$$P^r(t) = \sum_{m=1}^{M} U_{u,m}(t)P_m^r(t) \tag{7}$$

and $P^r(t)$denotes channel gain between the UAV and user 'm'.

In order to eliminate some of the intra-cell interference at the receiver side, the NOMA protocol employs SIC. Successfully identifying interference depends on identifying the optimal decoding sequence for SIC. Since the channel gain and inter-cell interference of each user may fluctuate due to movement, a dynamic decoding order must be utilized. To this end, the auxiliary term $A_m^u(t)$specified in equation (8) functions as a decoding order criterion and represents the comparable channel gain.

$$A_m^u(t) = \frac{U_{u,m}(t)g_m^u(t)}{\sum_{r=1,r\neq u}^{T} g_m^u(t)P^r(t) + \sigma_m^u(t)^2} \tag{8}$$

The user's data rate when linked to a UAV u can be computed using the formula

$$\Re_{\pi(m)}^u(t) = B\log_2\left(1 + \gamma_{\pi(m)}^u(t)\right) \tag{9}$$

where B is the UAV's bandwidth. As a result, the sum data rate at time 't' can be determined as

$$\Re(t) = \sum_{u=1}^{U}\sum_{m=1}^{M} \Re_{\pi(m)}^u(t) \tag{10}$$

As a result, the throughput is

$$\Re = \sum_{t=0}^{T} \Re(t) \tag{11}$$

III. Simulation Results

The study revealed a significant problem faced by users in multi-cell applications. Our models featuring 3, 7, and 19 cells discovered interference from the cells in the first or second tier with the core cell. Inter-cell interference was calculated using a distance-based method, where the interference was a function of the distance between the cell centre BS and the nearby BS. To assess system-level NOMA simulator capabilities, we examined three scenarios with randomly distributed users within each cell: the 3-cell, 7-cell, and 19-cell scenarios. Each cell's interference model was considered, incorporating the link performance model. For example, Fig. 2(a) illustrates the boundary shared by three cells, where the core cell is surrounded by two cells. In this situation, the interference caused by a user's two neighbouring cells is considered in the link performance model. Figure 2(b) shows a boundary shared by seven cells, and the link performance model accounts for the interference caused by a user's six neighbouring cells. Finally, Fig. 2(c) displays a scenario with 19 cells sharing a border, and the link measurement model determines intercell interference, SNR, and user capacity based on the interference caused by each of the cell's 18 adjacent cells on its users.

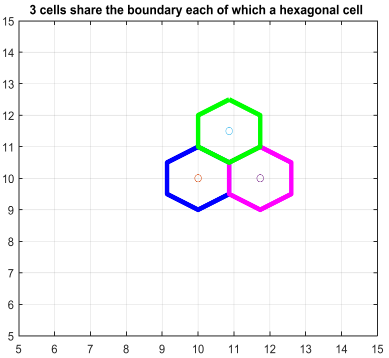


Fig. 2. (a) Three-Cell Scenario

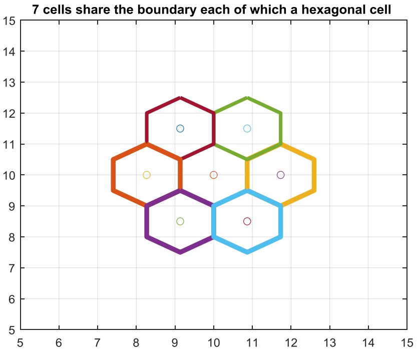


Fig. 2. (b) Seven-Cell Scenario

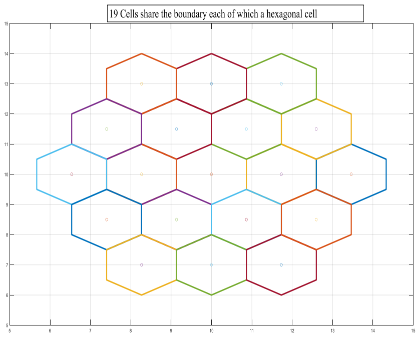


Fig. 2. (c) Nineteen-Cell Scenario

In this paper, a comparison is made between user 1 and user 2 rates at different heights of UAVs (100, 150, and 250 meters) in three environments - rural, urban, and suburban as shown in Figure 3-8. The performance of NOMA and OMA is assessed on the basis of SE and EE graphs at various UAV heights and in different scenarios. Fig. 3 illustrates the bit rates of users 1 and 2 in rural areas considering all UAV heights, indicating that NOMA performs better than OMA in all scenarios. Similarly, Figure 4 shows the SE versus EE performance in rural areas, again confirming the superiority of NOMA over OMA. The comparison between user 1 and user 2 rates in bits per second is shown in Figure 5 for the suburban area, while Figure 6 plots the SE versus EE for the same area, resulting in the same conclusion of NOMA outperforming OMA. In the urban area, Figures 7 and 8 show that NOMA is consistently better than OMA in terms of user rates and SE versus EE performance.

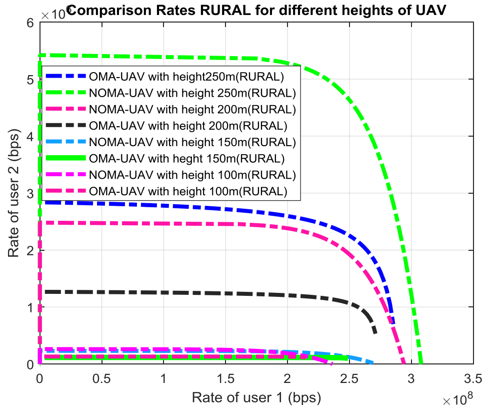


Fig. 3. Rate comparison of Rural area for different heights of UAV

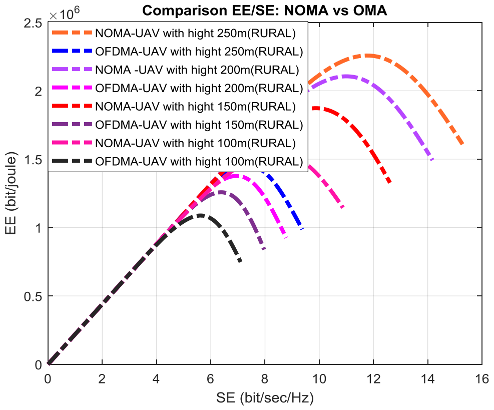


Fig. 4. Comparison of SE vs EE in a rural area for different heights of UAV

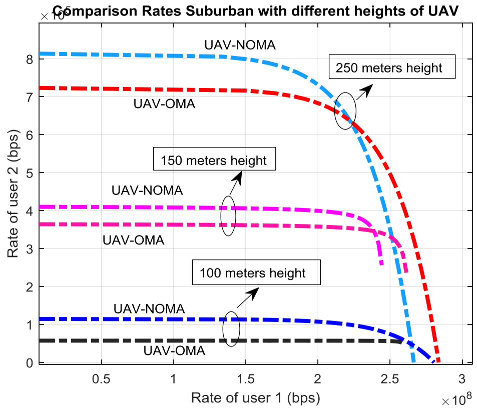


Fig. 5. Rate comparison of Sub-Urban area for different heights of UAV

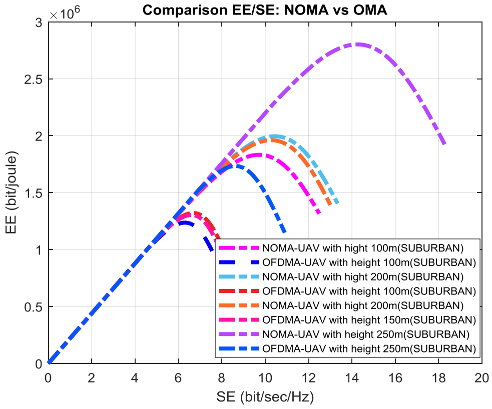


Fig. 6. Comparison of SE vs EE in Sub-Urban area for different heights of UAV

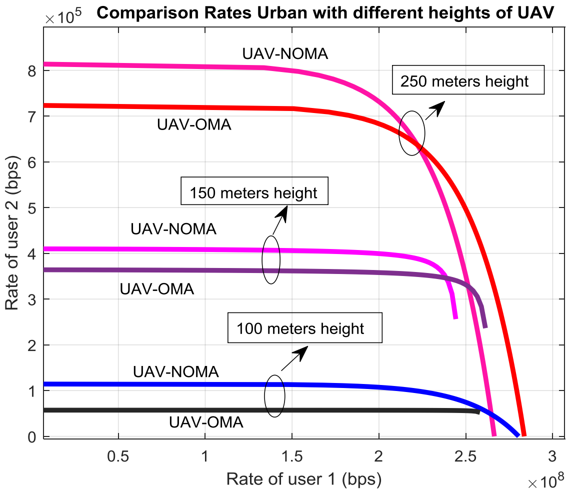


Fig. 7. Rate comparison of Urban area for different heights of UAV

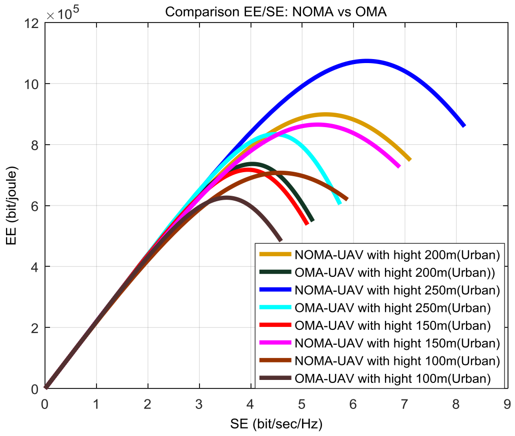


Fig. **8**. Comparison of SE vs EE in Urban area for 250m UAV height

IV. Conclusion

This paper provided a thorough comparative analysis of altitude-fixed UAV-NOMA and UAV-OMA systems in terms of their performance and efficiency. MATLAB is used to evaluate the performance for both systems UAV-OMA and UAV-NOMA and conducted a comprehensive examination of various scenarios, including 3-cell, 7-cell, and 19-cell setups. One significant contribution of the paper was the inclusion of multi-user scenarios, which extended beyond the typical analysis limited to two users in NOMA-related publications. By considering more than two users in a cell, the paper provided insights into the practical implications and performance of NOMA in real-world scenarios. Furthermore, the paper addressed the impact of inter-cell interference in multi-cell scenarios. This aspect is crucial for assessing the scalability and performance of UAV-NOMA and UAV-OMA systems, as interference between neighboring cells can have a significant impact on system-level analysis. Overall, the paper's findings contribute to the understanding of SE and EE UAV-NOMA and UAV-OMA systems, shedding light on their performance, efficiency, and practical implications. The thorough examination of different scenarios and considerations, such as multi-user setups and inter-cell interference, enhances the applicability and relevance of the study in real-world deployments.

References

[1] S. A. Abdel Hakeem, H. H. Hussein, and H. W. Kim, “Vision and research directions of 6G technologies and applications,” J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 6, pp. 2419–2442, 2022, doi: 10.1016/j.jksuci.2022.03.019.

[2] M. F. Sohail, C. Y. Leow, and S. Won, “Non-Orthogonal Multiple Access for Unmanned Aerial Vehicle Assisted Communication,” IEEE Access, vol. 6, pp. 22716–22727, 2018, doi: 10.1109/ACCESS.2018.2826650.

[3] M. Aldababsa, M. Toka, S. Gökçeli, G. K. Kurt, and O. Kucur, “A Tutorial on Nonorthogonal Multiple Access for 5G and Beyond,” Wirel. Commun. Mob. Comput., vol. 2018, 2018, doi: 10.1155/2018/9713450.

[4] A. Ebrahim, A. Celik, E. Alsusa, and A. M. Eltawil, “NOMA/OMA mode selection and resource allocation for beyond 5G networks,” IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC, vol. 2020-Augus, pp. 0–5, 2020, doi: 10.1109/PIMRC48278.2020.9217161.

[5] M. N. Boukoberine, Z. Zhou, and M. Benbouzid, “A critical review on unmanned aerial vehicles power supply and energy management: Solutions, strategies, and prospects,” Appl. Energy, vol. 255, no. September, 2019, doi: 10.1016/j.apenergy.2019.113823.

[6] A. A. Nasir, H. D. Tuan, T. Q. Duong, and H. V. Poor, “UAV-enabled communication using NOMA,” IEEE Trans. Commun., vol. 67, no. 7, pp. 5126–5138, 2019, doi: 10.1109/TCOMM.2019.2906622.

[7] M. Erdelj, E. Natalizio, K. R. Chowdhury, and I. F. Akyildiz, “Help from the Sky: Leveraging UAVs for Disaster Management,” IEEE Pervasive Comput., vol. 16, no. 1, pp. 24–32, 2017, doi: 10.1109/MPRV.2017.11.

[8] Ding, Z., Liu, C., Choi, J., & Elkashlan, M. (2020). Application of Non-Orthogonal Multiple Access in UAV Networks. IEEE Communications Magazine, 58(1), 112-118.

[9] Mozaffari, M., Saad, W., Bennis, M., Nam, Y. H., & Debbah, M. (2016). Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage. IEEE Communications Letters, 20(8), 1647-1650.

[10] Xie, Q., Yu, F. R., Liu, R., Tang, J., & Nallanathan, A. (2019). Non-orthogonal multiple access for 5G and beyond. IEEE Journal on Selected Areas in Communications, 37(1), 106-119.

[11] Liu, L., Dai, L., Zhang, X., & Wang, Z. (2017). UAV-aided communications for 5G and beyond. IEEE Communications Magazine, 55(5), 104-111.

[12] Ding, Z., Adachi, F., & Poor, H. V. (2017). The application of MIMO to non-orthogonal multiple access. IEEE Transactions on Wireless Communications, 15(1), 537-552.

[13] Hu, R. Q., Qian, Y., & Li, G. Y. (2019). 6G wireless networks: Vision, requirements, architecture, and key technologies. IEEE Vehicular Technology Magazine, 14(3), 28-41.

Elkashlan, M., Ding, Z., & Dai, L. (2020). Non-orthogonal multiple access for 6G networks: From theory to practice. IEEE Open Journal of Vehicular Technology, 1, 47-63.

Mounika Neelam is a research scholar in the department at NIT Warangal, India. She received her B.Tech and M.Tech degrees in Electronics and Communication Engineering from JNTUK, India in 2014 and 2016, She has been teaching for more than 5 years. Her current research interests are in the field of Wireless Communications, Cognitive Radio and Coding Theory, Fading channels, etc. Based on her research work she has published more than **30** research papers in various International Journals and conferences. A patent has been granted in her credit and she has authored four textbooks.

Dr. Anuradha Sundru is a professor of Electronics and Communication Engineering Dept, National Institute of Technology, Warangal, India. She received her B.Tech and M.Tech degrees in Electronics and Communication Engineering from the University of Nagarjuna in 1999 and Sri Venkateswara University in 2001, and obtained Ph.D. degree in Electronics and Communication Engineering from the Andhra University, Visakhapatnam in 2012. She has been teaching for more than 21 years. Her current research interests are in the field of Wireless Communications, cognitive Radio and Coding Theory, Fading channels, etc. Currently, **seven** scholars are working under her guidance and **six** scholars have completed their Ph.D. in her guidance. Based on her research work she has published more than **100** research papers in various International Journals and conferences. She has completed four research projects under MHRD, DRDO and DST-SERB and recently she received one DST-SERB project with the worth of **33 Lakhs**. She is the reviewer for various International Journals.

Session 4B: Artificial Intelligence 4

Chair: Prof. Otgonbayar Bataa, Mongolian University of Science and Technology, Mongolia

1 Paper ID: 20240477, 286~289

Multi-Class Document Classification using LayoutLMv1 and V2

Ms. Kounen Fathima, Mr. Athar Ali, Prof. Hee Cheol Kim,

Inje University. Korea(South)

2 Paper ID: 20240384, 290~294

Machine learning based techniques for the Prediction of axillary lymph node metastases in early breast cancer

Mr. Maisam Ali, Mr. Muhammad Yaseen, Mr. Sikandar Ali, Prof. Hee-Cheol Kim,

Inje University. Korea(South)

3 Paper ID: 20240479, 295~299

AI-based logistics system overview and a conceptual framework for digital freight forwarding in logistics

Mr. Md Ariful Islam Mozumder, Mr. Rashedul Islam Sumon, Mr. Ziaullah Khan, Mr. Shah Muhammad Imtiyaj Uddin, Mr. Muhammad Omair Khan, Prof. Hee-Cheol Kim,

Inje University. Korea(South)

4 Paper ID: 20240411, 300~304

The benefits of integrating AI, IoT, and Blockchain in healthcare supply chain management: A multi-dimensional analysis with case study

Mr. Tagne Poupi Theodore Armand, Ms. Kouayep Sonia Carole, Mr. Subrata Bhattacharjee, Mr. Md Ariful Islam Mozumder, Dr. Austin Oguejiofor Amaechi, Prof. Hee-Cheol Kim,

Inje University. Korea(South)

5 Paper ID: 20240321, 305~310

Knowledge-Prompted Estimator:A Novel Approach to Explainable Machine Translation Assessment

Dr. hao yang,

huawei. China

6 Paper ID: 20240393, 311~314

Integration of a Chatbot to facilitate access to educational content in digital universities

Mr. Birahim BABOU, Dr. Khalifa SYLLA, Mr. Mouhamadou Yaya Sow, Prof. Samuel OUYA,

UCAD. Senegal

Multi-Class Document Classification using LayoutLMv1 and V2

Kounen Fathima*, Ali Athar*, Hee-Cheol Kim*

*Department of Digital Anti-Aging Healthcare, Inje University, Gimhae, South Korea

Kounenfathima00@gmail.com, ali.athar1401@gmail.com, heeki@inje.ac.kr

Abstract— **In the age of information explosion, efficient and accurate information retrieval has become a pivotal task across numerous domains, from finance to healthcare and beyond. The LayoutLM model has enhanced the capabilities of existing NLP techniques by enabling them to extract information from complexly structured documents automatically. In this study, we employ a subset of the RVL-CDIP (Ryerson Vision Lab Complex Document Information Processing) dataset, consisting of 400,000 pictures of data organized into 16 groups. Out of the 16 classes, this subset includes all the classes but there was a limit set to 200 images per class which makes the total amount of images as 3200. Accordingly, 1920, 640, and 640 images make up the training, validation, and testing sets. This dataset is a rich collection of documents with diverse structures and content to demonstrate the effectiveness of our proposed method. LayoutLM, with its unique capability to analyze and understand document structures, has been a pivotal component of our methodology. We utilized LayoutLMv2 and version 1 for the purpose of classifying the documents into their respective categories and comparing the results accordingly. Accordingly, the two versions' accuracies are 80.94 and 68.75 percents.**

Keywords— **Natural Language Processing (NLP), LayoutLM Model, Document AI, Document Classification, Transformers**

I. Introduction

Digitizing data and extracting information are crucial in modern data centric operations. Business transactions generate much document data, which is unstructured, making extracting useful information challenging. Digitization provides access to cutting-edge data analysis, which can be quickly processed, analyzed, and displayed using advanced tools and algorithms. Information extraction, a crucial step in this process, turns unstructured data into organized, helpful knowledge. Automation is yet another significant advantage. Industry has a practical need for understanding how to digitize data and extract useful information from document data, and academic research on this subject has exploded in recent years.

The two main kinds of documents are scanned images of paper documents and digital documents created by computers, such as PDF, Excel, digital photographs, etc. These documents cover purchasing documents, industry reports, business emails, sales contracts, employment agreements, commercial invoices, personal resumes, and more. Automatically extracting adequate information from these unstructured documents has potential retail value for the development of enterprises [1].

This paper focuses on fine-tuning the LayoutLM model proposed by Huang, Y. et al [2]. This model specializes in Natural Language Processing (NLP) and is designed to perform text recognition and understanding in documents with complex layouts, such as forms, invoices, and receipts [2]. In recent years, deep learning techniques have revolutionized the field, enabling the development of highly accurate and adaptable document classification models. However, documents often exhibit intricate layouts, making it challenging to extract meaningful content for classification accurately. This is especially true in domains where document structure and formatting are pivotal, such as legal documents, financial reports, and medical records. This pre-trained model works on text-centric and image-centric document AI tasks, overcoming the problem. There are various noteworthy uses for this model beyond just categorizing documents. For example, it can also help in understanding documents, comprehending forms, answering visual questions about documents, and understanding receipts. There are other approaches for document classification such as categorizing reviews, tweets or articles with sentiment analysis [3].

This paper will review some related works in section II. Following that, we will delve into the workings of our model in section III. In section IV, we will discuss the results of the model. Finally, in section VI, we will conclude the paper by discussing future work.

II. Related Work

In recent years, document classification has been a well-researched task among others in natural language processing (NLP). Various approaches have been used, including deep learning techniques using CNNs (Convolutional Neural Networks), RNNs (Recurrent Neural Networks), RNTNs (Recursive Neural Tensor Networks), and RCNNs (Recurrent Convolutional Neural Networks) [4].

As more attention is being given to the analysis of visually engaging documents, various approaches have emerged, such as transformer-based, deep learning-based, and hybrid methods. One of the newer hybrid methods combines LayoutLM with BiLSTM and CRF to create a multi-modal fusion technique [1].

Some papers also proposed a hybrid model which works with a new approach which was based on more than just one model like this model named Layout Transformer (LiLT) for structured document understanding proposed in [5], To obtain the text bounding boxes and contents from an input document image, they made use of commercial OCR engines.

The relevant Transformer-based architecture is then fed with embedded text and layout data to achieve increased functionality. They also made an introduction of the bi-directional attention complementation mechanism (BiACM) which enables the cross-modal interplay of text and layout cues. The pre-training of this model was done on the IIT-CDIP dataset available which consists of 6 million documents and more than 11 million scanned documents [5].

In another research, the paper proposed made a comprehensive analysis between two models namely LayoutLM and Donut. The results showed that the implementation of the LayoutLM performed better with an accuracy of 0.88 than the donut, which obtained 0.74, the comparison was based on the performances of the LayoutLM and Donut in terms of analysis of image classification on different datasets [4].

In paper [6] , They proposed a Visual Grid Transformer in which they pre-train the model for 2 Dimensional level for tokens and also semantic understanding at the level of segments. This resulted in great results for the document layout analysis with the best results and the model was named D^4 LA.

There is also research that demonstrates the incorporation of both textual and visual input into a single classification job and pre-training this model using the freely accessible dataset RVL-CDIP. They achieved a high accuracy of 93.03% [7].

More research is available in the betterment of Visual Document Analysis like some more papers proposed different hybrid models like the model named SeRum [8], DocFormer [9] and Donut which is a OCR free document understanding transformer based model [10].

III.METHODOLOGY

In this paper, we intend to understand how document classification works with LayoutLM and analyze its results by comparing the old and newer versions. The LayoutLM proposed by [2] is a simple pretrained language model which works effectively on the layout analysis and text for better document layout understanding. The main applications of this language model are form understanding, information extraction from documents which have complex layout, and receipt understanding. Moreover, the recent version of LayoutLMv3 works by a word-patch alignment aim for predicting if the associated image patch of a text word is hidden while learning cross-modal alignment [11].

A. Data Description

The RVL-CDIP dataset [12], which is freely accessible, comprises document images used to test models for document categorization, information retrieval, and question-answering systems [7]. This dataset was involved in training many models including the LayoutLM model. This paper uses only a small subset of this dataset due to technical limitations.

B. Data Preparation

In this study, text is extracted from photographs and converted into machine-readable English using Tesseract OCR, an open-source optical character recognition (OCR) program. It supports a multitude of languages, making it a versatile tool for text recognition tasks. LayoutLmv2 uses text image alignments and text image matching tasks. It also used the attention mechanism in its transformer architecture which allows the model to understand the relationships between the blocks of text [13].

Hugging Face Transformers Library: Various parameters are used from the transformer library to preprocess the data before the extraction of features. The padding parameters can be used to pad the inputs to a maximum length and ensure that they are all the same length. The truncation parameter can be set to True, which means that if the input data is longer than the specified maximum length, it will be truncated. Label Indexing is done with the help of two dictionaries idx2label and label2idx. Which results in mapping between class labels and their corresponding numerical indices can be done. This indexing is commonly used when preparing data for machine learning and deep learning models, as models work with numerical labels.

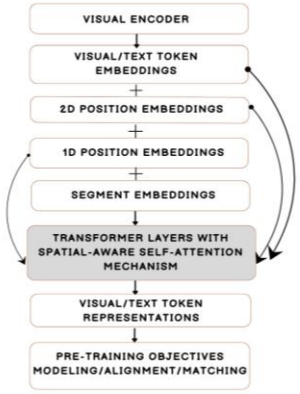

VISUAL ENCODER
VISUAL/TEXT TOKEN EMBEDDINGS
+
2D POSITION EMBEDDINGS
+
1D POSITION EMBEDDINGS
+
SEGMENT EMBEDDINGS
TRANSFORMER LAYERS WITH SPATIAL-AWARE SELF-ATTENTION MECHANISM
VISUAL/TEXT TOKEN REPRESENTATIONS
PRE-TRAINING OBJECTIVES MODELING/ALIGNMENT/MATCHING

Figure 1. Flow Diagram representing the workflow of LayoutLMv2 [13]

C. Model Architecture

The model basically uses the same architecture in both versions, is that version 2 additionally uses the tesseract OCR open engine and the incorporation of visual embeddings during pre-training in LayoutLMv2. This change suggests an enhancement in leveraging visual information from the document images earlier in the model's training process.

Since the LayoutLM uses transformer architecture and accepts input in the form of image, text, and layout embeddings. It also employs the self-attention technique for improved modelling.

1) ***Text Embeddings*:** These embeddings use models like BERT or RoBERTa to capture semantics and context while encoding textual material.

2) ***Visual Embeddings:*** To comprehend the visual structure of the document, visual aspects like layout and design are retrieved using convolutional neural networks (CNNs).

3) ***Layout Embeddings:*** To aid in exact location comprehension, these embeddings concentrate on spatial information, such as bounding boxes and coordinates.

After merging text, visual, and layout information through embeddings, the model generates feature representation. Depending on the applicable assignment, this full feature representation is subsequently used to either token-level classification or element-level classification. A loss function directs the training process when predictions from the model are contrasted with labels from the ground truth. Once trained, the model is assessed and adjusted as needed to get the best performance. The model is now prepared for inference, which involves analysing newer documents and speculating on their structure and content or categorizing them. This process makes sure that LayoutLMv2 is exceptional in analysing document layouts and associated duties.

Document images use the now trained model for predicting outcomes. The model makes a classification prediction for the document, and the outcomes are converted into class labels and corresponding probabilities. This approach enables the model to categorize the document into the appropriate classifications based on its content.

IV. Results

This experiment evaluates the performances of LayoutLM version 1 and 2. The first version uses the visual embeddings at the stage of fine tuning whereas version 2 adds the embeddings at the time of pre training.

The metrics used to calculate the values in table 1 are listed here:

$$\text{Accuracy} = (TP+TN)/(TP+FN)$$

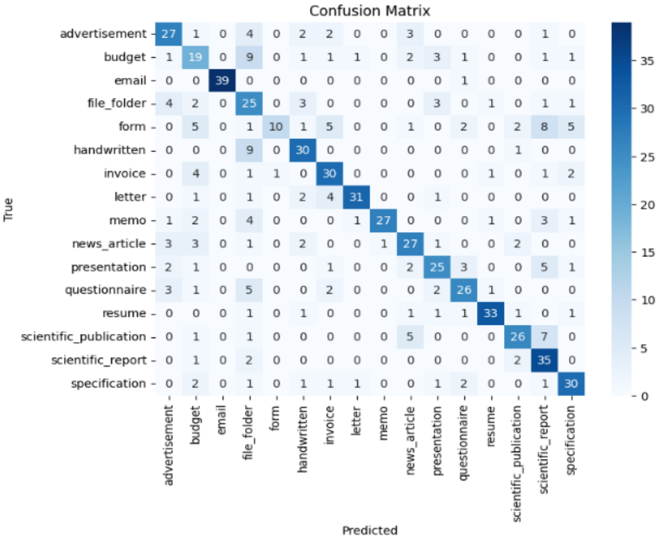
$$\text{Precision} = TP/(TP+FN)$$

$$\text{Recall} = TP/(TP+FP)$$

$$\text{F1Score} = 2*(\text{Precision}*\text{Recall})/(\text{Precision}+\text{Recall})$$

LayoutLM	*Precision*	*Recall*	*F1-score*	*Accuracy*
Version 1	0.73	0.69	0.69	68.75
Version 2	0.79	0.71	0.75	**80.94**

Table 1 Comparison of LayoutLM V1 and V2



	advertisement	budget	email	file_folder	form	handwritten	invoice	letter	memo	news_article	presentation	questionnaire	resume	scientific_publication	scientific_report	specification
advertisement	27	1	0	4	0	2	2	0	0	3	0	0	0	0	1	0
budget	1	19	0	9	0	1	1	1	0	2	3	1	0	0	1	1
email	0	0	39	0	0	0	0	0	0	0	0	1	0	0	0	0
file_folder	4	2	0	25	0	3	0	0	0	0	3	0	1	0	1	1
form	0	5	0	1	10	1	5	0	0	1	0	2	0	2	8	5
handwritten	0	0	0	9	0	30	0	0	0	0	0	0	0	1	0	0
invoice	0	4	0	1	1	0	30	0	0	0	0	0	1	0	1	2
letter	0	1	0	1	0	2	4	31	0	0	1	0	0	0	0	0
memo	1	2	0	4	0	0	0	1	27	0	0	0	1	0	3	1
news_article	3	3	0	1	0	2	0	0	1	27	1	0	0	2	0	0
presentation	2	1	0	0	0	0	1	0	0	2	25	3	0	0	5	1
questionnaire	3	1	0	5	0	0	2	0	0	0	2	26	1	0	0	0
resume	0	0	0	1	0	1	0	0	0	1	1	1	33	1	0	1
scientific_publication	0	1	0	1	0	0	0	0	0	5	0	0	0	26	7	0
scientific_report	0	1	0	2	0	0	0	0	0	0	0	0	0	2	35	0
specification	0	2	0	1	0	1	1	1	0	0	1	2	0	0	1	30

Figure 2. Confusion Matrix for LayoutLMv1

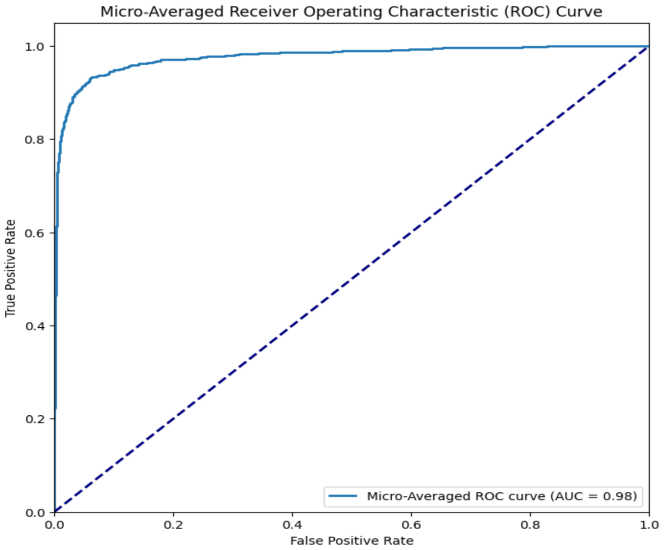
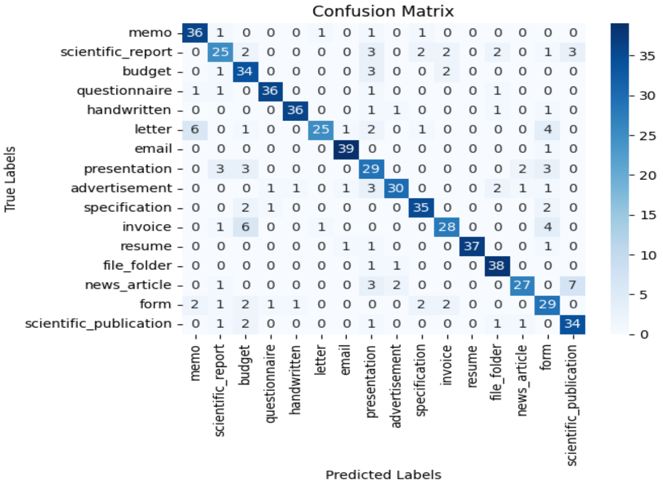


Figure 3. Micro Averaged ROC curve for LayoutLMv2



	memo	scientific_report	budget	questionnaire	handwritten	letter	email	presentation	advertisement	specification	invoice	resume	file_folder	news_article	form	scientific_publication
memo	36	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0
scientific_report	0	25	2	0	0	0	0	3	0	2	2	0	2	0	1	3
budget	0	1	34	0	0	0	0	3	0	0	2	0	0	0	0	0
questionnaire	1	1	0	36	0	0	0	1	0	0	0	0	1	0	0	0
handwritten	0	0	0	0	36	0	0	1	1	0	0	0	1	0	1	0
letter	6	0	1	0	0	25	1	2	0	1	0	0	0	0	4	0
email	0	0	0	0	0	0	39	0	0	0	0	0	0	0	1	0
presentation	0	3	3	0	0	0	0	29	0	0	0	0	0	2	3	0
advertisement	0	0	0	1	1	0	1	3	30	0	0	0	2	1	1	0
specification	0	0	2	1	0	0	0	0	0	35	0	0	0	0	2	0
invoice	0	1	6	0	0	1	0	0	0	0	28	0	0	0	4	0
resume	0	0	0	0	0	0	1	1	0	0	0	37	0	0	1	0
file_folder	0	0	0	0	0	0	0	1	1	0	0	0	38	0	0	0
news_article	0	1	0	0	0	0	0	3	2	0	0	0	0	27	0	7
form	2	1	2	1	1	0	0	0	0	2	2	0	0	0	29	0
scientific_publication	0	1	2	0	0	0	0	1	0	0	0	0	1	1	0	34

Figure 4. Confusion Matrix for LayoutLMv2

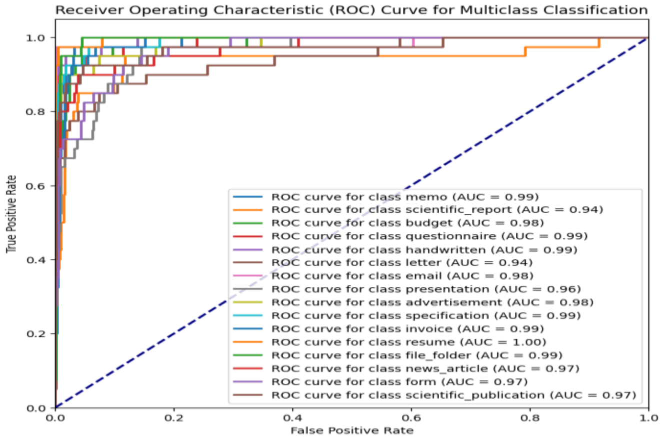


Figure 5 ROC curve for Multiclass classification using LayoutLMv2

In figures 3 and 5, the micro-averaged ROC curve and ROC curve for multiclass classification for all the classes present in the dataset using the LayoutLMv2 can be seen respectively.

V. Conclusion and Future work

We employed two versions, which are 1 and 2 of the LayoutLM model and deployed various parameters from the transformer library to prepare the data extracted from the document images. The version 1 of this model achieved an overall accuracy of 68.75% whereas, the Version 2 of this model which uses Tesseract OCR, an open engine to convert the scanned documents into encodings which the machine can read achieved the overall accuracy of 80.94%. We conclude that the LayoutLMv2 model has better performance for many tasks relating to document AI through the VRDs (Visually Rich Documents) compared to the initial version of the same model as it uses not only visual embeddings but also the spatial attributes which results in much better performance in various applications of document AI. Moreover, with the huge trend in digitizing the documents this technology surely has a large room to grow in through creating hybrid approaches to enhance accuracy from it. Future research may use this model's third version and examine how it has improved by utilizing different strategies.

Acknowledgment

This research was supported by the MSIT (Ministry of Science ICT), Korea, under the National Program for Excellence in SW, supervise by the IITP (Institute of Information & Communications Technology Planning & Evaluation) in 2022 (2022-0-01091, 1711175863)

References

[1] Boyang Wang and Lin Zhang, “Research on Document Oriented Entity Recognition,” Int. Core J. Eng., vol. 8, no. 5, May 2022, doi: 10.6919/ICJE.202205_8(5).0082.

[2] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, “LayoutLM: Pre-training of Text and Layout for Document Image Understanding,” in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, in KDD ’20. New York, NY, USA: Association for Computing Machinery, Aug. 2020, pp. 1192–1200. doi: 10.1145/3394486.3403172.

[3] A. Athar, S. Ali, S. Bhattacharjee, S. A. Shigri, H.-C. Kim, and M. Sheeraz, "Sentimental Analysis of Movie Reviews using Soft Voting Ensemble-based Machine Learning," in Proceedings of the 2022 IEEE/ACM International Conference on Social Networks and Applied Sciences (SNAMS), 2022, doi: 10.1109/SNAMS53716.2021.9732159.

[4] M. Bajrami, E. Zdravevski, P. Lameski, and B. Stojkoska, “A Comprehensive Analysis of LayoutLM and Donut for Document Classification,” Jul. 2023, Accessed: Oct. 11, 2023. [Online]. Available: https://repository.ukim.mk:443/handle/20.500.12188/27397

[5] J. Wang, L. Jin, and K. Ding, “LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding.” arXiv, Feb. 28, 2022. doi: 10.48550/arXiv.2202.13669

[6] C. Da, C. Luo, Q. Zheng, and C. Yao, “Vision Grid Transformer for Document Layout Analysis.” arXiv, Aug. 28, 2023. doi: 10.48550/arXiv.2308.14978.

[7] T. Dauphinee, N. Patel, and M. Rashidi, “Modular Multimodal Architecture for Document Classification.” arXiv, Dec. 09, 2019. doi: 10.48550/arXiv.1912.04376.

[8] H. Cao et al., “Attention Where It Matters: Rethinking Visual Document Understanding with Selective Region Concentration.” arXiv, Sep. 03, 2023. doi: 10.48550/arXiv.2309.01131.

[9] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, “DocFormer: End-to-End Transformer for Document Understanding.” arXiv, Sep. 20, 2021. doi: 10.48550/arXiv.2106.11539.

[10] G. Kim et al., “OCR-free Document Understanding Transformer.” arXiv, Oct. 06, 2022. doi: 10.48550/arXiv.2111.15664.

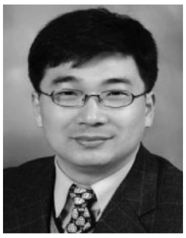
[11] “LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking | Proceedings of the 30th ACM International Conference on Multimedia.” Accessed: Oct. 11, 2023. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3503161.3548112

[12] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 2015, pp. 991-995, doi: 10.1109/ICDAR.2015.7333910.

[13] Y. Xu et al., “LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding.” arXiv, Jan. 09, 2022. doi: 10.48550/arXiv.2012.14740.

Kounen Fathima received her BE in IT Engineering from the Osmania University, India. Currently, she is pursuing her master’s degree in the Institute of Digital Anti-Aging Healthcare from Inje University. Her research interest aligns with Artificial Intelligence, Machine learning and Natural Language Processing.

Ali Athar received his BSSE degree Software Engineering from Government College University Faisalabad (GCUF), Pakistan. He received his MS degree from NUST, Pakistan. He is pursuing his Ph.D. degree from the Institute of Digital Anti-aging and healthcare at Inje University. His research areas include Text Mining, Machine learning, and Deep learning.



Hee-Cheol Kim received his BSc at the Department of Mathematics, MSc at the Department of Computer Science in Sogang University in Korea, and Ph.D. at Numerical Analysis and Computing Science, Stockholm University in Sweden in 2001. He is a Professor at the Department of Computer Engineering and Head of the Institute of. Digital Anti-aging Healthcare, Inje University in Korea. His research interests include machine learning, deep learning, Computer vision, and medical informatics.

Machine Learning Based Techniques for the Prediction of Axillary Lymph Node Metastases in Early Breast Cancer

Maisam Ali*, Muhammad Yaseen *, Sikandar Ali, Hee-Cheol Kim *

**Dept. of Digital Anti-Aging Healthcare, Inje University Gimhae, Republic of korea*

maisamali053@gmail.com, shigriyaseen@gmail.com, sikandershigri77@gmail.com heeki@inje.ac.kr

***Abstract*— One of the most significant considerations in determining the prognosis of breast cancer is the involvement of lymph nodes. Although non-invasive imaging techniques like ultrasound, computed tomography (CT), magnetic resonance imaging (MRI), and F-18 fluoro-2-deoxy-D-glucose (FDG) positron emission tomography (PET)/CT have been recommended for the assessment of the ALN status, their diagnostic performance lacks sufficient sensitivity in the detection of ALN metastasis. Accurate machine learning based techniques have been applied to the investigation of lymph node in the early stage of breast cancer. This occurrence is influenced by several tumor-specific elements. The primary objective of this research was to determine the clinical and pathological variables that validate the radiomics-based machine learning model can be used to predict whether individuals with early-stage breast cancer will have metastases in their axillary lymph nodes (ALNM). Various aspects were considered and analyzed while utilizing various machine learning algorithms to determine the involvement of the lymph node metastasis. For our experiments, we have examined three machine learning models such as Gradient Boosting Machine (GBM), KNNeighbor (KN) and Decision Tree (DT). Accuracy, specificity, sensitivity, and F1-score were used to evaluate the performance of all three models. The accuracy of GBM was 97%, followed by that of KNNeighbor (KNN) and Decision Tree (DT), which were 91% and 91%, respectively.**

***Keywords*— Lymph node metastasis, Breast cancer, Machine learning, Deep learning, Gradient Boosting Machine (GBM)**.

I. INTRODUCTION

Despite acknowledging that breast cancer is a heterogeneous disease with diverse prognoses, developments in research and public health initiatives regarding cancer patient identification and therapy have increased survival rates globally [1]. One of the most significant prognostic factors influencing the long-term prognosis of breast cancer is an accurate preoperative assessment of axillary lymph-node (ALN) status. Although non-invasive imaging methods such as ultrasonography, computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET)/CT are available to evaluate the status of axillary lymph nodes (ALNs), these techniques are insufficiently sensitive to identify ALN metastasis. In the 19th century, the biological concept of breast cancer predominated. It described the disease as a local occurrence inside the breast that propagated centrifugally along lymphatics to first and then second echelon nodes, with increasing systemic metastasis. A beneficial outcome of this strategy was radical local surgery, or the Halstead radical mastectomy [2]. To correctly stage the patient and achieve adequate local control, axillary lymph node dissection (ALND), which determines the ALN status, has been the standard of care for patients with invasive breast cancer. Except for T4 tumours, sentinel lymph node biopsy (SLNB), an alternative to ALND, has recently come to be the norm for axillary staging in all clinically node-negative patients (as determined by clinical examination, ultrasonography, and/or fine needle aspiration cytology). The staging of the axilla using the SLN approach has been shown to be possible, accurate, and adequate while avoiding the morbidity of an ALND. [3,4]. ALN dissection has been commonly used for the comprehensive examination of ALN status. However, it is unnecessary in patients without ALN metastases and can have serious adverse effects like lymphedema. Although sentinel lymph node biopsy, the standard procedure for determining the status of the axillary lymph nodes in patients with early breast cancer, has been associated to a false-negative rate of 5–10% [5]. The purpose of this study is to assess the probability of breast cancer will spread to the axillary lymph node. The method carried out by implementing the most well-known machine learning algorithms can produce an individual risk assessment for every patient, which helps doctors in making shrewd decisions about minimizing consequences from the spread of cancer to the lymph nodes and improving treatment according to each patient's specific needs.

II. RELATED WORKS

Several research attempts have been made by researchers to investigate axillary lymph node metastasis. All patients had managed to keep their blood glucose levels below 8.3 mmol/L (150 mg/dL) and had fasted for at least six hours before receiving F-18 FDG. The intravenous dose of 5.5 MBq/kg of F-18 FDG was followed by the PET/CT scan 60 minutes later. A

Discovery STE PET/CT scanner from GE Healthcare, Milwaukee, Wisconsin, was used to do the scans within a month of the surgical procedure. The cranial vertex to the proximal thighs were first imaged with a low-dose CT scan (peak voltage of 120 kVp, automatic tube current of 60–150 mA, and slice thickness of 3.75 mm) without contrast enhancement for attenuation correction. Following the acquisition of the CT scan, a 3-dimensional PET scan was conducted with an acquisition period of 3 minutes per bed position [6]. A retrospective review of all F-18 FDG PET/CT images was conducted using the AW server 3.2 (GE Healthcare, Milwaukee, WI, USA). To accurately identify genuine IDC lesions, the F-18 FDG PET/CT images were visually analyzed and compared with breast MRI images. The maximal SUV (SUVmax) was calculated using the following formula: SUVmax = maximum activity in region of interest (MBq/g)/ [injected dosage (MBq)/body weight (g)]. In cases of ALN, focally enhanced F-18 FDG avid ALNs with a high F-18 FDG uptake (SUVmax 2) were considered favorable findings, as a cut-off of 2.0 has been shown to exhibit the best accuracy in diagnosing ALN metastases in breast cancer [7]. The results showed that F-18 FDG PET/CT had a sensitivity of 55.8%, specificity of 93.0%, accuracy of 77.0%, positive predictive value (PPV) of 85.7%, and negative predictive value (NPV) of 73.6% for detecting ALN metastases.

III. MATERIALS AND METHODS

This section describes the materials and methods we used in our experiment.

A. *Data Source*

We used a data set from DACON for this research study. The data set contains 549 patients' information with 28 different features.

B. *Data Preparation*

Data preparation is one of the key methods used in deep learning and machine learning. The degree to which the data has been pre-processed typically determines how accurate your model will be. In other words, the quality of your model is strongly correlated with the preprocessing of the data. Our dataset contained categorical and non-categorical values as well as null values and outliers. We eliminated outliers and null values. The categorical values and non-categorical values were converted using one-hot encoding and label encoding. We obtained balanced and clean data for further processing after the dataset underwent preprocessing. So that models could be used, we converted the raw data into an appropriate form. To choose the most pertinent features, we used an additional technique for tree feature selection. In our experiment, we used 21 attributes to facilitate the implementation of different machine learning models and to further develop explainable AI for the support and better representation of this research study. Table 1 below displays the features information and descriptions.

TABLE 1. Feature descriptions of the dataset

	Features	**Feature type**
1	Age	Continuous
2	Diagnostic Name	Categorical
3	Cancer Location	Continuous
4	Cancer Size	Continuous
5	NG	Continuous
6	HG	Continuous
7	HG_score_1	Continuous
8	HG_score_2	Continuous
9	HG_score_3	Continuous
10	DCIS_or_LCIS	Categorical
11	DCIS_or_LCIS_type	Categorical
12	N_category	Categorical
13	ER	Categorical
14	ER_Allred_score	Continuous
15	PR	Categorical
16	PR_Allred_score	Continuous
17	KI-67 LI percent	Continuous
18	HER2	Categorical
19	HER2_IHC	Categorical
20	HER2_SISH,	Categorical
21	HER2_SISH_ratio	Categorical

C. *Machine Learning Algorithms*

For the development of our prediction model, we used three different machine learning models. Based on input features, a lymph node metastatic decision tree classifier determines if a patient has lymph node metastasis. The decision tree algorithm develops a set of rules to categorize patients into two groups based on factors including tumour size, tumour location, and patient age. These groups are either patients with lymph node metastases or not. With a combination of decision trees that determine the mode or mean/average of each tree, RF is an ensemble learning approach for classification, regression, and other use cases [8]. The Gradient Boosting Machine (GBM) [9], by iteratively training decision trees to compensate for prediction errors, GBM for lymph node metastasis forecasts the likelihood of metastasis. Starting with a single prediction, it then aggregates forecasts, fits a tree to residuals, and repeats. The model is improved through this iterative process, which captures complicated patterns and yields a reliable assessment of the chance of lymph node metastasis. In the evaluation of lymph node metastasis, K-Nearest Neighbours (KNN) [10], predicts if a patient has metastasis by seeking the 'k' nearest patients in the dataset with comparable variables (such as tumour size, grade, etc.). The assumption is based on the dominant class among these neighbours. KNN is non-parametric, logical, and makes few assumptions about the data, enabling simple interpretation. In the case of lymph node metastasis investigations, however, choosing an adequate "k" value and distance metric is essential for reliable predictions. We compared these models' performance metrics, including accuracy, specificity, sensitivity, and F1-score.

D. *Performance Measures*

The performance of machine learning models can be evaluated using certain metrics and instruments. The performance of models is typically measured using metrics like accuracy, specificity, sensitivity, and F1-score. These criteria were also used to assess how well our models performed, and the findings are fairly positive. Accuracy is the degree to which measurement results are close to the actual value. In other words, it is only the proportion of observations that were correctly predicted to all the observations. The percentage of relevant occurrences among the examples that were retrieved is known as specificity, which is exactness. The ratio of accurately anticipated positive observations to all the actual class observations is known as sensitivity. The sensitivity and specificity weighted average is known as the F1-score.

Accuracy=TP+ TN / TP + FP + FN + TN (1)

$Specificity = TN / TN + FP$ (2)

$Sensitivity = TP / TP + FN$ (3)

$F1{-}Score = 2(Recall * Precision) / (Recall + Precision)$ (4)

Where TP, TN, FP, FN denoted as true positive, true negative, false positive, false negative, respectively.

IV. OVERALL EXPERIMENTAL WORKFLOW

In this research, we initially retrieved the data from the Dacon dataset and performed preprocessing to remove outliers and null values from the data. After eliminating data abnormalities, we received 549 records. When preprocessing the data, we used label encoding and one hot encoding. We also chose 21 features for our experiment after selecting other features. The data was then divided into 20% for testing and 80% for training. We used Decision Tree, KNNeighbor, and Gradient Boosting Machine. It was determined how to measure performance using metrics like accuracy, sensitivity, specificity, and F1-scores. In this study, the performance measures' graphical representation and ROC were also computed. Fig. 1 displays the overall process used in this research.

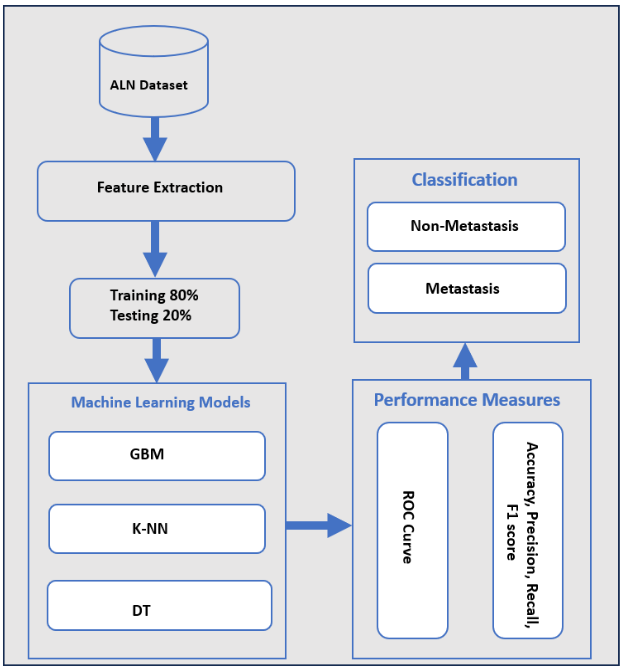


Figure 1. Overall workflow of the classification of the axillary lymph node metastasis.

V. RESULTS AND DISCUSSION

After removing the data that had missing values and did not satisfy the criteria set out by our models, we included 549 patient records in the analysis. After preprocessing the data, we chose 21 features for our investigation. A training set comprised 80% of the dataset, and a testing set comprised 20%. The gradient boosting machine, KNneighbor, and decision tree were the three machine learning models we used. Interesting findings were obtained. Each model's accuracy, specificity, and recall were calculated. The performance metric for each of the three used machine learning algorithms is displayed in Table. 2

Table.2 The Overall Performance of all the Three Applied Machine Learning Models being Used for the Prediction and Classification of ALN metastasis.

Classifiers	Accuracy	Specificity	Sensitivity	F1 score
GBM	0.97	0.93	0.89	0.91
KNN	0.91	0.96	0.67	0.79
DT	0.91	0.90	0.87	0.88

In this study, we trained the machine learning ML models to predict the ALN metastasis of breast cancer patients using clinical informations. The GBM model outperformed the other three machine learning models, with an AUC of 0.97%, demonstrating strong discrimination performance. Fig 2 shows the ROC curve of all the three applied machine learning algorithms.

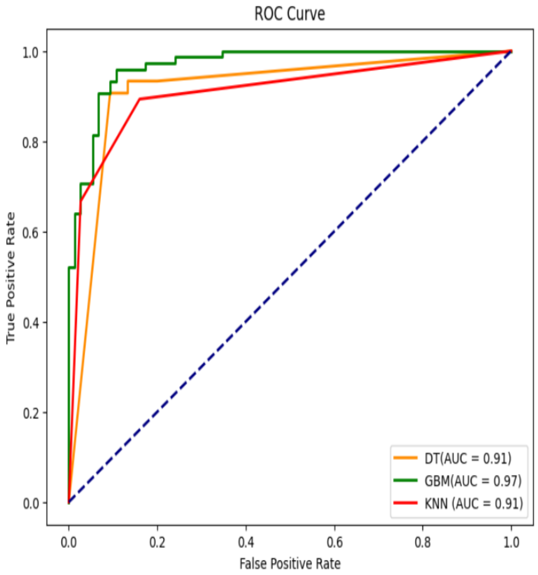


Figure 2. Receiver operating characteristics (ROC) of Classification Models

VI. Conclusion

In this study, the classification, and the prediction of the involvement of an axillary lymph node metastasis could be precisely assessed by three different machine learning models. In a conclusion, our models offer a precise and understandable prediction result that takes into consideration the distinctive characteristics of each patient and can assist doctors in customizing the treatment of patients with axillary lymph node metastasis in accordance with the patients' specific prognoses.

Acknowledgment

This research was supported by a grant from Korea Health Technology R&D Project through the Korea Health Industry Development Institute validation (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant No: HI21C0977).

References

[1] Arnold M, Rutherford MJ, Bardot A, Ferlay J, Andersson TM-L, Myklebust TÅ, et al. Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (ICBP SURVMARK-2): a population-based study. *Lancet Oncol.* (2019) 20:1493–505. doi: 10.1016/S1470-2045(19)30456-5

[2] Fisher, B. "Seminars of Bernard Fisher 1960-nature of cancer as systemic disease." *Bull Soc Int Chir* 31.6 (1972): 604-9.

[3] Peeters, M., et al. "Guideline for the esophageal and gastric cancer: scientific support of the College of Onclogy." *Good Clinical Practice (GCP). Brussels: Belgian Health Care Knowledge Centre (KCE)* 21.03 (2008): 2008.

[4] Rietman, J. S., Dijkstra, P. U., Geertzen, J. H., Baas, P., De Vries, J., Dolsma, W., ... & Hoekstra, H. J. (2003). Short-term morbidity of the upper limb after sentinel lymph node biopsy or axillary lymph node dissection for Stage I or II breast carcinoma. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *98*(4), 690-696.

[5] Krag DN, Anderson SJ, Julian TB, Brown AM, Harlow SP, Ashikaga T, et al. Technical outcomes of sentinel-lymph-node resection and conventional axillary-lymph-node dissection in patients with clinically node-negative breast cancer: results from the NSABP B-32 randomised phase III trial. Lancet Oncol. 2007;8:881–8.Song, Bong-Il. "A machine learning-based radiomics model for the prediction of axillary lymph-node metastasis in breast cancer." *Breast Cancer* 28 (2021): 664-671.

[6] Taira N, Ohsumi S, Takabatake D, Hara F, Takashima S, Aogi K, et al. Determination of indication for sentinel lymph node biopsy in clinical node-negative breast cancer using preoperative 18F-fluorodeoxyglucose positron emission tomography/computed tomography fusion imaging. Jpn J Clin Oncol. 2009;39:16–21.

[7] Carkaci S, Adrada BE, Rohren E, Wei W, Quraishi MA, Mawlawi O, et al. Semiquantitative analysis of maximum standardized uptake values of regional lymph nodes in inflammatory breast cancer: is there a reliable threshold for differentiating benign from malignant? Acad Radiol. 2012;19:535–41.

[8] Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32. [Google Scholar] [CrossRef]

[9] Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [Google Scholar]

[10] Xu, Z., Xie, Y., Wu, L., Chen, M., Shi, Z., Cui, Y., ... & Liu, Z. (2023). Using Machine Learning Methods to Assess Lympho vascular Invasion and Survival in Breast Cancer: Performance of Combining Preoperative Clinical and MRI Characteristics. *Journal of Magnetic Resonance Imaging*.

Mr. Maisam Ali received his B.E degree in Electrical and communication Engineering from Hamdard University, Pakistan. He is currently pursuing his master's degree from Inje University, South Korea. His research interests are artificial intelligence, machine learning, deep learning, computer vision.

Mr. Muhammad Yaseen received his B.E degree in Electrical Engineering from Hamdard University, Pakistan. He is currently pursuing his master's degree from Inje University, South Korea. His research interests are artificial intelligence, machine learning, deep learning, computer vision.

Mr. Sikandar Ali received his B.E. degree in Computer Engineering from Mehran University of Engineering & Technology, Pakistan. He got his MS from the Department of Computer Science from Chung Buk National University, the Republic of Korea. Furthermore, he is now a Ph.D. candidate at Inje University South Korea in the department of Digital Anti-Aging Healthcare. His research interests include artificial intelligence, data science, big data, machine learning, deep learning, reinforcement learning, Computer vision, and medical imaging.

Prof. Hee-Cheol Kim received his BSc at the Department of Mathematics, MSc at the Department of Computer Science in Sogang University in Korea, and Ph.D. at Numerical Analysis and Computing Science, Stockholm University in Sweden in 2001. He is a Professor at the Department of Computer Engineering and Head of the Institute of Digital Anti-aging Healthcare, Inje University in Korea. His research interests include machine learning, deep learning, Computer vision, and medical informatics.

AI-based logistics system overview and a workflow for digital freight forwarding in logistics.

Md Ariful Islam Mozumder*, Rashedul Islam Sumon*, Ziaullah Khan*, Shah Muhammad Imtiyaj Uddin*,
Muhammad Omair Khan* Hee Cheol Kim*

Department of Computer Engineering/Institute of Digital Anti-Aging Healthcare/u-HARC, Inje University, South Korea
arifulislamro@gmail.com, sumon39.cst@gmail.com, zkhan.msee19seecs@seecs.edu.pk, imtiyaj.dream@gmail.com,
mumairkhan690@gmail.com, heeki@inje.ac.kr

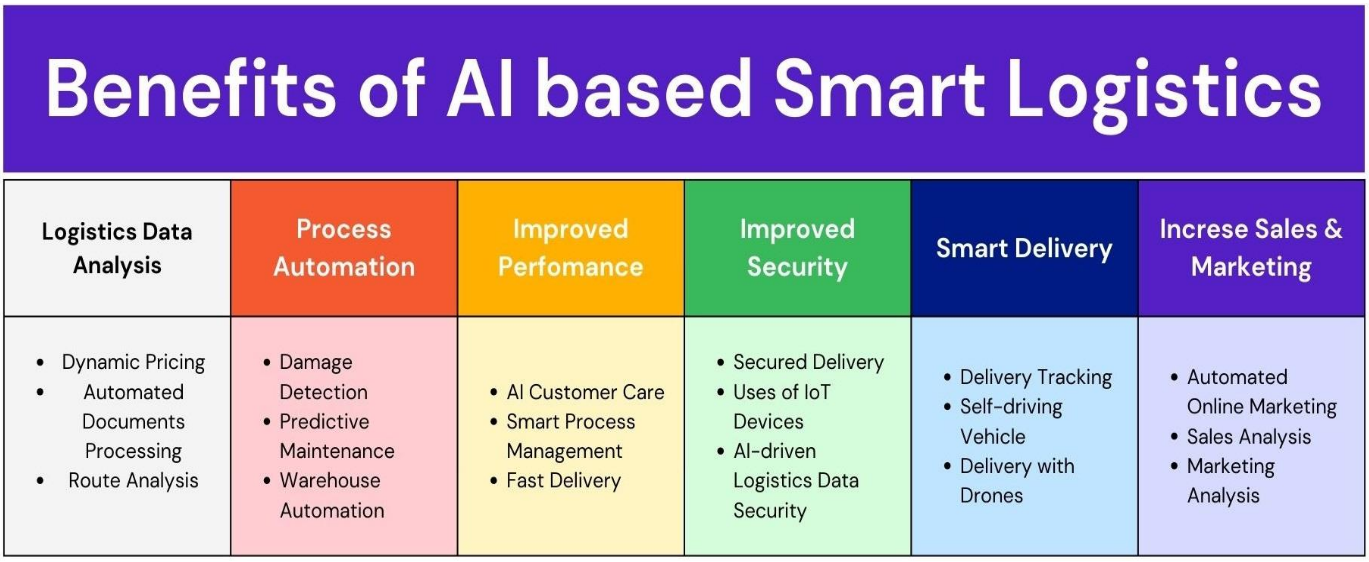
***Abstract*— In the realm of global business, logistics stands out as a cornerstone, and the ongoing development of Artificial Intelligence (AI) is shaping logistics into a secure and intelligent domain. The digitization of freight forwarding involves converting traditional logistic procedures into streamlined, digitized processes within the freight forwarding system. This paper provides a concise exploration of AI applications in logistics systems, delving into the transformative impact on supply chain management. Focusing on key components such as machine learning and predictive modeling, it offers a brief overview of AI's role in optimizing logistics processes and enhancing efficiency. Also, we will show a framework for digital freight forwarding in logistics considering AI applications with the explanation.**

***Keywords*— Artificial intelligence, IoT, supply chain, logistics, and digital freight forwarding.**

I. Introduction

In recent years, the synergy between Artificial Intelligence (AI) and logistics systems has propelled the evolution of supply chain management into uncharted territory. The integration of AI technologies, ranging from advanced machine learning algorithms to natural language processing and computer vision, has ushered in a new era of efficiency, adaptability, and precision within the logistics domain [1]. Beyond the technical aspects, we delve into the ethical considerations inherent in AI adoption within logistics, contemplating issues of data privacy, algorithmic bias, and the imperatives of transparent decision-making [2]. In the dynamic landscape of contemporary logistics, the synergy between Artificial Intelligence (AI) and the pervasive trend of digitalization is catalyzing a profound transformation. The integration of AI in logistics represents a pivotal evolution, offering unprecedented opportunities to enhance efficiency, accuracy, and adaptability across the supply chain. From real-time data analytics to predictive modeling, AI applications permeate every facet of logistics, revolutionizing decision-making processes and ushering in a new era of intelligent logistics systems [3]. Simultaneously, the conceptual framework outlined herein addresses the digital transformation of freight forwarding critical component within the logistics ecosystem. As traditional procedures undergo a paradigm shift towards digital processes, the framework serves as a guiding structure, illuminating the intricate web of technological advancements, connectivity, and streamlined operations that characterize the envisioned future of freight forwarding [4]. There are several benefits for AI-based smart logistics systems, in figure 1 we have shown some core benefits [5].

By weaving together these strands, this overview aspires to furnish researchers, practitioners, and policymakers with a



Benefits of AI based Smart Logistics

Logistics Data Analysis	Process Automation	Improved Perfomance	Improved Security	Smart Delivery	Increse Sales & Marketing
• Dynamic Pricing • Automated Documents Processing • Route Analysis	• Damage Detection • Predictive Maintenance • Warehouse Automation	• AI Customer Care • Smart Process Management • Fast Delivery	• Secured Delivery • Uses of IoT Devices • AI-driven Logistics Data Security	• Delivery Tracking • Self-driving Vehicle • Delivery with Drones	• Automated Online Marketing • Sales Analysis • Marketing Analysis

Figure 1. The benefits of smart logistics based on artificial intelligence

comprehensive understanding of the current state and future trajectories of AI-based logistics systems. Through this exploration, we aim to contribute to the ongoing dialogue on the harmonious integration of artificial intelligence into the intricate tapestry of modern supply chain management. This paper embarks on a comprehensive exploration of AI-based logistics systems, seeking to elucidate the transformative influence of AI across the entire logistics spectrum. From the rudimentary components of AI implementation to its sophisticated applications in demand forecasting, route optimization, and inventory management, we dissect the mechanics behind the efficiency gains witnessed in modern logistics. As we traverse the logistics lifecycle, the integration of AI in transportation, warehousing, and last-mile delivery emerges as a critical focal point, reshaping traditional methodologies and mitigating challenges that have long plagued supply chain operations. On the other hand, we will provide a conceptual framework for digital freight forwarding in logistics.

II. AI IN SMART LOGISTICS

Integrating AI into Smart Logistics is propelling a transformative wave across the logistics and supply chain management landscape. Smart Logistics harnesses the power of the Internet of Things (IoT) by embedding sensors and GPS trackers in various logistical elements, including vehicles, warehouses, and packages. This seamless integration enables the continuous collection of real-time data on location, condition, and environmental factors, forming the bedrock for precise tracking and meticulous inventory management, thereby ensuring goods' quality and timely delivery [6]. The dynamic synergy of AI and machine learning is pivotal in the Smart Logistics ecosystem, where these technologies process the substantial volume of data generated by IoT devices. Through sophisticated analysis, AI optimizes current logistical operations, predicts future demand patterns, refines delivery routes, and enhances stock-level management. This predictive capability becomes a strategic asset, empowering logistics operations to adapt to market changes and meet evolving customer needs with agility and precision [7]. Smart Logistics, therefore, represents more than just optimizing logistics processes; it embodies a paradigm shift toward a holistic, data-driven approach to logistics management. Beyond its operational benefits of enhanced efficiency and cost reduction, Smart Logistics significantly elevates customer satisfaction by providing faster and more reliable delivery services [8]. Below the figure, we have shown the application of AI for logistics with its explanation.

A. AI-based Planning

AI is reshaping the landscape of logistics planning, offering transformative solutions across critical domains. In the realm

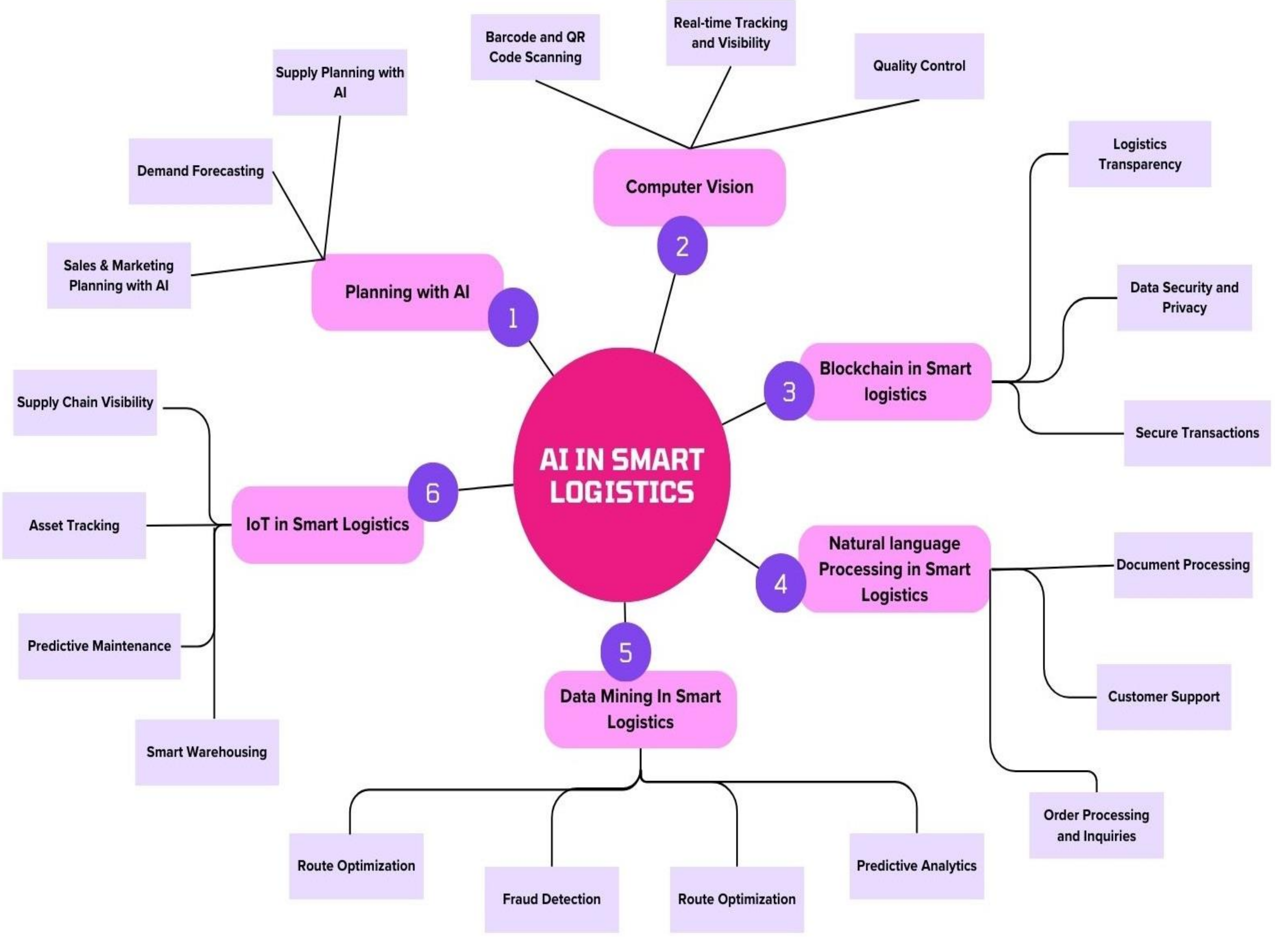


Figure 2. Artificial intelligence application for logistics

of supply planning, AI algorithms analyze historical data, market trends, and supplier behaviors to optimize inventory levels, ensuring a seamless flow of goods. Demand forecasting, another pivotal area, witnesses the power of AI-driven predictive modeling, which anticipates customer needs with unprecedented accuracy, mitigating the risk of overstock or stockouts. Sales and marketing strategies, too, benefit significantly from AI in logistics planning. Through data-driven insights, AI refines sales projections, allowing for dynamic adjustments based on market fluctuations.

B. Computer Vision

Computer Vision is reshaping logistics with precision and speed. Barcode and QR Code Scanning by advanced vision algorithms e.g., Deep Learning-based Image Recognition and Optical Character Recognition, streamline inventory checks, reducing errors and boosting efficiency [9]. Real-time Tracking and Visibility get a significant upgrade with Computer Vision. Automated tracking of shipments and goods provides instant insights, empowering quick decisions and enhancing operational transparency e.g., Light Detection and Ranging, Simultaneous Localization and Mapping. Quality Control becomes meticulous as Computer Vision assesses product quality, identifying defects with accuracy e.g., Hyperspectral Imaging, Texture Analysis, and Pattern Recognition.

C. Blockchain

Supply Chain Transparency gets a boost, providing stakeholders with real-time visibility. Traceability and Provenance ensure accurate tracking of product journeys, instilling trust in the supply chain. Secure Transactions form the backbone of blockchain in logistics, preventing tampering and fraud [10]. This decentralized approach fosters resilience and trust in financial processes. Simultaneously, Data Security and Privacy are prioritized, leveraging cryptographic techniques to safeguard sensitive information, and addressing modern concerns around confidentiality.

D. Natural Language Processing

NLP, a cornerstone in smart logistics, redefines customer-centric processes. In Supply Customer Support, chatbots driven by NLP ensure swift and precise assistance. Order Processing benefits as NLP deciphers natural language, automating inquiries and improving responsiveness [11]. Meanwhile, Document Processing sees a boost as NLP extracts vital information from unstructured documents, streamlining logistics documentation for increased efficiency.

E. Data Mining

In smart logistics, data mining is the silent efficiency booster. Supply Predictive Analytics leverages data for proactive inventory management. Route Optimization crunches historical and real-time data for cost-effective logistics. Fraud Detection sharpens security by unveiling irregularities in transactions. For Energy Efficiency, data mining identifies optimization opportunities, aiding informed decisions.

F. IoT in smart Logistics

The Internet of Things (IoT) is a vast network of individually identifiable, highly networked items, including mechanical and digital machinery. These devices can send data across a network. IoT devices can sense, collect, and transfer data via the Internet. IoT has completely changed logistics and transportation fields by connecting large objects. IoT has improved transportation and logistics, including condition monitoring, online tracking, traffic control, avoiding traffic jams, effective supply chains, and prompt decision-making [10]. IoT enables real-time tracking of assets, enhancing their visibility and management. IoT sensors anticipate equipment issues, allowing for proactive maintenance and minimizing disruptions. IoT optimizes warehouse operations, automating processes and improving overall efficiency. IoT provides end-to-end visibility, ensuring real-time insights into the entire supply chain.

III. FRAMEWORK FOR DIGITAL FREIGHT FORWARDING IN LOGISTICS

AI-based freight forwarding in logistics involves the application of AI technologies to optimize and automate various aspects of freight forwarding processes. It includes the

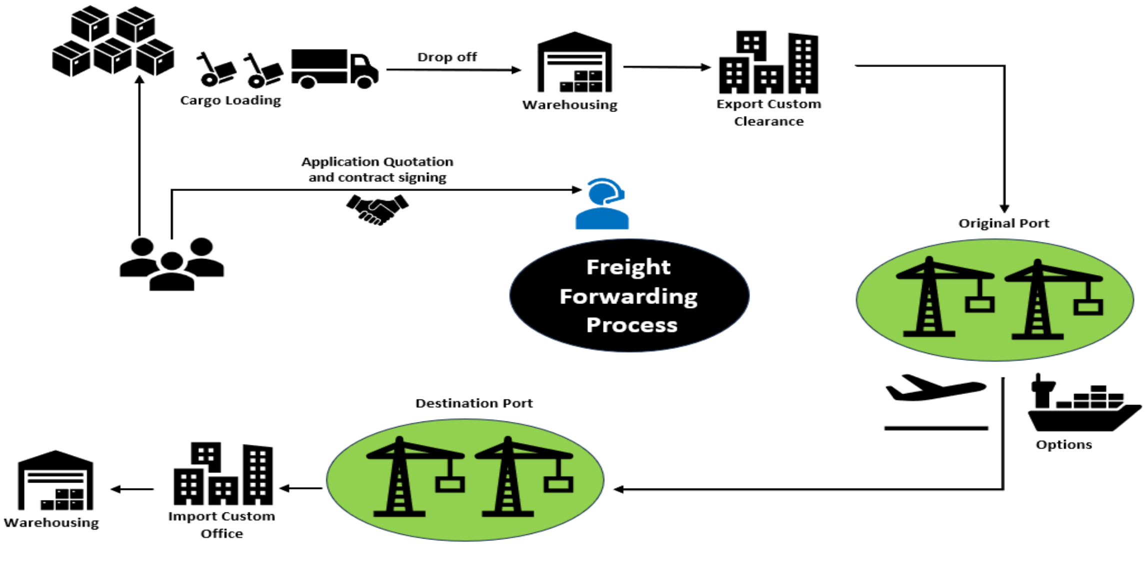


Figure 3. Complete workflow for digital freight forwarding in logistics.

use of AI algorithms for route optimization, demand forecasting, real-time tracking, document processing, and decision-making. AI enhances efficiency, reduces costs, and improves overall operational effectiveness in logistics freight forwarding [12-14]. In Figure 3 we have shown the workflow of AI-based digital freight forwarding.

AI-based automated systems facilitate the loading of goods onto a transport vehicle during the cargo loading phase. Smart tracking algorithms monitor the movement of the truck symbolizing the transportation of loaded cargo to the warehouse. Predictive analytics determine the optimal timing for goods to be ready for shipping, factoring in customs clearance and consolidation needs. Natural Language Processing (NLP) algorithms assist in generating and finalizing quotations and contracts for shipping, ensuring clarity on costs and terms of service. Machine learning models aid in export customs clearance, automatically checking cargo against documentation and ensuring compliance with export regulations. Intelligent data analytics optimize the selection of the original port based on various factors such as transport efficiency and cost-effectiveness. Automation and optimization algorithms drive the freight forwarding process, integrating, and streamlining all preceding steps for efficiency. Predictive modeling anticipates the arrival time and condition of goods at the destination port, facilitating smoother logistics planning. Automated systems assist in import customs clearance by quickly verifying adherence to destination country import regulations upon the goods' arrival. AI-driven logistics planning determines the most suitable warehousing strategy post-clearance, considering delivery timelines and other logistical factors. Decision support systems recommend the optimal mode of transport (air, road, sea) for the final leg based on destination-specific factors and urgency.

IV. Conclusions

The paper provided a comprehensive overview of AI-based logistics systems and introduced a framework for digital freight forwarding within the logistics domain. The synthesis of AI technologies in logistics showcased their transformative impact on efficiency, transparency, and decision-making. The conceptual framework outlined sets the stage for a more streamlined and responsive logistics ecosystem, emphasizing the pivotal role of AI in shaping the future of freight forwarding. This contribution aims to inspire further exploration and integration of AI in logistics, fostering a progressive and technologically advanced industry.

Acknowledgment

This research was supported by the MSIT (Ministry of Science ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) in 2022 (2022-0-01091, 1711175863).

References

[1] Luo, Jinhua. "Research on the Application of Artificial Intelligence in Smart Logistics in Australia." 2023 3rd International Conference on Public Management and Intelligent Society (PMIS 2023). Atlantis Press, 2023.

[2] Balfaqih, Hasan. "Artificial Intelligence and Smart Logistics Systems in Industry 4.0." Proceedings of the International Conference on Industrial Engineering and Operations Management. 2023.

[3] Woschank, Manuel, Erwin Rauch, and Helmut Zsifkovits. "A review of further directions for artificial intelligence, machine learning, and deep learning in smart logistics." Sustainability 12.9 (2020): 3760.

[4] Pan, Shenle, et al. "Smart city for sustainable urban freight logistics." International Journal of Production Research 59.7 (2021): 2079-2089.

[5] (2023) “Top 15 Use Cases and Applications of AI in Logistics in 2023” [Accessed on 25.11.2023]. [Online]. Available https://research.aimultiple.com/logistics-ai/

[6] Efficiency is Key: How Courier Dispatch Systems Can Help You Streamline Your Business | Discover News, Travel, Sports, Fashion, and Events with Fragnewz. https://fragnewz.com/efficiency-is-key-how-courier-dispatch-systems-can-help-you-streamline-your-business/

[7] Zhou, L. and C.X. Lou. Intelligent cargo tracking system based on the internet of things. in 2012 15th International Conference on Network-Based information systems. 2012. IEEE.

[8] Alahi, M.E.E.; Sukkuea, A.; Tina, F.W.; Nag, A.; Kurdthongmee, W.; Suwannarat, K.; Mukhopadhyay, S.C. Integration of IoT-Enabled Technologies and Artificial Intelligence (AI) for Smart City Scenario: Recent Advancements and Future Trends. Sensors 2023, 23, 5206. https://doi.org/10.3390/s23115206

[9] Di Capua, M., A. Ciaramella, and A. De Prisco. "Machine learning and computer vision for the automation of processes in advanced logistics: The integrated logistic platform (ILP) 4.0." Procedia Computer Science 217 (2023): 326-338.

[10] Humayun, M., et al., Emerging smart logistics and transportation using IoT and blockchain. IEEE Internet of Things Magazine, 2020. 3(2): p. 58-62

[11] Garg, Rachit, et al. "i-Pulse: A NLP based novel approach for employee engagement in logistics organization." International Journal of Information Management Data Insights 1.1 (2021): 100011.

[12] Herold, David M., Behnam Fahimnia, and Tim Breitbarth. "The digital freight forwarder and the incumbent: A framework to examine disruptive potentials of digital platforms." Transportation Research Part E: Logistics and Transportation Review 176 (2023): 103214.

[13] Heinbach, Christoph, et al. "Data-driven forwarding: a typology of digital platforms for road freight transport management." Electronic Markets 32.2 (2022): 807-828.

[14] Rana, Sumit Kumar, et al. "Blockchain-based model to improve the performance of the next-generation digital supply chain." Sustainability 13.18 (2021): 10008.

Md Ariful Islam Mozumder is pursuing his Ph.D. in the Institute of Digital Anti-Aging Healthcare at Inje University. He has previously worked on multiple real-life projects related to computer vision, data sciences, Smart IoT systems, and text mining. His research interest aligns with Computer Vision, Medical Image Processing, Metaverse, Artificial Intelligence, Sensor Data Analysis, Bio Signal Processing, Algorithms, and Smart Logistics.

Rashedul Islam Sumon is pursuing his Ph.D. in the Institute of Digital Anti-Aging Healthcare at Inje University. His research interest aligns with Computer Vision, Medical Image Processing, Metaverse, Artificial Intelligence, and Bio Signal Processing.

Ziaullah Khan is a Ph.D. student at the Institute of Digital Anti-aging and Healthcare at Inje University. His research interest area Machine Learning and deep Learning, includes Smart Logistics, Metaverse, and Natural Language Processing.

Shah Muhammad Imtiyaj Uddin received his BSc in Computer Science and engineering from the World University of Bangladesh in 2017. Currently, he is pursuing his Master's in the Institute of Digital Anti-Aging Healthcare from Inje University. His research interests include Computer Vision, Machine Learning, and Deep Learning.

Muhammad Omair Khan is a Ph.D. student at the Institute of Digital Anti-aging and Healthcare at Inje University. His research interest’s area Machine Learning, Deep Learning, including Image Processing, Logistics, and Natural Language Processing.

Hee-Cheol Kim Ph.D. at Numerical Analysis and Computing Science, at Stockholm University in Sweden. He is a professor and Head of Institute of the Digital Anti-aging Healthcare, Inje University, South Korea. His research interests include Machine learning, Text mining, Bioinformatics, Blockchain, Metaverse, and XAI for Finance.

The benefits of integrating AI, IoT, and Blockchain in healthcare supply chain management: A multi-dimensional analysis with case study

Tagne Poupi Theodore Armand*, Kouayep Sonia Carole*, Subrata Bhattacharjee**, Md Ariful Islam Mozumder*, Austin Oguejiofor Amaechi***, Hee-Cheol Kim*

*Institute of Digital Anti-Aging Healthcare, Inje University, Gimhae 50834, Republic of Korea

**Department of Computer Engineering, u-AHRC, Inje University, Gimhae 50834, Republic of Korea

*** Department of Information and Communication Technology, The ICT University USA, Cameroon Campus.

poupiarmand2@gmail.com, carolesonia39@gmail.com, subrata_bhattacharjee@outlook.com, arifulislamro@gmail.com, austinhanz@gmail.com, heeki@inje.ac.kr

Abstract— **As time goes on, rapid development happens in the healthcare industry, and most of the significant challenges healthcare professionals and stakeholders face is supply chain management. With an excessive increase in demand for healthcare services and the need for efficient, cost-effective, and high-quality healthcare delivery, healthcare supply chain management has become a crucial factor in considering success in healthcare structures. Recently, Artificial Intelligence (AI), the Internet of Things (IoT), and blockchain have shown some potential to revolutionize healthcare supply chain management. In this research, we explore the benefits associated with integrating the abovementioned technologies in healthcare for more effective and efficient supply chain management in this industry. By leveraging these technologies, we explore the potential benefits of their integration into the healthcare supply chain using the eventual existing challenges. In this paper, we highlight the problems faced by conventional supply chain management and show how integrating AI, IoT, and BC can serve as a powerful tool to overcome these challenges. To achieve our goal, we carried out a multi-dimensional analysis with case studies that considered crucial aspects such as visibility, efficiency, data-driven decision-making, security, and trust in the supply chain. We proposed a healthcare supply chain management system that incorporates AI, IoT, and blockchain to raise awareness among healthcare providers about the benefits of an intelligent supply chain management system.**

Keywords— Artificial intelligence, Blockchain, Healthcare supply chain management, Internet of Things

I. INTRODUCTION

As in many other industries, the logistics of goods and services are a necessity of the first order. The Healthcare Supply Chain Management (HCSM) monitors and manages the flow of healthcare products and services from the manufacturing plant to the end user. It is a critical aspect of the healthcare industry that involves planning, procuring, storing, and distributing medical supplies, drugs, and equipment [1]. An effective healthcare supply chain system can deliver medications and treatment protocols to patients at any time that suits their needs. From the onset, the supply chain management system in healthcare involves getting the best suppliers with the best deals for healthcare-related products, managing stock at all levels of the delivery process, including the tracking of expiration dates, ensuring efficient movement of products and their effective delivery, ensuring the quality of the products in compliance with the industrial regulations while maintaining cost control and supplier relationship [2].

Effective supply chain management ensures patients receive timely, high-quality care while reducing costs and waste [3]. However, healthcare supply chain management faces several challenges that hinder its efficiency and effectiveness [2, 4]. These challenges include the lack of transparency and visibility, fragmented systems and processes, limited interoperability between stakeholders, inefficient inventory management, and inadequate data management and analytics. Given the challenges faced in healthcare supply chain management, there is a need for innovation that will help to improve efficiency and effectiveness in processes [5].

Addressing these challenges calls for innovation in the supply chain management system. A significant improvement can be noticed in HCSM over the years due to the integration of digital technologies and data analytic techniques driving toward patient-centred and value-based care. Emerging technologies such as Artificial Intelligence (AI), Internet of Things (IoT), and Blockchain have the potential to revolutionize healthcare supply chain management by addressing some of the challenges mentioned above [6- 9]. An example of using AI can be the improvement of demand forecasting. At the same time, IoT is the backbone for tracking deliveries, and blockchain provides a transparent platform for stakeholder operations. By leveraging these technologies, healthcare supply chain management can become more efficient, cost-effective, and patient-centric.

This study highlights the benefits of integrating emerging technologies (AI, IoT, and blockchain) into the HSCM for more effectiveness. Current challenges of HCSM are discussed; the integration of emerging technologies demonstrated a promising way to address these challenges. We adopted a multi-dimensional analysis of vital and critical aspects of the supply chain that are fundamental for its optimal operation. As depicted in Figure 1, these key aspects include visibility, efficiency, data-driven decision-making, security, and trust in the supply chain. By analyzing the challenges and solutions of the HSCM based on these aspects, we demonstrate the potential of AI, IoT, and blockchain to improve HSCM and enhance patient outcomes.

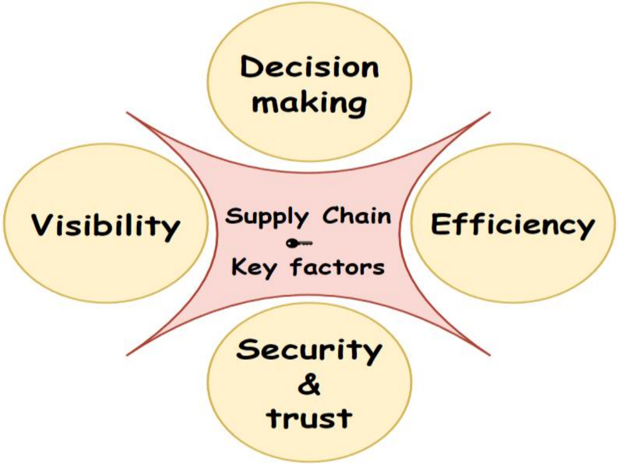


Figure 1. Key aspects considered in supply chain innovation

II. Challenges in healthcare supply chain management

Several studies have identified the HCSM system's challenges. These challenges are numerous, and a few will be discussed in this section. Figure 2 shows some challenges of conventional HCSM.

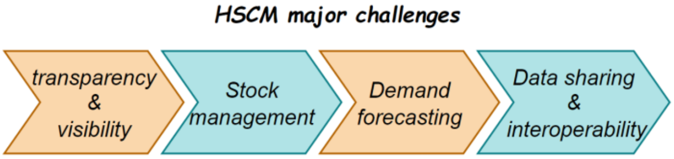


Figure 2. Challenges of the conventional healthcare supply chain management

- *Lack of transparency and visibility*: Tracking and monitoring products as they move on the supply chain system is crucial to give stakeholders insights into their delivery schedules and inventories. In healthcare systems, disruption can be observed in real-time operations, leading to delayed and errored deliveries. A more visible and transparent supply chain system is necessary to boost business effectiveness while optimizing transactions. The lack of visibility in the HSCM can result in inefficient operations, unmitigated supply chain disruptions, increased risk of theft or loss, inaccurate data, and poor relationships between customers and suppliers [4].
- *Stock management challenges*: Maintaining inventory levels to avoid stockouts and overstocking is significant in healthcare supply chain management. The lack of accurate data on the real-time inventory can lead to overstocking or stockouts with considerable financial implications. The supply and demand should be controlled in the HCSM to avoid unexpected issues such as expiration. Expiration is a waste that causes many businesses to collapse. Inappropriate stock management is caused by poor medicine selection, inaccurate forecasting, missed demand quantification, and poor warehouse management. It can lead to storing more products that can't be consumed before the expiration date, generating waste of storage and financial resources [4, 10].
- *Demand forecasting*: Anticipating the demand for medical supplies, equipment, pharmaceuticals, and services to ensure the availability of resources when needed is a complex and challenging task in HCSM. The variability in patient demand and clinical practices are key factors that make demand forecasting complex. Moreover, lack of real-time quality data on patient demographics and solicited products, regulatory changes, eventual disruptions, and interconnected processes in the supply chain equally contribute to poor demand forecasting that leads to low quality of healthcare services and deliveries with associated high costs [11-12].
- *Data sharing and interoperability*: The ability to exchange data seamlessly and efficiently among various stakeholders is essential for efficient HCSM. It can be noticed that the traditional HCSM is facing a lot of challenges in achieving data sharing and interoperability due to data silos across healthcare organizations, lack of standardization, interoperability gaps, privacy and security concerns, cost and resources constraints, regulatory, cultural, and organizational challenges. These challenges must be addressed to allow stakeholders to seamlessly exchange data over different healthcare platforms for effective and efficient HCSM [13].

The list of challenges can't be exhaustive; some other challenges include temperature control, considering its importance in drug supply, shipment visibility for real-time tracking, data-driven decision-making, trust, and security issues [14-16]. Given the challenges facing healthcare supply chain management, innovation is needed to improve its efficiency and effectiveness [5, 17]. Integrating Artificial Intelligence (AI), Internet of Things (IoT), blockchain, and other emerging technologies in the HSCM can potentially address some of the abovementioned challenges, thus revolutionizing it.

III. Integration of AI, IoT, and Blockchain in HCSM

Globally, Artificial Intelligence (AI) is the simulation of human intelligence processes by computers to perform problem-solving, decision-making, and prediction tasks. On the other hand, The Internet Of Things (IoT) enables the connection between objects (drugs, medical devices, animals, people) and devices over the Internet to facilitate data exchange. Blockchain is a distributed ledger technology that provides secure, transparent, and tamper-proof record-keeping without the need for approval of a central authority. In healthcare, AI predicts a disease based on patient data [18]; IoT sensors help collect the patient's vital signs [19], while blockchain assures the authenticity and security of healthcare data and transactions [20].

The convergence of AI, IoT, and blockchain in healthcare supply chain management is a transformative development. AI processes data to predict demand and optimize inventory, while IoT devices monitor the condition and location of medical supplies. Blockchain provides a secure and transparent ledger to track the flow of medical products. These technologies work together to create a more efficient and safe supply chain.

The benefits of the integration of these technologies in HCSM can be summarized as follows:

- *Improved Visibility*: AI analyzes historical data to predict demand accurately, while IoT sensors provide real-time tracking of product conditions and locations. Blockchain ensures transparency and data integrity throughout the supply chain.
- *Enhanced Efficiency*: Automation through AI and IoT reduces manual intervention, streamlining tasks like order processing and inventory management. Blockchain smart contracts automate transaction processes, reducing administrative overhead.
- *Data-Driven Decision-Making*: AI analyzes data for informed procurement, distribution, and demand forecasting decisions. IoT-generated data is valuable for monitoring the condition of products, enabling proactive measures. Blockchain provides a trustworthy audit trail for data-driven decisions.
- *Enhanced Security and Trust*: AI can identify anomalies or suspicious activities, reducing the risk of fraud or theft. IoT's real-time monitoring prevents unauthorized access to supplies. Blockchain's immutability ensures the authenticity of products and builds trust among stakeholders.

Figure 3 below describes the potential benefits of integrating AI, IoT and blockchain in conventional HSCM with some applications use cases.

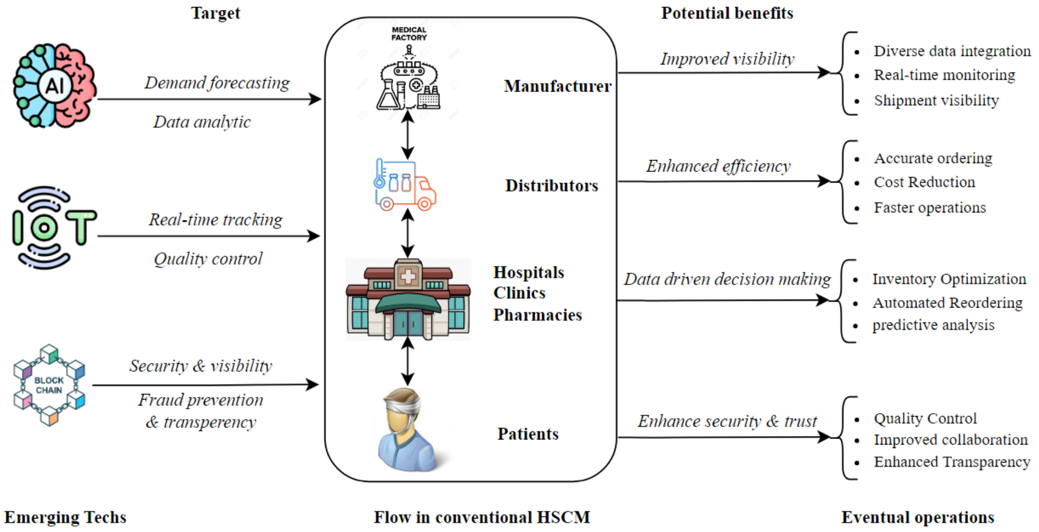


Figure 3. Integration of AI, IoT and blockchain in healthcare supply chain

IV. Case study of integrating AI, IoT, and Blockchain in HSCM

A. *Pharmaceutical Distributor Efficiency*

AI in predictive analytics and inventory management can significantly improve pharmaceutical distributor efficiency. By analyzing vast amounts of data, AI-powered predictive analytics can enable proactive identification and mitigation of potential risks in the drug supply chain. This can help distributors optimize inventory levels, reduce waste, and ensure timely delivery of drugs to healthcare providers and patients. With the help of AI, pharmaceutical distributors can achieve greater efficiency and accuracy in their supply chain management, ultimately benefitting patients and healthcare providers alike.

Implementing blockchain technology can provide secure and transparent transactions, ensuring the integrity and authenticity of the supply chain. Blockchain can enable end-to-end traceability of drugs, allowing for better tracking of products from manufacturers to patients. By leveraging blockchain for transparency and security, pharmaceutical distributors can enhance their supply chain management, reduce the risk of counterfeit products, and ensure the safe and timely delivery of drugs to patients. Blockchain technology can also facilitate data sharing across the supply chain, enabling better stakeholder collaboration and coordination.

Leveraging IoT can provide real-time visibility and monitoring of supply chain processes, enabling pharmaceutical distributors to identify and address any issues that may arise quickly. IoT devices can track the location and condition of drugs during transportation, ensuring that they are stored and transported under appropriate conditions. IoT can help distributors optimize their operations, reduce waste, and improve overall efficiency by providing real-time visibility into the supply chain. IoT in supply chain management can also enable better communication and collaboration among different stakeholders, improving the overall effectiveness of the supply chain.

B. Hospital Inventory Optimization

Stock management is a challenge that can generate a significant loss of money and resources when attention is not adequately paid. Healthcare institutions such as hospitals, pharmacies, and clinics in charge of patients' care must be cautious when ordering supplies. The traditional HSCM has proven limits in managing accurate orders that suit the population's needs. AI uses data to forecast demand, considering the population's needs accurately. From the manufacturing plant to the complete order delivery, the supplier and the customers can get real-time information during shipping. The IoT sensors can be attached to the parcel to capture real-time and vital information, such as the ambient temperature, and then transmit it to a server for visualization. Getting such data can help confirm the product's quality, a crucial parameter in a drug delivery scenario. While controlling the supplies from natural hazards using IoT sensors, blockchain technology can help secure procurement records. This integration will lead to rational stockouts, improving waste management and ensuring substantial cost savings.

C. Healthcare Network Data-Driven Decisions

Data-driven decision-making is a process that involves using data to inform and guide decision-making processes. In healthcare supply chain management, data-driven decision-making can help to optimize the flow of goods and services, reduce costs, and improve patient outcomes. By leveraging data analytics tools and techniques, healthcare organizations can gain insights into their supply chain operations and make informed inventory management, logistics, and distribution decisions.

AI, IoT, and blockchain can help to automate and streamline supply chain processes, reduce errors and inefficiencies, and improve data security and privacy. For example, AI can analyze large amounts of data and identify patterns and trends. At the same time, blockchain can provide a secure and transparent way to store and share data across different entities. IoT devices can also track and monitor inventory in real-time, providing greater visibility and control over the supply chain. By integrating these technologies into their supply chain operations, healthcare organizations can make more informed and data-driven decisions, improving patient outcomes and increasing efficiency and profitability.

Data-driven decision-making in HSCM leads to more optimized supply chain operations, causing a reduction in costs, improving profitability and patient outcomes/satisfaction, increasing efficiency and productivity, enhancing data security and privacy, and enhancing inventory and logistics management.

V. Challenges and prospects

While AI, IoT, and blockchain show the potential to revolutionize healthcare supply chain management, challenges such as data privacy, implementation cost, regulatory compliance, and platform interoperability must be acknowledged. The future of the HCSM is more promising with the integration of more emerging technologies such as Edge computing, 5G, digital twins, and others that can advance the capabilities of AI, IoT, and blockchain and ensure patient outcomes while increasing efficiency and profitability.

VI. Conclusion

This paper presents conventional healthcare supply chain management challenges and suggests solutions using AI, IoT, and blockchain technologies. A multi-dimensional analysis approach was adopted for key aspects like visibility, efficiency, data-driven decision-making, security, and trust. The paper discusses three case studies demonstrating the facilities that these technologies can bring into the HSCM. In pharmaceutical distribution, inventory optimization, or data-driven decision-making, AI technology is used for demand forecasting, data analytics, and other predictions. IoT sensors provide real-time information that can inform supplies' condition while blockchain ensures security in automated contracts and verification of pharmaceutical provenance. Though some challenges can be observed in fully implementing these three technologies in the HCSM, the product of this convergence will be helpful for the healthcare industry and is worth implementation to achieve user satisfaction.

Acknowledgment

This research was supported by a grant from Korea Health Technology R&D Project through the Korea Health Industry Development Institute validation (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant No: HI21C0977)

References

[1] Kwon, Ik-Whan G., Sung-Ho Kim, and David G. Martin. "Healthcare supply chain management; strategic areas for quality and financial improvement." Technological forecasting and social change 113 (2016): 422-428.

[2] Mathur, Bhavana, et al. "Healthcare supply chain management: literature review and some issues." *Journal of Advances in Management Research* 15.3 (2018): 265-287.

[3] Schneller, Eugene, et al. *Strategic management of the healthcare supply chain*. John Wiley & Sons, 2023.

[4] Privett, Natalie, and David Gonsalvez. "The top ten global health supply chain issues: perspectives from the field." *Operations Research for Health Care* 3.4 (2014): 226-230.

[5] Habidin, Nurul Fadly, et al. "A review of supply chain innovation and healthcare performance in healthcare industry." *International Journal of Pharmaceutical Sciences Review and Research* 35.1 (2015): 195-200.

[6] Pal, Kamalendu. "Blockchain-integrated internet-of-Things architecture in privacy preserving for large-scale healthcare supply chain data." *Blockchain Technology and Computational Excellence for Society 5.0*. IGI Global, 2022. 80-124.

[7] Gupta, Anil Kumar, et al. "Digital Supply Chain Management Using AI, ML and Blockchain." *Innovative Supply Chain Management via Digitalization and Artificial Intelligence*. Singapore: Springer Singapore, 2022. 1-19.

[8] Hu, Hui, et al. "Vaccine supply chain management: An intelligent system utilizing blockchain, IoT and machine learning." *Journal of business research* 156 (2023): 113480.

[9] Aich, Satyabrata, et al. "A review on benefits of IoT integrated blockchain based supply chain management implementations across different sectors with case study." *2019 21st international conference on advanced communication technology (ICACT)*. IEEE, 2019.

[10] Landry, Sylvain, and Martin Beaulieu. "The challenges of hospital supply chain management, from central stores to nursing units." *Handbook of healthcare operations management: Methods and applications*. New York, NY: Springer New York, 2013. 465-482.

[11] Singh, Sudhanshu, Rakesh Verma, and Saroj Koul. "Managing critical supply chain issues in Indian healthcare." *Procedia computer science* 122 (2017): 315-322.

[12] Jayaraman, Raja, Fatima AlHammadi, and Mecit Can Emre Simsekler. "Managing product recalls in healthcare supply chain." *2018 IEEE international conference on industrial engineering and engineering management (IEEM)*. IEEE, 2018.

[13] Khan, Athar Ajaz, and János Abonyi. "Information sharing in supply chains-Interoperability in an era of circular economy." *Cleaner Logistics and Supply Chain* (2022): 100074.

[14] Musamih, Ahmad, et al. "A blockchain-based approach for drug traceability in healthcare supply chain." *IEEE access* 9 (2021): 9728-9743.

[15] Bhatia, Ambika, and Prabhat Mittal. "Big data driven healthcare supply chain: understanding potentials and capabilities." *Proceedings of International Conference on Advancements in Computing & Management (ICACM)*. 2019.

[16] Mondal, Sanjana, and Kaushik Samaddar. "Reinforcing the significance of human factor in achieving quality performance in data-driven supply chain management." *The TQM Journal* 35.1 (2023): 183-209.

[17] Ageron, Blandine, Omar Bentahar, and Angappa Gunasekaran. "Digital supply chain: challenges and future directions." *Supply Chain Forum: An International Journal*. Vol. 21. No. 3. Taylor & Francis, 2020.

[18] Rong, Guoguang, et al. "Artificial intelligence in healthcare: review and prediction case studies." *Engineering* 6.3 (2020): 291-301.

[19] Armand, Tagne Poupi Theodore, et al. "Developing a Low-Cost IoT-Based Remote Cardiovascular Patient Monitoring System in Cameroon." *Healthcare*. Vol. 11. No. 2. MDPI, 2023.

[20] Sharma, Pratima, Malaya Dutta Borah, and Suyel Namasudra. "Improving security of medical big data by using Blockchain technology." *Computers & Electrical Engineering* 96 (2021): 107529.

Tagne Poupi Theodore Armand is a Ph.D. research scholar at the Institute of Digital Anti-aging Healthcare at Inje University. He received his M.Sc. in Information Systems and Networking at ICT University USA, Cameroon Campus. His research interests include artificial intelligence in healthcare, image processing focusing on medical image analysis, deep learning, machine learning, and metaverse.

Kouayep Sonia Carole is Ph.D. candidate in digital anti-aging and healthcare at Inje University, South Korea. She received her Master of Engineering degree from Pukyong University in Busan, South Korea. Previously, she earned her B.S degree in Computer Science from Dschang in Cameroon. Her research interests include image processing, computer vision, Artificial intelligence, and Business intelligence.

Subrata Bhattacharjee received his B.S. in information technology (IT) from the University of Derby, UK 2016. He is pursuing his M.S. leading Ph.D. from the Graduate School of Computer Engineering, Inje University, Korea. He is currently a Researcher and Teaching Assistant of the Medical Image Technology Laboratory (MITL) at Inje University. His research interests include digital image processing, multimodal medical imaging, tissue and cell image analysis, digital pathology, and AI applications in healthcare.

Md Ariful Islam Mozumder is pursuing his Ph. D. in the Institute of Digital Anti-Aging Healthcare at Inje University. He has previously worked on multiple real-life projects related to computer vision, data sciences, Smart IoT systems, and text mining. His research interest aligns with computer vision, medical image processing, metadata, AI, sensor data analysis, and blockchain.

Austin Oguejiofor Amaechi (Ph.D.) is a practitioner (over 20 years) and senior faculty member at Design Thinking and Innovation, Complex Systems, and Cybersecurity at ICT University, USA, Cameroon Campus. His current research interests include technology, artificial intelligence, management, and innovation in organizations.

Hee-Cheol Kim received his BSc at the Department of Mathematics, MSc at the Department of Computer Science at SoGang University in Korea, and Ph.D. in Numerical Analysis and Computing Science at Stockholm University in Sweden in 2001. He is a professor at the Department of Computer Engineering and Head of the Institute of Digital Anti-aging Healthcare Inje University in Korea. His research interests include machine learning, deep learning, Computer viion, and medical informaics.

Knowledge-Prompted Estimator: A Novel Approach to Explainable Machine Translation Assessment

1st Hao Yang
2012 Labs
Huawei Technologies CO., LTD
yanghao30@huawei.com

2nd Min Zhang
2012 Labs
Huawei Technologies CO., LTD
zhangmin186@huawei.com

3rd Shimin Tao
2012 Labs
Huawei Technologies CO., LTD
taoshimin@huawei.com

4rd Minghan Wang
2012 Labs
Huawei Technologies CO., LTD
wangminghan@huawei.com

5rd Daimeng Wei
2012 Labs
Huawei Technologies CO., LTD
weidaimeng@huawei.com

6rd Yanfei Jiang
2012 Labs
Huawei Technologies CO., LTD
jiangyanfei@huawei.com

***Abstract*—Cross-lingual Machine Translation (MT) quality estimation plays a crucial role in evaluating translation performance. GEMBA, the first MT quality assessment metric based on Large Language Models (LLMs), employs one-step prompting to achieve state-of-the-art (SOTA) in system-level MT quality estimation; however, it lacks segment-level analysis. In contrast, Chain-of-Thought (CoT) prompting outperforms one-step prompting by offering improved reasoning and explainability. In this paper, we introduce Knowledge-Prompted Estimator (KPE), a CoT prompting method that combines three one-step prompting techniques, including perplexity, token-level similarity, and sentence-level similarity. This method attains enhanced performance for segment-level estimation compared with previous deep learning models and one-step prompting approaches. Furthermore, supplementary experiments on word-level visualized alignment demonstrate that our KPE method significantly improves token alignment compared with earlier models and provides better interpretability for MT quality estimation.** [1]

***Index Terms*—Machine Translation Quality Estimation Chain-of-Thought Prompting Large Language Models**

I. Introduction

Large Language Models (LLMs), such as GPT-3 [1], ChatGPT [2], GPT-4 [3], and LLaMA [4], have been successfully validated in typical NLP scenarios, including question answering, search, summarization, and keyword extraction [5]. In the domain of multilingual NLP tasks, [6] [7] have demonstrated that utilizing a prompt-based approach, instead of fine-tuning models, can enable LLMs to perform machine translation tasks. This method has yielded impressive results for high-resource language pairs. However, for low-resource language pairs, the performance may be unsatisfactory due to insufficient training data or lower quality of available data.

Going further, GEMBA [8] explores the application of LLMs not only for translation tasks but also as translation quality estimators with one-step prompting. This approach targets two main scenarios: system-level quality estimation and segment-level quality estimation. By using one-step prompting, LLMs can assign scores to different systems or sentence pairs. Three scoring modes have been designed: scalar (0-100 points), 5-star (0-5 points), and 5-category (5-category classification). For evaluation, system-level assessment relies on accuracy as the metric, while segment-level assessment employs the Kendall ranking system as the evaluation metric. Results indicate that, under the prompting model and 5-category evaluation, state-of-the-art (SOTA) results are achieved at the system-level assessment. However, for segment-level quality

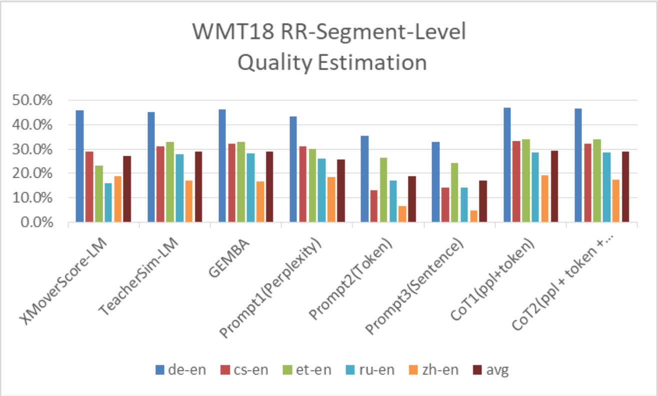


Fig. 1: The Kendall correlation of cross-lingual MT quality estimation for WMT18 segment-level metrics shows that the KPE CoT1 (CoT of perplexity and token) metric outperforms other metrics, including traditional deep learning metrics such as XMoverScore-LM and TeacherSim-LM, as well as LLM-based single-step metrics like GEMBA (perplexity, token, and sentence). Additionally, KPE CoT1 surpasses LLM-based CoT2, which is the CoT of perplexity, token, and sentence.

[1] Code will be released upon publication.

assessment, LLMs still fall short compared with dedicated machine translation quality assessment models [9].

This study introduces two innovative considerations: (1) transitioning from single-dimensional evaluation to multi-dimensional evaluation. Machine translation Quality Estimation (QE) [10] now involves more than just assigning a single score; it adopts a multi-dimensional evaluation approach (MQM) [11]. This approach assesses fluency and accuracy separately. Fluency can be evaluated by having LLMs measure the perplexity of a sentence, while accuracy can be assessed by having LLMs evaluate sentence-level similarity [12] or word-level similarity [13]. (2) To implement multi-dimensional evaluation, the CoT prompting method for LLMs can be employed [14] [15]. This approach guides LLMs to consider fluency first, followed by word-level and sentence-level accuracy, before finally combining 2-3 feature aspects to produce the best results. In the end, the study demonstrates that, in the WMT QE task, our KPE system achieves the best performance in 80% of segment-level tasks. Additionally, in terms of interpretability [16], token-level alignment exhibits better results.

In summary, this study can be characterized by the following key points:

- Introducing Knowledge-Prompted Evaluator, including 3 one-step prompting evaluations for perplexity, token-level similarity, and sentence-level similarity; and 2 CoT prompting evaluations for perplexity-token prompting and perplexity-token-sent prompting.
- Experimental validation demonstrates that KPE achieves positive gains at each step of segment-level evaluation. Furthermore, one-step prompting is competitive and CoT1 prompting achieves SOTA segment-level evaluation performance, even better than CoT2 prompting.
- In terms of interpretability analysis, we compare the outcomes of BertScore, TeacherSim, and KPE. Our findings reveal that KPE-based token alignment exhibits substantially better accuracy and more discriminative power than previous results, thus offering enhanced interpretability capabilities.

II. Related Work

A. Machine Translation Quality Estimation

Machine translation assessment can be classified into two categories based on the presence or absence of reference translations: quality evaluation as metrics, and Quality Estimation (QE) as metrics. Quality evaluation is a method that predicts the accuracy of machine translations based on a triplet of source text (src), machine translation output (mt), and reference translation (ref). In contrast, QE predicts the accuracy of machine translations solely based on the source text (src) and machine translation output (mt). Quality evaluation as metrics tend to have higher accuracy than QE. However, due to the need for human-provided reference translations for each source text, quality evaluation may be less efficient in practical applications. In contrast, QE as metrics, which does not require reference translations, has a wider range of applications despite its lower accuracy.

QE as metrics can be divided into two categories based on the evaluation methodology: system-level evaluation and segment-level evaluation.

System-level evaluation employs a calculation method similar to that of Learning-to-Rank (LTR), within two steps: For each system, it gets a system-level score with the source list (src list) and machine translation list (mt list). It then calculates the pairwise accuracy based on the machine-ranked and human labeled system scores. Segment-level evaluation, on the other hand, generates Relative Ranking (RR) segment-level data for all system pairs of source text (src) and machine translation (mt) using Direct Assessment (DA) or Multidimensional Quality Metrics (MQM). This data consists of triplets (src, mt1, mt2), indicating that the human evaluation score for mt1 is higher than that for mt2. Segment-level evaluation calculates correlations using the Kendall ranking system.

$$\begin{aligned} QE =&fun(src, mt)\\ &\quad \text{– segment-level } fun\text{, input (src, mt) return score}\\ =&fun(srclist, mtlist)\\ &\quad \text{– system-level } fun\text{, input (src list, mt list) return score} \end{aligned} \tag{1}$$

B. Translation Quality Estimation Metrics

Pairwise accuracy is the most commonly employed metric for system-level translation QE.

$$\begin{aligned} Accuracy =&\frac{|sign(metric(\triangle) - sign(human(\triangle))|}{|allsystempair|}\\ &\quad - metric(\triangle) \text{ metric score pair difference}\\ &\quad - human(\triangle) \text{ human labels pair difference}\\ &\quad - sign \text{ return 1 if first is better, else 0} \end{aligned} \tag{2}$$

where Kendall's Tau [17] is the most commonly employed metric for segment-level translation QE.

$$\begin{aligned} \tau =&\frac{Concordant - Discordant}{Concordant + Discordant}\\ &\quad \text{– “Concordant” - count of the set agrees with human labels}\\ &\quad \text{– “Discordant” - count of the set disagrees with human labels} \end{aligned} \tag{3}$$

5*Human		$s1 < s2$	$s1 = s2$	$s1 < s2$
	$s1 < s2$	Concordant	Discordant	Discordant
	$s1 = s2$	–	–	–
	$s1 > s2$	Discordant	Discordant	Concordant

TABLE I: Kendall metrics

C. LLM-based Quality Estimation Metrics

Prompt engineering is the key component of success when it comes to using modern AI conversation tools and language models like ChatGPT and GPT-4. It is the art of crafting a statement or question that returns accurate and valuable results.

Different from traditional deep learning QE models, LLM QE or GEMBA is a prompt-tuning method based on LLMs. It consists of three steps: (1) finding an appropriate prompt template to complete the task, which includes a predefined response format; (2) filling the template with the source sentence, translation sentence, and other parameters; and (3) parsing results from the response.

The LLMs QE formula is defined as follows:

$$\begin{aligned} \text{LLMs QE} =& \text{fun(prompt, src list, mt list)} \\ & - prompt\text{ , template string for quality estimation} \\ & - src\ list\text{, source sentences} \\ & - mt\ list\text{, translation sentences} \end{aligned} \tag{4}$$

The only difference between traditional QE and LLM QE lies in whether $prompts$ are used.

The generation of prompts can begin by asking counter-questions to LLMs. After obtaining multiple candidate templates, a suitable one can be manually selected and slightly modified. An example of a GEMBA prompt is as follows:

Classify the quality of machine translation into one of following classes: "No meaning preserved", "Some meaning preserved, but not understandable", "Some meaning preserved and understandable", "Most meaning preserved, minor issues", "Perfect translation".
source: "source_seg"
machine translation: "target_seg"
Class:

III. Proposed Approach

Prompt engineering can be categorized into two types: one-step prompting and Chain-of-Thought (CoT) prompting. One-step prompting is the most basic and straightforward prompt type, similar to zero-shot learning. Essentially, models like ChatGPT and GPT-3 complete the given prompt, which is the primary mechanism behind their functionality. One-step prompting works like a simple question-answer process, and even an initial statement can suffice, as the AI model will always attempt to complete it.

In contrast, CoT prompting not only provides AI with context for completion, but also offers a "chain of thought" process that demonstrates how the correct answer to a question should be reached. This type of prompting encourages reasoning and can even improve arithmetical results, which AI language models sometimes struggle with. Furthermore, due to its step-by-step reasoning approach, CoT prompting is much more explainable.

Our KPE tries to improve segment-level QE performance with CoT prompting. We first design three one-step prompting metrics, and then chain one-step prompting metrics as CoT metrics.

A. KPE One-Step Prompting Metric

Drawing inspiration from SOTA segment-level QE systems such as CometKiwi and TeacherSim, QE can be divided into three aspects, perplexity, token-level similarity and sentence-level similarity. We designed three one-step prompting metrics.

The one-step perplexity QE formula is defined as follows:

$$\begin{aligned} \text{Prompt1} =& \text{fun(perplexity\_prompt, mt)} \\ \text{Prompt2} =& \text{fun(token\_sim\_prompt, src, mt)} \\ \text{Prompt3} =& \text{fun(sent\_sim\_prompt, src, mt)} \end{aligned} \tag{5}$$

where $perplexity\_prompt$ is the prompt to estimate perplexity only based on mt, whereas $token\_sim\_prompt$ and $sent\_sim\_prompt$ are the prompts to estimate token level similarity and sentence level similarity.

B. KPE CoT Prompting Metric

KPE CoT prompting metrics are step-by-step prompting metrics based on three one-step prompting metrics.

$$\begin{aligned} \text{CoT1} =& \text{fun(comb\_prompt1, Prompt1, Prompt2)} \\ \text{CoT2} =& \text{fun(comb\_prompt2, Prompt1, Prompt2, Prompt3)} \end{aligned} \tag{6}$$

where $comb\_prompt1$ is quality estimation based on perplexity and token-level similarity, whereas $comb\_prompt2$ is quality estimation plus sentence-level similarity.

C. KPE Prompt Design

We follow the approach used in GEMBA for generating prompts. For the three one-step prompts and the two CoT prompts, we first have the system recommend five candidate prompt templates. We then manually select and edit these templates to create the optimal prompts. One-step prompts and CoT prompts are created like those in Figure 2.

IV. Experiments

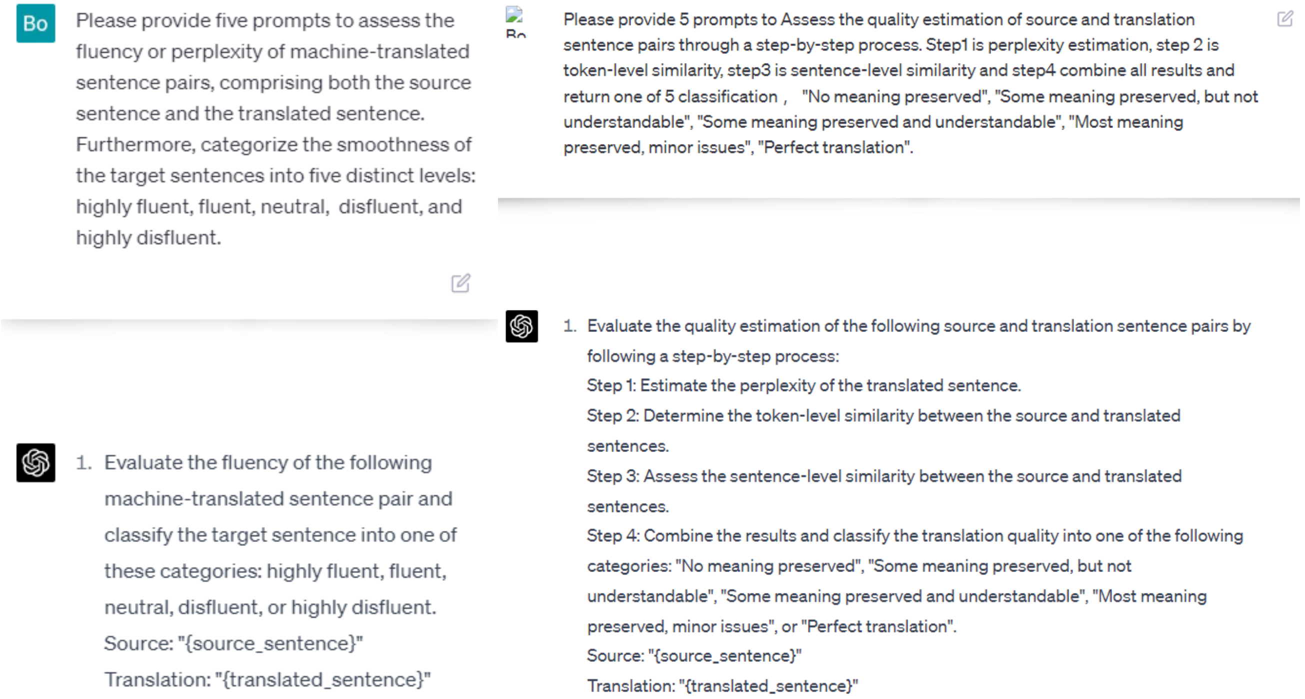
In this study, the proposed method is evaluated using the WMT18 metrics segment-level task data. For multilingual reference-free evaluation, the (source, candidate) pair is selected as a metric, similar to the QE approach.

A. Datasets

Our experiments employ the WMT18 news dataset as the evaluation dataset, containing 3,000 source sentences. The evaluation dataset also encompasses translated sentences from all teams that participated in WMT18. These target sentences are divided into 14 language pairs, such as DE-EN, FI-EN, and ZH-EN. Among them, DE-EN sentences are the largest in number (77,811), while CS-EN sentences are the smallest in number (5,110). Professional multilingual experts conduct pairwise comparisons on the target sentences before utilizing them as test data.

TABLE II: WMT18 segment-level to-En datasets.

LP	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en
RR_Systems	5	16	14	9	8	5	14
Dev Datasize	5110	77811	56712	15648	10404	5525	33357



Please provide five prompts to assess the fluency or perplexity of machine-translated sentence pairs, comprising both the source sentence and the translated sentence. Furthermore, categorize the smoothness of the target sentences into five distinct levels: highly fluent, fluent, neutral, disfluent, and highly disfluent.

1. Evaluate the fluency of the following machine-translated sentence pair and classify the target sentence into one of these categories: highly fluent, fluent, neutral, disfluent, or highly disfluent.
   Source: "{source_sentence}"
   Translation: "{translated_sentence}"

One-step Prompt for perplexity metric (Prompt1)

Please provide 5 prompts to Assess the quality estimation of source and translation sentence pairs through a step-by-step process. Step1 is perplexity estimation, step 2 is token-level similarity, step3 is sentence-level similarity and step4 combine all results and return one of 5 classification , "No meaning preserved", "Some meaning preserved, but not understandable", "Some meaning preserved and understandable", "Most meaning preserved, minor issues", "Perfect translation".

1. Evaluate the quality estimation of the following source and translation sentence pairs by following a step-by-step process:
   Step 1: Estimate the perplexity of the translated sentence.
   Step 2: Determine the token-level similarity between the source and translated sentences.
   Step 3: Assess the sentence-level similarity between the source and translated sentences.
   Step 4: Combine the results and classify the translation quality into one of the following categories: "No meaning preserved", "Some meaning preserved, but not understandable", "Some meaning preserved and understandable", "Most meaning preserved, minor issues", or "Perfect translation".
   Source: "{source_sentence}"
   Translation: "{translated_sentence}"

CoT Prompt for perplexity + token-level + sentence-level metric (CoT1)

Fig. 2: One-step prompt for perplexity QE and CoT prompt for perplexity + token-level similarity and sentence-level similarity.

B. Models

The comparison systems comprise several strong baseline systems, including:

- **M-BERT [18] and LASER [19]** as the multilingual pre-trained model.
- **XMoverScore and XMoverScore+LM [20] [21]** based on token-level similarity.
- **TeacherSim and TeacherSim-LM [12]** as the strong baseline based on sentence-level similarity.
- **GEMBA [8]** as the strong baseline based on one-step prompt and LLMs.

In addition, we tested our method, which includes (1) three single-step approaches for perplexity, token similarity, and sentence similarity, and (2) two CoT methods: perplexity + token similarity; perplexity + token similarity + sentence similarity.

C. Main Results

Using multiple datasets such as the WMT18 to-En datasets, we compare the similarity of the source and translated sentences and estimate their quality by utilizing the Kendall rank correlation coefficient. The results reveal that:

- (1) One-step prompts achieve comparable performance, with Prompt1 (25.7%), Prompt2 (18.8%), and Prompt3 (17.1%) performing better than traditional deep learning single models like M-BERT (11.0%), LASER (15.0%), XMoverScore (15.0%), and TeacherSim (19.0%);
- (2) CoT1 (29.1%) and CoT2 (28.9%) outperform the combined method XMoverScore-LM (27.0%) and LLM GEMBA (28.8%);
- (3) CoT1 (29.1%) achieves SOAT performance, surpassing CoT2 (28.9%) and Teacher-LM (29.0%), which indicates that increasing steps in prompting does not definitely improve the performance.

D. Scorer Analysis

Based on GEMBA's analysis, the scoring methods for QE can be divided into scalar scoring, 5-star scoring, and 5-category scoring. For segment-level evaluation, LLMs do not have high accuracy in scalar scoring, while 5-star and 5-category scoring methods perform better.

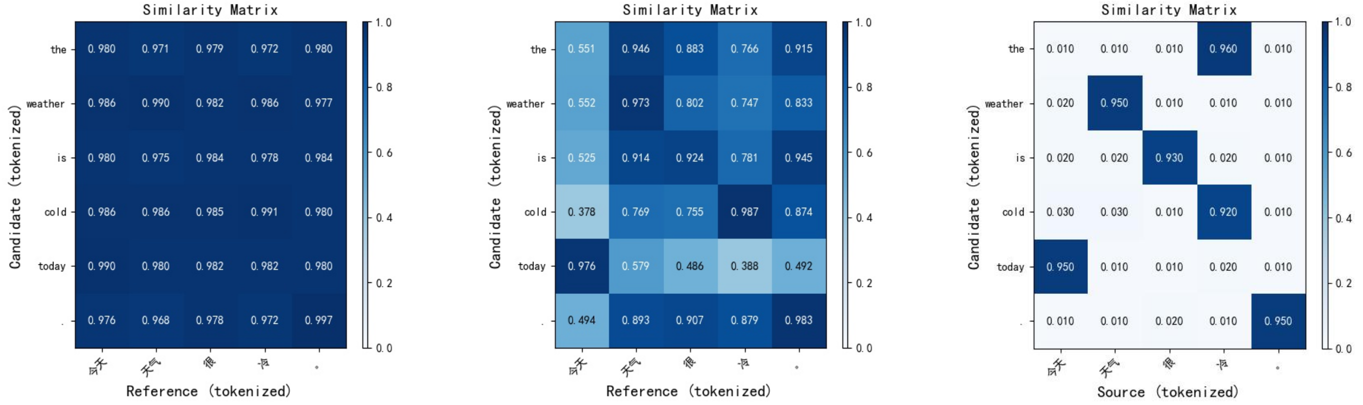
In order to analyze the scoring capabilities of large models, we further examined the differences in scoring quality between 3-category and 5-category methods. We found that as machine translation quality improves, the 3-category model tends to gather more than 30% in the neutral category, while the 5-category model has a much larger distinction, as shown in Figure 4.

E. Explainable Analysis

We also attempted to use large models for explainable analysis of QE in a similar manner to BERTScore or TeacherSim. We conducted visualization experiments for token-level alignment, calculating the similarity score for each token in the candidate sentence with each token in the reference sentence.

TABLE III: KPE result analysis for WMT18 segment-level to-En datasets.

model	category	de-en	cs-en	et-en	fi-en	ru-en	zh-en	tr-en	avg
M-BERT	PLMs	23.0%	1.0%	18.0%	12.0%	10.0%	8.0%	4.0%	11.0%
LASER	PLMs	32.0%	7.0%	25.0%	16.0%	10.0%	6.0%	9.0%	15.0%
XMoverScore	Token	28.0%	8.0%	21.0%	15.0%	15.0%	12.0%	9.0%	15.0%
XMoverScore-LM	Token + PLMs	46.0%	29.0%	23.0%	32.0%	16.0%	19.0%	16.0%	27.0%
TeacherSim	Sentence	17.0%	13.0%	17.0%	23.0%	26.0%	15.0%	21.0%	19.0%
TeacherSim-LM	Sentence+PLMs	45.0%	31.0%	33.0%	**24.0%**	28.0%	17.0%	**22.0%**	29.0%
GEMBA	One Step LLMs	46.3%	32.1%	32.9%	23.9%	28.3%	16.7%	20.7%	28.8%
Prompt1(Perplexity)	One Step LLMs	43.2%	31.1%	29.9%	18.8%	25.9%	18.5%	12.3%	25.7%
Prompt2(Token)	One Step LLMs	35.5%	13.0%	26.3%	16.4%	17.2%	6.7%	16.6%	18.8%
Prompt3(Sentence)	One Step LLMs	32.9%	14.0%	24.3%	14.7%	14.2%	4.9%	14.7%	17.1%
CoT1(ppl+token)	CoT LLMs	**47.1%**	**33.4%**	**33.8%**	22.6%	**28.7%**	**19.1%**	19.3%	**29.1%**
CoT2(ppl + token + sent)	CoT LLMs	46.7%	32.1%	33.8%	23.9%	28.4%	17.3%	20.4%	28.9%



(a) Token-level alignment visualization based on BertScore over M-BERT

(b) Token-level alignment visualization based on BertScore over TeacherSim

(c) Token-level alignment visualization based on KPE

Fig. 3: A comparison of explainable QE based on token-level alignment visualization for M-BERT, TeacherSim, and KPE demonstrates that KPE performs significantly better in terms of QE explanation.

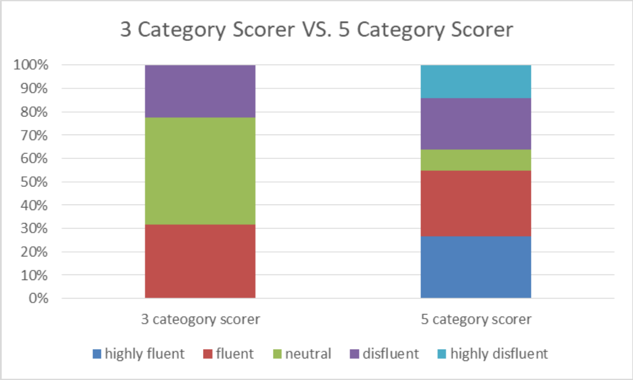


Fig. 4: A comparison of the KPE 3-category scorer and the 5-class scorer shows that the 5-category scorer performs significantly better, primarily because fewer samples are labeled as neutral.

This paper compares token-level alignment visualization for the multilingual pre-trained model XLM-R, the TeacherSim model, and KPE.

A typical case in Fig.3a shows that the similarity scores of all words in the multilingual pre-trained model M-BERT exceed 90%, making it difficult to differentiate them. The token-level similarity distribution based on TeacherSim, as shown in Fig.3b, is considerably uneven. Although it is much more accurate than the multilingual pre-trained model, it still exhibits leaky probability and aligns to the last token, such as *("the", ".")*, with an alignment probability of 91.5%.

Compared with these models, KPE's token alignment avoids both the issue of alignment probability exceeding 90% in XLM-R and the leaky probability phenomenon in TeacherSim. For instance, the alignment probability of *("the", ".")* is 1%, while that of *(".", ".")* is 95%, as shown in Fig3c. The accuracy and explainability of KPE's token alignment are significantly better than those of previous models.

V. Results

In this paper, we address the issue of low segment-level accuracy in quality estimation with single step prompts. We propose KPE, a QE method based on LLMs and CoT, which focuses on three core dimensions of quality estimation: perplexity, token similarity, and sentence similarity. Experimental results show that the CoT-based QE system outperforms both previous deep learning systems and single-step large model estimation methods in segment-level quality assessment. Moreover, KPE's token alignment visualization experiments demonstrate its clear superiority over multilingual pre-trained models and specialized sentence-level QE systems in terms of explainability [22] [23].

As for future research directions, on the one hand, we will attempt to fine-tune LLMs of various sizes, such as 6B, 13B, or 65B LLaMA models [24] [25] [4], to explore the upper limit of large models in performing QE. On the other hand, for language pairs with lower Kendall coefficients, such as zh-en, we will try to incorporate knowledge from knowledge-graphs to improve the metrics for the relevant language pairs.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] A. Kasirzadeh, "Chatgpt, large language technologies, and the bumpy road of benefiting humanity," 2023.

[3] OpenAI, "Gpt-4 technical report," 2023.

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[5] D. Vilar, M. Freitag, C. Cherry, J. Luo, V. Ratnakar, and G. Foster, "Prompting palm for translation: Assessing strategies and performance," *arXiv preprint arXiv:2211.09102*, 2022.

[6] W. Jiao, W. Wang, J. tse Huang, X. Wang, and Z. Tu, "Is chatgpt a good translator? yes with gpt-4 as the engine," 2023.

[7] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla, "How good are gpt models at machine translation? a comprehensive evaluation," *arXiv preprint arXiv:2302.09210*, 2023.

[8] T. Kocmi and C. Federmann, "Large language models are state-of-the-art evaluators of translation quality," *arXiv preprint arXiv:2302.14520*, 2023.

[9] H. Yang, S. Tao, M. Wang, M. Zhang, D. Wei, S. Zhao, M. Ma, and Y. Qin, "CCDC: A Chinese-Centric Cross Domain Contrastive Learning Framework," in *Knowledge Science, Engineering and Management*, G. Memmi, B. Yang, L. Kong, T. Zhang, and M. Qiu, Eds. Cham: Springer International Publishing, 2022, pp. 225–236.

[10] T. Kocmi, C. Federmann, R. Grundkiewicz, M. Junczys-Dowmunt, H. Matsushita, and A. Menezes, "To ship or not to ship: An extensive evaluation of automatic metrics for machine translation," *arXiv preprint arXiv:2107.10821*, 2021.

[11] R. Rei, J. G. de Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. F. Martins, "Comet-22: Unbabel-ist 2022 submission for the metrics shared task," in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 2022, pp. 578–585.

[12] H. Yang, M. Zhang, S. Tao, M. Ma, Y. Qin, and D. Wei, "Teachersim: Cross-lingual machine translation evaluation with monolingual embedding as teacher," in *2023 25th International Conference on Advanced Communication Technology (ICACT)*. IEEE, 2023, pp. 283–287.

[13] M. Zhang, H. Yang, Y. Zhao, X. Qiao, S. Tao, S. Peng, Y. Qin, and Y. Jiang, "Implicit cross-lingual word embedding alignment for reference-free machine translation evaluation," *IEEE Access*, 2023.

[14] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic chain of thought prompting in large language models," *arXiv preprint arXiv:2210.03493*, 2022.

[15] Y. Fu, H. Peng, A. Sabharwal, P. Clark, and T. Khot, "Complexity-based prompting for multi-step reasoning," *arXiv preprint arXiv:2210.00720*, 2022.

[16] S. Tao, S. Chang, M. Miaomiao, H. Yang, X. Geng, S. Huang, M. Zhang, J. Guo, M. Wang, and Y. Li, "Crossqe: Hw-tsc 2022 submission for the quality estimation shared task," in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 2022, pp. 646–652.

[17] C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan, "Findings of the 2011 workshop on statistical machine translation," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 22–64.

[18] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01502*, 2019.

[19] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 597–610, 2019.

[20] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, "Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, August 2019.

[21] W. Zhao, G. Glavaš, M. Peyrard, Y. Gao, R. West, and S. Eger, "On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 1656–1671. [Online]. Available: https://aclanthology.org/2020.acl-main.151

[22] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung *et al.*, "Multilingual universal sentence encoder for semantic retrieval," *arXiv preprint arXiv:1907.04307*, 2019.

[23] H. Yang, Y. Deng, M. Wang, Y. Qin, and S. Sun, "Humor Detection based on Paragraph Decomposition and BERT Fine-Tuning," in *AAAI Workshop 2020*, 2020.

[24] H. Yang, M. Zhang, S. Tao, M. Ma, and Y. Qin, "Chinese asr and ner improvement based on whisper fine-tuning," in *2023 25th International Conference on Advanced Communication Technology (ICACT)*. IEEE, 2023, pp. 213–217.

[25] M. Wang, "Whislu: End-to-end spoken language understanding with whisper," pp. 213–217, 2023.

Integration of a Chatbot to facilitate access to educational content in digital universities

Birahim BABOU*, Khalifa SYLLA**, M.Y. SOW***, Samuel OUYA*

* LITA Laboratory, Polytechnic High School, Cheikh Anta Diop University of Senegal

**Department of Applied Mathematics and Computer Science, Cheikh Hamidou Kane Digital University (UNCHK) of Senegal

***Infrastructure and Information Systems Department, UNCHK of Senegal

birahim.babou@ucad.edu.sn, khalifa.sylla@unchk.edu.sn, mouhamadouyaya.sow@unchk.edu.sn, samuel.ouya@gmail.com

Abstract— **Digital universities have been developed in several countries, particularly on the African continent, to meet the need for massification in the higher education sector. However, the lack of physical space is a major drawback, preventing learners from succeeding and increasing the drop-out rate compared with a conventional university.**

In these digital universities, learners use distance learning platforms to complete their training. For a good training, mastery of the fundamental modules is essential.

With the frequent use of messaging applications, the integration of Artificial Intelligence (AI) could promote and facilitate access to educational content and enhance their learning experience.

In this article, we propose a model for integrating a chatbot that will enable learners to access training modules to increase their knowledge and master core modules through formative skills assessments.

The model we propose is based on the use of Machine Learning (ML) with the Rasa open-source framework and the Moodle Learning Management System (LMS) platform.

Keywords— **Chatbot, Moodle, Artificial Intelligence, Machine Learning, Rasa**

I. Introduction

Today, digital universities have been developed in several African countries to meet the need for massification. To mitigate these problems, some virtual or digital universities [1][2][3] have proposed several pedagogical approaches often based on open digital spaces (ODS) to complement the virtual space and offer students a place to work and solve pedagogical, technical, administrative and social problems [3][4].

In these digital learning institutions, learners use distance learning platforms to take their courses.

To reduce the dropout rate and facilitate access to educational content, some universities are integrating a number of solutions into their pedagogical model, including social media and pedagogical supervision by tutors.

Faced with this problem, largely linked to the framing and ergonomics of certain Learning Management System (LMS) platforms, we propose in this paper the integration of a chatbot to facilitate access to educational content and improve the learning experience for students.

A chatbot is a new form of automated contextual interaction enabling communication between users and machines or systems. This system exploits a conversational approach based on natural language [6].

This article will be structured as follows: after the introduction, we'll review related work in the field, followed by some basic chatbot concepts. After that, we'll talk about the design and implementation of the solution. We'll then conclude with a look ahead.

II. Related Work

A number of researchers have worked in this field, making pertinent proposals. Among them are K. Gaglo, B. M. Degboe, G. M. Kossingou and S. Ouya [5] who have proposed a chatbot, enabling, in the context of covid19, learners to self-train on the parts of the course they have not mastered. This chatbot is connected to a Moodle platform and enables users to continue learning while at home. This proposed chatbot, is developed as a plugin that can be used in other platforms.

K. Souali, O. Rahmaoui, M. Ouzzif and I. El Haddioui [6], who propose and describe a new recommendation approach centered mainly on the use of a chatbot linked to the Moodle platform.

N. Nenkov, G. Dimitrov, Y. Dyachenko and K. Koeva [7], who proposed an intelligent agent in the form of a chatbot on the IBM Bluemix platform. This agent automates interaction between users and the Moodle training platform. This is a very interesting proposal, but it is specific to a technology belonging to IBM.

Authors T. Kita, C. Nagaoka, N. Hiraoka and T. Molnár [8], in their article entitled "Development of a Moodle UI Using LINE Chat for Casual Learning as a Part of a Learner Assistive LMS", set up a chatbot for a mobile application enabling interaction between users and a Moodle LMS platform. This tool is used on a specific LINE Chat application and meets a need in the Japanese community.

W. Kaiss, K. Mansouri and F. Poirier [9], to improve the quality of online learning, have proposed a methodology, chatbot architectural design, to help learners self-regulate their learning by accompanying them via a chatbot within the Moodle platform, which constitutes a metacognitive virtual assistant.

S. Kesarwani, Titiksha and S. Juneja [10], with their chatbot in place, have enabled their institution's administration to reduce the amount of work they have to do in providing the required information to students, thus reducing their workload in continuing to answer all student questions. They also confirm that chatbot systems can be used in a wide range of sectors, including education, healthcare and marketing.

B. Rawat et al [11] carried out a detailed survey of recent deep learning techniques for chatbots. This has brought developers' understanding closer to good chatbot design.

J. J. Sophia and T. P. Jacob [12], in their article, demonstrate and present the design and development of Student Chatbot software integrated with a user website that manages student queries via defined intents. The article covers the chatbot system with the Recurrent Neural Network (RNN) to handle the language part, the Convolutional Neural Network (CNN) to handle the image part, Dialogflow, accurately illustrating the intention and entity representation and keyword matching techniques.

S. Ondáš, M. Pleva and D. Hládek [13], have developed three chatbots to support teaching at their university's Department of Electronics and Multimedia Telecommunications. The first is the KEMTbot, which is a robot located on the department's web page. It provides information from the Web and about the department's staff. The second chatbot is a bot that accompanies students during exercises in the subject "Databases". The last is an Amazon Alexa chatbot skill, which answers questions about the department on Amazon Echo devices.

III. Basic concepts & tools used

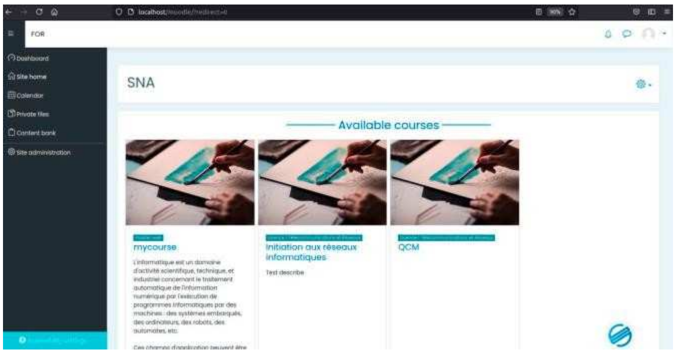
With the progress made in the field of Artificial Intelligence (AI) and Machine Learning (ML), conversational agents are playing a key role in various business sectors. Many organizations are opting for this type of solution, on the one hand to reduce the number of physical staff in the company, and on the other to enable a rapid, automatic response based on criteria defined at the time of implementation [14].

A conversational agent, chatbot or dialogue system is a tool for interacting with users in natural language. As a result, this system is able to understand and converse with a user as if it were humans talking. These types of systems can be text-based or voice-based [15].

These systems are applied and used in a wide range of sectors, including human resources, healthcare and education. [14][15].

A. Moodle

Moodle (Modular Object-Oriented Dynamic Learning Environment) is a free Learning Management System distributed under the GNU General Public License. It is developed in PHP. In addition to the possibility of creating courses with integrated tools and categorizing content by course, cohort level, sub-category, etc., the platform offers the possibility of being interconnected with external tools via secure APIs [16].



L'informatique est un domaine d'activité scientifique, technique, et industriel concernant le traitement automatique de l'information numérique par l'exécution de programmes informatiques par des machines : des systèmes embarqués, des ordinateurs, des robots, des automates, etc.

Ces champs d'application peuvent être

Test describe

Figure 1: Moodle platform home page

B. Interoperability between the chatbot and the Moodle platform using APIs

An API (Application Programming Interface) is a tool enabling different systems to communicate with each other. It defines the methods by which the two systems can communicate.

Moodle offers several APIs for interaction between the chatbot and its system. To retrieve data from the Moodle platform, authentication is required via a time-limited Tocken. To enable the chatbot to access the APIs, an authentication function must be implemented [16].

C. Natural Language Processing (NLP)

NLP (Natural Language Processing) is a branch of computer science focused on developing systems that enable computers to communicate with people using everyday language [17][18].

The intelligent conversation system is the foundation on which all Chatbots are built. It enables us to understand user requests and respond in a relevant way. This type of system is often built on top of an understanding and categorization algorithm. Let's now focus on the different elements of language processing: NLG (Natural Language Generation) and NLU (Natural Language Understanging).

Most chatbots operate on a basic model of these three properties, namely: Entities, Intentions, Response [18].

D. Key stages in the learning process

The first part consists of creating the NLU and discussion models, commonly known as the training phase. As Rasa is based on Machine Learning, it requires training data [19].

- For the NLU part (Rasa-NLU), the training data are sample sentences that the user might utter, in which intent and entities are specified. A configuration file is also required to set the algorithm parameters.
- For the discussion part (Rasa-CORE), a set of stories must be defined so that the agent learns to choose its next action. The configuration file accompanying the

stories contains lists of intentions, entities, slots and actions.

IV. Solution implementation

The implementation of a conversational agent involves several stages, including preparation and selection of the solution, development, and finally management and continuous improvement.

A. *Overall solution architecture*

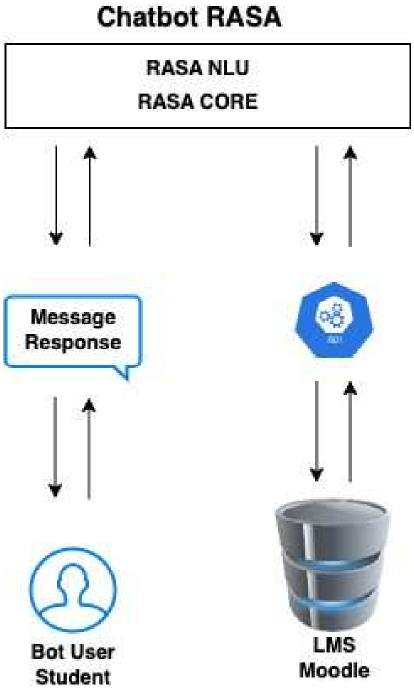
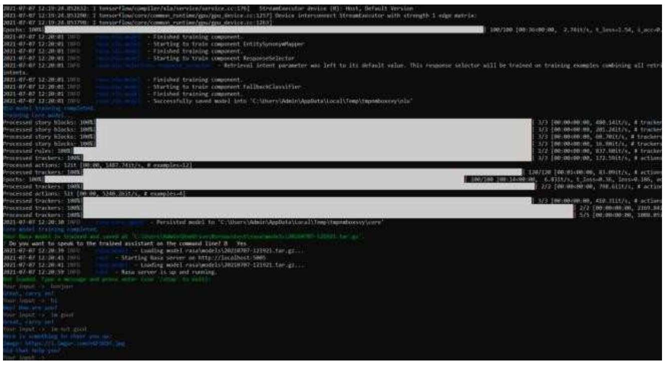


Figure 2: System architecture

B. *Solution development*

There are several stages in the development of the solution:
Step 1: Installing Rasa
Step 2: Project creation
Step 3: Defining intentions and examples
Step 4: Defining responses
Step 5: Creation of dialogue stories
Step 6: Model training and testing



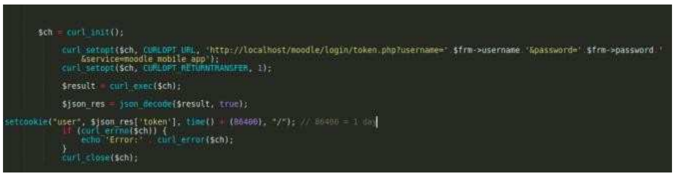
```
2021-07-07 12:19:24.852632: I tensorflow/compiler/xla/service/service.cc:176]   StreamExecutor device (0): Host, Default Version
2021-07-07 12:19:24.853290: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1257] Device interconnect StreamExecutor with strength 1 edge matrix:
2021-07-07 12:19:24.853798: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1263]
Epochs: 100%|                                                              | 100/100 [00:36<00:00,  2.76it/s, t_loss=1.56, i_acc=0.
2021-07-07 12:20:01 INFO     [illegible]   - Finished training component.
2021-07-07 12:20:01 INFO     [illegible]   - Starting to train component EntitySynonymMapper
2021-07-07 12:20:01 INFO     [illegible]   - Finished training component.
2021-07-07 12:20:01 INFO     [illegible]   - Starting to train component ResponseSelector
2021-07-07 12:20:01 INFO     [illegible]   - Retrieval intent parameter was left to its default value. This response selector will be trained on training examples combining all retri
intents.
2021-07-07 12:20:01 INFO     [illegible]   - Finished training component.
2021-07-07 12:20:01 INFO     [illegible]   - Starting to train component FallbackClassifier
2021-07-07 12:20:01 INFO     [illegible]   - Finished training component.
2021-07-07 12:20:01 INFO     [illegible]   - Successfully saved model into 'C:\Users\Admin\AppData\Local\Temp\tmpomboxsvy\nlu'
NLU model training completed.
Training Core model...
Processed story blocks: 100%|                                          | 3/3 [00:00<00:00, 480.14it/s, # tracker
Processed story blocks: 100%|                                          | 3/3 [00:00<00:00, 205.26it/s, # tracker
Processed story blocks: 100%|                                          | 3/3 [00:00<00:00, 60.70it/s, # trackers
Processed story blocks: 100%|                                          | 3/3 [00:00<00:00, 16.98it/s, # trackers
Processed rules: 100%|                                                 | 2/2 [00:00<00:00, 837.58it/s, # tracker
Processed trackers: 100%|                                              | 3/3 [00:00<00:00, 172.59it/s, # actions
Processed actions: 12it [00:00, 1487.74it/s, # examples=12]
Processed trackers: 100%|                                              | 120/120 [00:01<00:00, 83.09it/s, # action
Epochs: 100%|                                                          | 100/100 [00:14<00:00,  6.83it/s, t_loss=0.36, loss=0.186, ac
Processed trackers: 100%|                                              | 2/2 [00:00<00:00, 798.61it/s, # action
Processed actions: 5it [00:00, 5246.26it/s, # examples=4]
Processed trackers: 100%|                                              | 3/3 [00:00<00:00, 410.31it/s, # actions
Processed trackers: 100%|                                              | 2/2 [00:00<00:00, 2169.3
Processed trackers: 100%|                                              | 5/5 [00:00<00:00, 1000.0
2021-07-07 12:20:10 INFO     [illegible]   - Persisted model to 'C:\Users\Admin\AppData\Local\Temp\tmpomboxsvy\core'
Core model training completed.
Your Rasa model is trained and saved at '[illegible]\models\20210707-121921.tar.gz'.
? Do you want to speak to the trained assistant on the command line? 🤖  Yes
2021-07-07 12:20:39 INFO     [illegible]   - Loading model rasa\models\20210707-121921.tar.gz...
2021-07-07 12:20:41 INFO     [illegible]   - Starting Rasa server on http://localhost:5005
2021-07-07 12:20:41 INFO     [illegible]   - Loading model rasa\models\20210707-121921.tar.gz...
2021-07-07 12:20:59 INFO     [illegible]   - Rasa server is up and running.
Bot loaded. Type a message and press enter (use '/stop' to exit):
Your input ->  bonjour
Great, carry on!
Your input ->  hi
Hey! How are you?
Your input ->  im good
Great, carry on!
Your input ->  im not good
Here is something to cheer you up:
Image: https://i.imgur.com/nGF1K8f.jpg
Did that help you?
Your input ->
```

Figure 3: Model training

Once the prerequisites have been set up, the next step is to train the model and test it in console mode.

Step 7: Creating the graphical interface

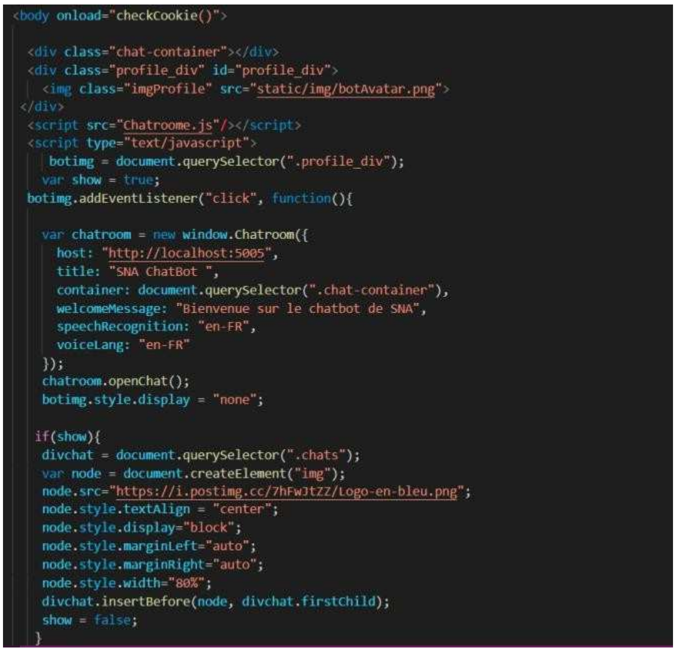
After configuration and testing, it's important to set up a graphical interface enabling users to interact with the system. This interface defines the access parameters for the Rasa chatbot.



```
        $ch = curl_init();

            curl_setopt($ch, CURLOPT_URL, 'http://localhost/moodle/login/token.php?username=' .$frm->username. '&password=' .$frm->password.'
                &service=moodle_mobile_app');
            curl_setopt($ch, CURLOPT_RETURNTRANSFER, 1);

            $result = curl_exec($ch);

            $json_res = json_decode($result, true);

setcookie("user", $json_res['token'], time() + (86400), "/"); // 86400 = 1 day
            if (curl_errno($ch)) {
                echo 'Error:' . curl_error($ch);
            }
            curl_close($ch);
```

Figure 4: Moodle authentication and token recovery function

After finalizing the creation of the chatbot in console mode, we created the graphical interface enabling us to interact more easily with the chatbot.



```
<body onload="checkCookie()">

  <div class="chat-container"></div>
  <div class="profile_div" id="profile_div">
    <img class="imgProfile" src="static/img/botAvatar.png">
 </div>
  <script src="Chatroome.js"/></script>
  <script type="text/javascript">
   botimg = document.querySelector(".profile_div");
   var show = true;
  botimg.addEventListener("click", function(){

    var chatroom = new window.Chatroom({
      host: "http://localhost:5005",
      title: "SNA ChatBot ",
      container: document.querySelector(".chat-container"),
      welcomeMessage: "Bienvenue sur le chatbot de SNA",
      speechRecognition: "en-FR",
      voiceLang: "en-FR"
    });
    chatroom.openChat();
    botimg.style.display = "none";

   if(show){
    divchat = document.querySelector(".chats");
    var node = document.createElement("img");
    node.src="https://i.postimg.cc/7hFwJtZZ/Logo-en-bleu.png";
    node.style.textAlign = "center";
    node.style.display="block";
    node.style.marginLeft="auto";
    node.style.marginRight="auto";
    node.style.width="80%";
    divchat.insertBefore(node, divchat.firstChild);
    show = false;
   }
```

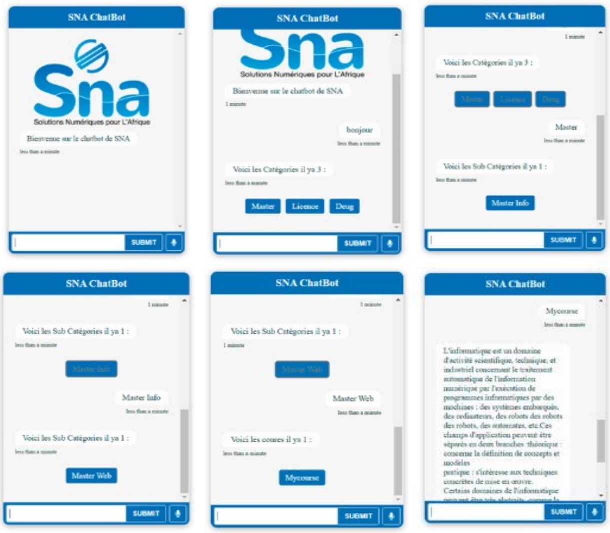
Figure 5: Code of Chatbot graphical interface

C. *Steering* and Continuous Improvement

Depending on the indicators set to measure the system's performance, it is important to measure the rate of understanding of the entities' performance, and to make continuous improvements by adding responses as the system is used.

V. Results and analysis

After implementing the various tools and tests required, the chatbot was set up and the following captures were taken.



Bienvenue sur le chatbot de SNA
less than a minute

Bienvenue sur le chatbot de SNA
1 minute
bonjour
less than a minute
Voici les Catégories il ya 3 :
less than a minute
Master Licence Deug

1 minute
Voici les Catégories il ya 3 :
less than a minute
Master Licence Deug
Master
less than a minute
Voici les Sub Catégories il ya 1 :
less than a minute
Master Info

1 minute
Voici les Sub Catégories il ya 1 :
less than a minute
Master Info
Master Info
less than a minute
Voici les Sub Catégories il ya 1 :
less than a minute
Master Web

1 minute
Voici les Sub Catégories il ya 1 :
1 minute
Master Web
Master Web
less than a minute
Voici les cours il ya 1 :
less than a minute
Mycourse

Mycourse
less than a minute
L'informatique est un domaine d'activité scientifique, technique, et industriel concernant le traitement automatique de l'information numérique par l'exécution de programmes informatiques par des machines : des systèmes embarqués, des ordinateurs, des robots des robots des robots, des automates, etc.Ces champs d'application peuvent être séparés en deux branches théorique : concerne la définition de concepts et modeles
pratique : s'intéresse aux techniques concrètes de mise en œuvre.
Certains domaines de l'informatique

Figure 6: Chatbot graphical interface

In these images, after clicking on the chatbot icon, the interface opens, displaying the different course categories available on the platform.

After selecting a given category, the corresponding sub-categories are displayed in turn.

After that, if there are no sub-categories, the course content is displayed on the chatbot, with the option of scrolling through the content.

VI. Conclusion and Perspectives

A chatbot is a new form of automated contextual interaction enabling communication between users and machines or systems. This system exploits a conversational approach based on natural language.

The chatbot implemented in this article with Artificial Intelligence (AI), Machine Learning (ML) and Rasa tools, then interconnected with the Moodle training platform, enables learners to self-train by considering the basic modules for training in the IT field.

With the advancement of AI, our next objective will be to interconnect this chatbot with ChatGPT to be able to answer questions from learners who are not on the Moodle training platform. This will make the tool complete and accessible to users in all training streams.

References

[1] UVBF, "*Université Virtuelle du Burkina Faso*", 2023. [Online]. Available: https://uv.bf/

[2] UVCI, "*Université Virtuelle de la Cote d'Ivoire* ", 2023. [Online]. Available: https://www.uvci.edu.ci/

[3] UNCHK, "*Université Numérique Cheikh Hamidou Kane,*" 2023. [Online]. Availaible: https://www.unchk.sn/

[4] Sylla, K., Nkwetchoua, G. M. M., & Bouchet, F. (2022). "How Does the Use of Open Digital Spaces Impact Students Success and Dropout in a Virtual University?". *International Association for Development of the Information Society*.

[5] K. Gaglo, B. M. Degboe, G. M. Kossingou and S. Ouya, "Proposal of conversational chatbots for educational remediation in the context of covid-19," *2022 24th International Conference on Advanced Communication Technology (ICACT)*, PyeongChang Kwangwoon_Do, Korea, Republic of, 2022, pp. 354-358, doi: 10.23919/ICACT53585.2022.9728860.

[6] K. Souali, O. Rahmaoui, M. Ouzzif and I. El Haddioui, "Recommending Moodle Resources Using Chatbots," *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Sorrento, Italy, 2019, pp. 677-680, doi: 10.1109/SITIS.2019.00110.

[7] N. Nenkov, G. Dimitrov, Y. Dyachenko and K. Koeva, "Artificial intelligence technologies for personnel learning management systems," *2016 IEEE 8th International Conference on Intelligent Systems (IS)*, Sofia, Bulgaria, 2016, pp. 189-195, doi: 10.1109/IS.2016.7737420.

[8] T. Kita, C. Nagaoka, N. Hiraoka and T. Molnár, "Development of a Moodle UI Using LINE Chat for Casual Learning as a Part of a Learner Assistive LMS," *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Takamatsu, Japan, 2020, pp. 927-929, doi: 10.1109/TALE48869.2020.9368321.

[9] W. Kaiss, K. Mansouri and F. Poirier, "Chatbot Design to Help Learners Self-Regulte Their Learning in Online Learning Environments," *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, Orem, UT, USA, 2023, pp. 236-238, doi: 10.1109/ICALT58122.2023.00075.

[10] S. Kesarwani, Titiksha and S. Juneja, "Student Chatbot System: A Review on Educational Chatbot," *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2023, pp. 1578-1583, doi: 10.1109/ICOEI56765.2023.10125876.

[11] B. Rawat, A. S. Bist, U. Rahardja, Q. Aini and Y. P. Ayu Sanjaya, "Recent Deep Learning Based NLP Techniques for Chatbot Development: An Exhaustive Survey," *2022 10th International Conference on Cyber and IT Service Management (CITSM)*, Yogyakarta, Indonesia, 2022, pp. 1-4, doi: 10.1109/CITSM56380.2022.9935858.

[12] J. J. Sophia and T. P. Jacob, "EDUBOT-A Chatbot For Education in Covid-19 Pandemic and VQAbot Comparison," *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2021, pp. 1707-1714, doi: 10.1109/ICESC51422.2021.9532611.

[13] S. Ondáš, M. Pleva and D. Hládek, "How chatbots can be involved in the education process," *2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, Starý Smokovec, Slovakia, 2019, pp. 575-580, doi: 10.1109/ICETA48886.2019.9040095.

[14] M. G. C. P, A. Srivastava, S. Chakraborty, A. Ghosh and H. Raj, "Development of Information Technology Telecom Chatbot: An Artificial Intelligence and Machine Learning Approach," *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*, London, United Kingdom, 2021, pp. 216-221, doi: 10.1109/ICIEM51511.2021.9445354.

[15] R. Singh, M. Paste, N. Shinde, H. Patel and N. Mishra, "Chatbot using TensorFlow for small Businesses," *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore, India, 2018, pp. 1614-1619, doi: 10.1109/ICICCT.2018.8472998.

[16] Moodle. (2023, Oct). *Moodle: Web Services API Functions – MoodleDocs* [Online]. Available: https://docs.moodle.org/dev/Web_service_API_functions

[17] H. K. K., A. K. Palakurthi, V. Putnala and A. Kumar K., "Smart College Chatbot using ML and Python," *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Pondicherry, India, 2020, pp. 1-5, doi: 10.1109/ICSCAN49426.2020.9262426.

[18] M. Ganesan, D. C., H. B., K. A.S. and L. B., "A Survey on Chatbots Using Artificial Intelligence," *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Pondicherry, India, 2020, pp. 1-5, doi: 10.1109/ICSCAN49426.2020.9262366.

[19] R. Agrawal, S. Kumar, S. Kumar, N. Goyal and S. Sinha, "WASABI Contextual BOT," *2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, Goa, India, 2022, pp. 481-485, doi: 10.1109/ICCCMLA56841.2022.9989007.

Session 4C: Computer Vision & Appliance Software 1

Chair: Prof. Nattagit Jiteurtragool , King Mongkuts University of Technology, Thailand

1 Paper ID: 20240312, 315~320

Explainable Rip Current Detection and Visualization with XAI EigenCAM

Mr. Juno Choi, Mr. Muralidharan Rajendran, Mr. Yong Cheol Suh,

Pukyong National University. Korea(South)

2 Paper ID: 20240076, 321~326

A Test Method for the Convergence of the Metaverse and Blockchain

Prof. Tae-gyu Lee,

Pyeongtaek University. Korea(South)

3 Paper ID: 20240429, 327~331

Time-frequency Analysis for Validating Prognostics Algorithms of Rolling Element Bearings

Dr. Guanhua Zhu, Dr. Xiaoling Xu, Mr. Qing Zhong , Dr. Bing-Yuh Lu, Mr. Yushen Lu, Mr. Guangming Xu, Mr. Yumeng Zhou, Mr. Ziyi Jiang, Mr. Kai Sun, Mr. Minhao Wang,

Guangdong University of Petrochemical Technology. China

4 Paper ID: 20240435, 332~336

Evaluation System for Dancing Enlightenment Posture Training Using the Skeleton Tracking of Microsoft Common Objects in Context

Mr. Ruilong Huang, Ms. Huifang Deng, Prof. Ruei-Yuan Wang, Prof. Bing-Yuh Lu, Prof. Hongwei Ren, Mr. Yiheng Chen, Mr. Jianwen Ye, Mr. Jinhui Chen, Mr. Yingbo Jia, Ms. Leyang Lang,

Guangdong University of Petrochemical Technology. China

5 Paper ID: 20240061, 337~342

Computer Vision-Based Structural Deformation Monitoring System on Android Smartphones: Design and Implementation

Ms. Xiang DONG, Mr. Maokai LAI, Ms. Hui LIANG, Mr. Peng WU, Ms. Chaoxia WANG, Mr. Ting PENG,

Chang'an University. China

Explainable Rip Current Detection and Visualization with XAI EigenCAM

Juno Choi [1], Muralidharan Rajendran [2], Yong Cheol Suh [3, *]

[1] UR Interdisciplinary Program of Hydrography, Pukyong National University, Busan 48513, Korea.

[2] Research and Development Team, GreenBlue Inc., Busan 48934, Korea.

[3] Department of Civil Engineering, Pukyong National University, Busan 48513, Korea.

jochoi@green-blue.co.kr, murali@green-blue.co.kr, suh@pknu.ac.kr

Abstract— **Rip currents have long posed a serious threat to beachgoers and swimmers. Despite numerous preventive measures throughout the period, the fatality rate [1] underscores the need for a robust rip current detection system. Recently deep learning models have shown promising results in rip current detection, outperforming traditional methods. However, these models still exhibit some accuracy limitations due to insufficient data distribution. To address this challenge, we incorporate a novel largest dataset [2] comprising over 110,215 Korean coastline images. Through the comparative study of the state-of-the-art models, we aim to analyze the detection accuracy of each model and gain a deeper understanding of their intensity over rip current detection. In comparison to the other rip current datasets, the evaluation results on our proposed dataset demonstrate a remarkable elevation in accuracy, reaching 79.4 mAP. Further, we employ the EigenCAM (Eigen Class Activation Maps) to interpret the intense regions of the rip currents and to gain a deeper comprehension of rip current explainability. This comprehensive analysis marks a significant step toward improving rip current safety and understanding.**

Keywords— **Rip current detection, computer vision, class activation maps, largest rip current dataset, state-of-the-art models comparison.**

I. Introduction

Rip currents are swift and narrow channels of water flowing from the shoreline out to the open sea. It presents a significant and potentially life-threatening hazard along coastlines worldwide. Rip currents have been responsible for numerous fatalities and near-drowning incidents. These currents exhibit variable strength, influenced by factors such as wave size, tide, and beach configuration. These factors render them unpredictable and challenging to identify through human visual observation. The water funneled through these channels can attain remarkable speed, reaching up to 2.43 ms^{-1}, which is faster than an average Olympic swimmer [3]. Predominant causes of rip currents are alterations in underwater topography such as sandbars and submerged depressions, and the configuration of breaking waves. Additionally, man-made structures such as piers and jetties can also contribute to the formation of rip currents. Traditional approaches to rip current detection primarily rely on human lifeguards for monitoring, yet accidents occur due to human errors and the unpredictability of rip currents. Consequently, an emergent need for automated rip current detection solutions has arisen. A subset of automated systems employs machine learning and numerical algorithms to scrutinize the complex data patterns associated with rip currents. These automated systems offer a notable advantage of continuous surveillance and expedited alerts, benefiting both beachgoers and lifeguards. Although these automated methods for rip current detection represent a marked improvement over traditional approaches, it is evident that further refinements and enhancements are necessary. In recent years, the coastal engineering community has leveraged deep neural networks to address numerous intricate challenges. Though deep learning models have proven successful in the realm of object detection techniques, there are several distinctive challenges exist in the case of rip current detection. Unlike traditional object detection scenarios like automobiles or buildings which have clearly defined object boundaries, rip currents are amorphous, lacking well-defined shapes or boundaries. These intrinsic complexities make rip currents detection exceptionally challenging.

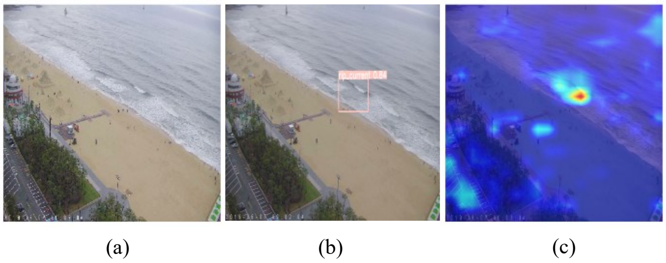


Figure 1. Illustration of predictions of visible rip currents on the coastline. Eigen-CAM method interprets the predicted rip currents on Haeundae Beach. (a) The original image, (b) Bounding box prediction of YOLOv8 model, (c) Visual explanations of Eigen-CAM over the same model and image.

One significant issue present in the literature of rip current detection is the absence of an appropriate dataset. This paper investigates the rip current detection on a specific coastline of the Republic of Korea and leverages an extensive dataset comprising over 110,215 images of Haeundae Beach, Busan. Every year a certain number of people get pulled away by these rip currents in Haeundae Beach. In 2017, a rip tide at this location carried over 70 persons 100 meters out to sea [4]. This distinctive dataset not only facilitates the training of detection models but also empowers the elucidation of rip current characteristics and explainability as the images are of a particular coastline. To make a suitable rip current preventive system, training the model and identifying the currents alone are insufficient. It is crucial to not only build these models but also comprehend their predictions. However, with the rapid evolution of deep learning solutions, the models are becoming increasingly complex in terms of layers and parameters used. It is impossible to explicitly transform each computation into a logical explanation since the deep learning models are black-box solutions. Therefore, it is essential to utilize XAI tools to extract the information from different layers and transform it into comprehensible knowledge. In this paper, we employ the Eigen-CAM to explain the intensity of the rip current occurred regions. Furthermore, this study endeavors to conduct a comparative analysis of state-of-the-art object detection models concerning rip channels. This development of automated detection methodologies represents a significant stride in advancing beach safety measures. Additionally, it offers a more holistic and efficient approach to mitigating the persistent risks associated with rip currents. The major contributions of this paper can be summarized as:

- We leverage 110,215 region-specific coastline images both with and without rip current for the first time.
- We perform a comparative analysis with SoTA models of single-stage and two-stage detection models.
- We implement the XAI model explainability tool named Eigen-CAM to interpret the intensity of rip currents.
- We perform a careful data augmentation procedure to avoid over-fitting on specific region images.
- We demonstrate a high accuracy compared to other benchmark rip current detection datasets.

II. Related Works

Following the development of Machine Learning (ML) and Artificial Intelligence (AI), rip currents have become a well-investigated phenomenon. To meet the critical need for increased coastal safety, numerous studies have aided in building rip current detecting techniques. Traditional approaches to rip current identification have encompassed various methodologies such as marine radar systems for remote sensing observations [5], the deployment of dye tracking techniques [6] (as illustrated in Figure 2), and the integration of wave buoys [7]. Subsequently, researchers explored in-situ instrumental systems like Lifeguarding Operational Camera Kiosk System (LOCKS) [8], and current profiling meters [9]. These advancements have significantly enhanced precision and reliability. However, the prevailing challenge of these systems is the cost-effectiveness of implementing them on a widespread scale across various beaches and coastal regions.

The age of ML and AI began after the development of traditional numerical techniques. Maryan et al. [10] employed the Viola-Jones algorithm [11] and Convolution Neural Networks (CNN) for rip current detection using time-exposed images. Likewise, Rashid et al. [12] also used the same dataset as Maryan et al. but with a different rip current identification method. They proposed a Fully Convolutional Auto-Encoder architecture as RipNet. Later they proposed another network named RipDet [13] which is a more lightweight framework than RipNet as it utilizes the Tiny-YOLO architecture. Subsequently, the same authors introduced a new residual-based architecture named RipDet+ [14] for enhanced feature detection on the same dataset as the prior version. However, the problem with these data and methodologies is that they are neither easily accessible nor easily understood by the average layman as a beachgoer. Hence, we use class activation maps to improve the explainability of the results of rip current detection.

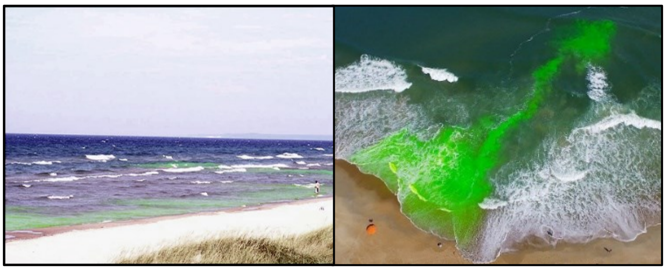


Figure 2. Traditional dye tracking method for rip current identification [26].

de Silva et al. proposed a novel dataset [15], designed for detecting rip currents in individual frames. They opted for the Faster Region-based Convolutional Neural Networks (Faster RCNN) [16], which surpassed prior methods. Furthermore, they integrated a specialized temporal aggregation stage to enable detection from both still images and videos. Following this, de Silva introduced a novel feature detection model RipViz [17] aimed at extracting the essential features of rip currents from stationary videos. To gain insights into water flow behavior, they employed an LSTM autoencoder framework combined with pathline sequences. In the same way, Zhu et al. [18] significantly enhanced the performance by their YOLO-Rip model which is a lightweight network based on YOLOv5s [19] architecture. They incorporated the de Silva rip current dataset along with their collection of rip current images from the South China coastlines. However, some of the results of these studies exhibit false negative reports, due to the insufficient data distribution for training the model. We surpass this data distribution challenge by leveraging a novel rip current dataset with 110,215 coastline images. Further, we employ the Explainable AI tools to improve the explainability of the rip current detection. These combined enhancements significantly elevate the detection accuracy and ultimately safeguard beachgoers along coastal regions.

III. Dataset

The availability of meticulously annotated datasets remains a fundamental prerequisite for enhancing the efficacy of rip current detection. In the extant literature, several substantial datasets have been introduced for rip current detection. de Silva et al. [15] presented a novel dataset of 2440 images and 23 videos both with and without rip currents. A limitation of this dataset is that most of the images are of top-elevated aerial perspective, lacking diversity in terms of beach scene perspectives. In recognition of this limitation, Zhu et al. [18] supplemented an additional 1352 images to the de Silva dataset featuring varied perspectives. Similarly, Dumitriu et al. [20] leveraged the Zhu dataset and meticulously annotated 2466 rip current images for instance segmentation.

A. Proposed Dataset: Training Data

A substantial training dataset enables the extraction of intricate and nuanced patterns of an object in the image. The quantity of training images in a dataset directly correlates with the performance of the models. As the number of training images remains relatively modest in the existing literature, it potentially gives rise to suboptimal models with limited generalization and potential inaccuracies. To circumvent these problems and enhance the efficacy of rip current detection, we leverage a novel dataset [2] encompassing 110,215 coastline images of Haeundae Beach, Korea. It is 30x times larger than the Zhu et al. dataset which is the current largest rip current dataset with 3792 coastline images. It is a publicly available dataset that is collected by coastline CCTVs, marine observation buoys, and tide stations. It contains both rip current and non-rip current images and is meticulously refined through crowdsourced workforces. While there are other Korean coastlines rip current data also accessible on the same site, we specifically incorporate the Haeundae data as it has a larger number of images comparatively. The Haeundae rip current dataset features four distinct sets of data captured from varying camera perspectives and elevations, namely GLORY, PARA1, PARA2, and SEAC1. In total, our utilization comprises 89,181 coastline images depicting rip currents and 21,034 images devoid of rip currents. All these images are presented in High-Definition (HD) at dimensions of 1920 × 1080 pixels. The bounding box annotations are meticulously aligned with the x and y axes of an image, ensuring consistent reference points. The sample rip current images are shown in Figure 3.

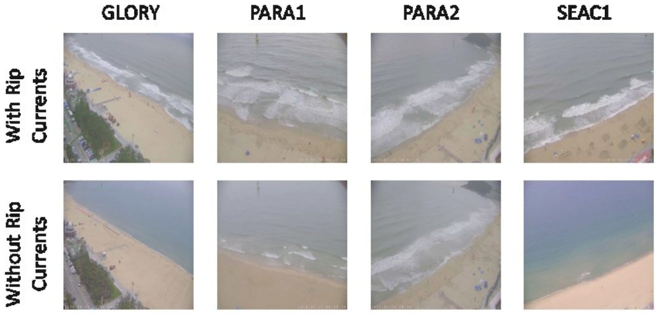


Figure 3. Sample images of both with and without rip currents drawn from the proposed dataset. Images displayed in all four columns are of Haeundae Beach captured in different perspectives.

B. Test Data

A crucial phase in the evaluation of model effectiveness, accuracy, and generalization entails testing the data. Our testing data incorporates the rip current videos sourced from the test datasets of de Silva [15] and Dumitriu [20]. The de Silva dataset comprises 23 video clips, collectively comprising 18,042 frames, while the Dumitriu dataset features 17 videos comprising a total of 24,295 frames. Each testing video has a 20-second duration and a resolution of 1920 × 1080 pixels on average. The trained models underwent rigorous testing, encompassing a comprehensive assessment of their capabilities and performance.

IV. METHODOLOGY

Identifying rip currents in complex and turbulent ocean flow dynamics is undeniably challenging. While several alternative methods exist for rip current identification, deep learning approaches have emerged as the preeminent choice. In this section, we investigate several SoTA object detection models.

A. Faster RCNN

Faster R-CNN [16] is one among the family of region-based CNNs in the realm of computer vision. These object detection models typically comprise a shared feature extraction network alongside independent classification and localization networks. Faster R-CNN introduced the Region Proposal Network (RPN), a dedicated network that performs feature region extraction. RPN employed anchor boxes to extract region proposals more precisely. Furthermore, Faster R-CNN seamlessly integrated Fast R-CNN [21] as the final object detector in its architecture. This design enables the entire model process to execute seamlessly on the GPU, eliminating bottlenecks and enabling end-to-end network training.

B. YOLO v5

The You Only Look Once (YOLO) model series is a widely recognized family of object detection models, distinguished by its single-stage architecture. Among the series of models, the fifth version, the YOLOv5 model was developed by Jocher et al. from Ultralytics [19]. Notably, YOLOv5 exhibits faster convergence, requiring less time per epoch in comparison to its successors [22]. The architecture of YOLOv5 is delineated into three core components: the backbone, the neck, and the head. YOLOv5 employs a Cross Stage Partial Network (CSPNet) bottleneck [23] as the CNN backbone. The CSP models are based on DenseNet, a design aimed at reducing network parameters. These features are combined in the model neck and subsequently forwarded to the head. The model head interprets the combined features to predict the object classes in an image.

C. YOLO v8

A recent state-of-the-art version of the YOLO family is YOLOv8 [24] from the same creators of YOLOv5. YOLOv8 has emerged as a great choice for real-time object detection thanks to its updated architecture, refined convolutional layers in the backbone, and more sophisticated detection head and neck. The YOLOv8 model leverages the CSPDarknet-53 model as the backbone network, which significantly enhances both speed and precision compared to previous iterations. As an anchor-free model, it directly predicts the center of an object rather than its offset. This anchor-free paradigm effectively reduces the number of box predictions, consequently expediting the Non-Maximum Suppression (NMS) process. In the neck of the architecture, features are directly concatenated without necessitating uniform channel dimensions. This divergent architecture results in the reduction of parameter count and the overall size of the tensors. Similar to YOLOv5, YOLOv8 embraces online image augmentation during training. The mosaic data augmentation technique is consistently applied, affording the model exposure to objects in various positions.

V. Experiments

This section encompasses the delineation of the experimental environment setup including hyper-parameter fine-tuning, computational resources utilized, the metrics employed for evaluation, and the utilization of the proposed dataset. In addition, we also investigate the methods of explainable AI with class activation maps to interpret the rip current detection results

A. Environmental Setup

The implementation of all baseline models outlined in the methodology section adheres to Python 3.10.9, PyTorch 2.0.1, and CUDA 11.8.0. The computational tasks, both training and inference, were executed on a dedicated GPU of NVIDIA GeForce RTX 3060 with 12GB of memory. The hyperparameters utilized in the foundational model were retained for training all the baseline models. Depending on the model size, training occurred with batches ranging from 16 to 64, spanning a duration of 100 epochs. Furthermore, we conducted experiments with varying threshold values to gain insights into their influence on the final detection performance.

B. Data Utilization

To facilitate the training process, the images are partitioned into a training set and a validation set. These images are accompanied by ground truth annotations and possess dimensions of 1920 × 1080 with three channels. For the sake of training efficiency, all images are uniformly resized to dimensions of 640 × 640 pixels while training the models. As all images of the dataset are of the same data distribution that is Haeundae Beach, we incorporate data augmentation techniques to prevent the model from overfitting to the original dataset and to enhance the diversity of the training dataset. The augmentation methods we utilized are as follows:

- **Rotation:** By applying random rotations to the images, models can learn about rip currents from different angles.
- **Flip:** Horizontal and vertical flipping help the model to recognize different orientations of rip currents.
- **Scaling:** Resizing images to scales such as 640 × 640 and 320 × 320, allows the models to learn to identify rip currents of varying sizes.
- **Brightness and Contrast Adjustment:** This method aids the models in detecting rip currents under varying levels of illumination and lighting conditions.

C. Evaluation Metrics

To assess the performance and detection efficacy of the baseline models, we have employed four common evaluation metrics, namely Precision (P), Recall (R), mean Average Precision (mAP), and Frames Per Second (FPS). Precision and Recall are computed using the formulations presented in Equations (1) and (2).

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

In the context of our evaluation, TP (True Positive) represents the count of correctly detected rip currents, FP (False Positive) corresponds to the number of erroneously detected rip currents, and FN (False Negative) accounts for the missed rip currents. The primary evaluation metric for our model is mean Average Precision (mAP), with a focus on mAP50. This metric is based on the Intersection over Union (IoU), derived from the Jaccard index which signifies the similarity between two sets. Higher mAP values typically signify superior detection performance. The mean Average Precision is calculated as Equation (3), where k denotes the number of IoU thresholds and N denotes the total number of target categories.

$$mAP = \frac{\sum_{k=1}^{k=N}(AP_k)}{N} \tag{3}$$

Moreover, we assess the model's efficiency by considering training speed and inference time. In the evaluation of the test videos, we analyze the number of frames with precisely detected rip currents, employing the mAP50 metric to gauge the accuracy of each video while scrutinizing cases of failure. The detection speed is quantified in terms of Frames Per Second (FPS) which represents the number of images processed by the model in one second. Higher FPS values indicate faster image processing capability of the model.

D. Class Activation Mapping

The complexity of computer vision models can sometimes make it challenging to extract meaningful insights from their layers. Fortunately, there are XAI tools available to help make sense of these model layers. One such tool is Class Activation Maps (CAM), which combines activation maps from selected convolution layers to provide model explainability. These maps effectively visualize substantial features and are consistent with how humans perceive vision. CAM-based XAI methods

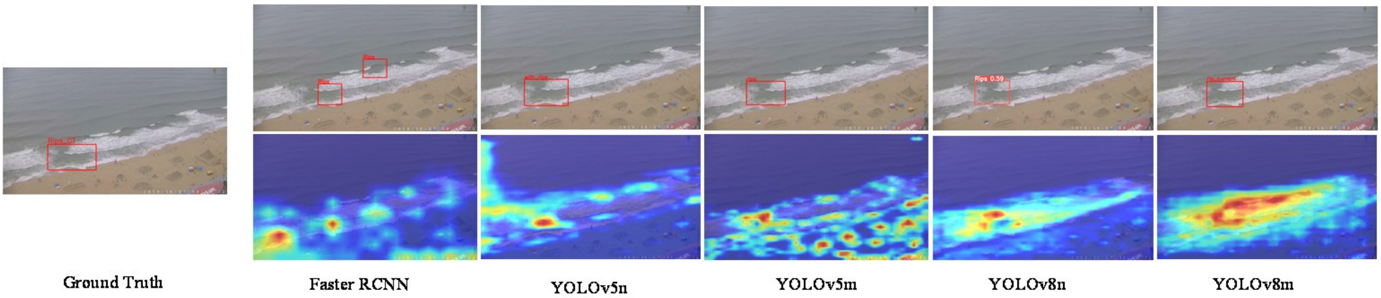


Figure 4. Visualization of obtained results of the state-of-the-art object detection models. For model comparison, same image has been used for inferencing. Each column represents different model outputs except Column 1 which shows the ground truth of the original image. While Row 1 displays the bounding box annotation predictions, Row 2 depicts the class activation map of the corresponding model.

provide insight into how the model is learning and whether the training setup needs any improvement. These methods use partial derivatives to weigh each activation map, making them faster than Region-based XAI methods. However, these saliency maps rely solely on feature maps and typically contain irrelevant features.

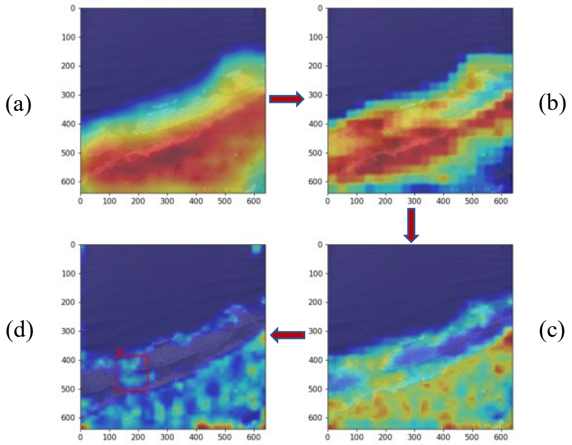


Figure 5. Different stages of activation maps from various layers of the model. Image (a) refers the activated map from bottom layers of the model where Image (d) refers the activated map from top layers of the model.

In this study, we incorporate the Eigen-CAM method, introduced by Muhammad et al. in [25]. Unlike other CAM approaches like Grad-CAM, Grad-CAM++, and XGrad-CAM, the Eigen-CAM stands out for its ease of implementation and seamless integration. It extends CAM by visualizing the principal components of learned features from various convolutional layers (as depicted in Figure 5). By overlaying this heatmap on the input image, the areas that generated the highest levels of activation from the convolutional layers can be emphasized (as shown in Figure 5(d)). It is important to note that as convolutional layers become more compact, they shift from acquiring more comprehensive aspects across the image to capturing tiny features. Consequently, the Eigen-CAM proposed using the Eigenvector and the primary weight matrix to create a heatmap. The n^{th} eigenvector next to the singular value decomposition generates various intriguing activation maps. As a result, this makes it possible to comprehend the amorphous boundaries of a detected rip current in the input image. The authors have expressed the Eigen-CAM as the equation below (Equation 4), where the class activation map, $L_{\text{Eigen-CAM}}$ is obtained by projecting $O_{L=k}$ onto the first eigenvector V_1 in the V matrix.

$$L_{Eigen-CAM} = O_{L=k}V_1 \tag{4}$$

VI. RESULTS

This section discusses the comprehensive analysis of the various object detection SoTA models and their associated performance on our dataset. We engage in a comparative analysis of single-stage object detection models from the YOLO series and a two-stage object detection model of Faster R-CNN. The performance of the models has been assessed both with and without the presence of rip currents.

Table 1. Evaluation loss comparison between the models.

Models	Box loss ↓		Object loss ↓	
	Train	Val	Train	Val
F-RCNN	1.5147	0.8806	1.3456	0.8880
YOLOv5n	0.0348	0.0305	0.0133	0.0146
YOLOv5m	0.9020	0.6651	0.4136	0.2808
YOLOv8n	0.0210	0.0072	0.0154	0.0060
YOLOv8m	0.8077	0.4044	0.2449	0.1218

As Table 1 provides insights into various losses over training and validation datasets, the YOLOv8 nano model has lesser loss values compared to other models. The rip currents detected by the trained models on our dataset are visually depicted in Figure 4. It's imperative that the models should accurately detect rip currents from the given data. However, as the rip currents lack clearly defined boundaries, it causes a significant amount of uncertainty in the scale of actual ground truth bounding boxes [15]. Accordingly, to classify a detection as accurate, the detected bounding box need not have a striking resemblance to the labeled bounding box. Given these characteristics, we adopt a lower IoU threshold for the detection. The Faster RCNN model has demonstrated expected results in detecting rip currents as shown in Figure 4. However, as the RCNN models operate as a two-stage detection algorithm, they encounter notable challenges as the dataset size increases. While the size of the dataset grew, the detection time increased substantially. Timely detection of rip currents holds paramount importance as it directly impacts the safety of swimmers and beachgoers. In response to these challenges, we have explored the capabilities of the state-of-the-art YOLO (You Only Look Once) detection frameworks. YOLO offers distinct advantages compared to the Faster RCNN network, which aligns with the real-time demands of rip current detection, further enhancing beach safety. The advantages of YOLOv5 are more pronounced as illustrated in Tables 1 and 2, where it surpasses Faster RCNN. Notably, the YOLOv5 nano-size model performs better than the YOLOv5 medium-size model.

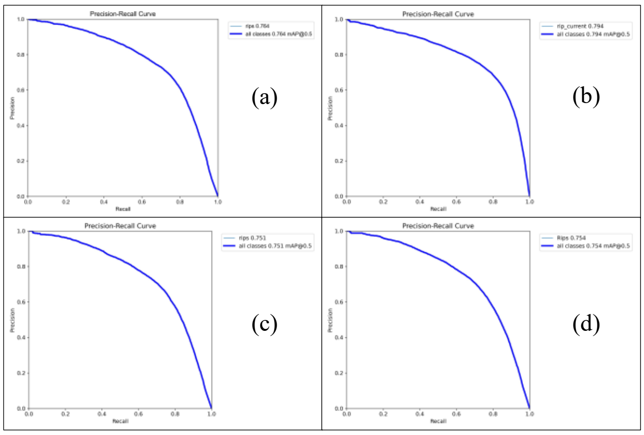


Figure 6. Plots showing the mean Average Precision values of different models with different sizes. In this PR curve plot, while the x axis denotes recall, y axis refers precision. Plot (a) denotes the value of YOLOv5n, Plot (b) denotes the value of YOLOv8n, Plot (c) denotes the value of YOLOv5m, and Plot (d) denotes the value of YOLOv8m.

Table 2. Experimental results of the object detection SoTA models.

Models	Precision	Recall	mAP50	mAP50:95
F-RCNN	0.6434	0.6457	0.6645	0.3482
YOLOv5n	0.7047	0.7271	0.7637	0.4147
YOLOv5m	0.7035	0.7041	0.7512	0.4127
YOLOv8n	0.7139	0.7683	0.7940	0.4259
YOLOv8m	0.7074	0.7004	0.7536	0.4159

The latest model YOLOv8 demonstrates exceptional capabilities in rip current detection, even when multiple rip currents are present. In our comprehensive analysis, the YOLOv8 outperforms the Faster RCNN and YOLOv5 models. As shown in Table 2, the YOLOv8n model attains an overall mAP score of 0.794. Furthermore, the YOLOv8 model size is 14.6 MB, which underscores the compact model size in comparison to other models. This feature is crucial when considering using the model in a coastal early warning system, as a smaller model size has advantages.

VII. Conclusion

We introduced a novel dataset for rip current detection, setting a new benchmark in terms of dataset scale. Employing state-of-the-art models including Faster-RCNN, YOLOv5, and YOLOv8, we conducted a comprehensive comparative analysis encompassing precision, recall, mAP50, and mAP50:95 performance metrics. Among other models, YOLOv8 emerged as the most effective model and demonstrated its suitability for detecting the amorphous nature of rip currents. Additionally, we explored the model explainability tools using the Explainable AI method namely Eigen-CAM. The activation maps interpret the black box model computations at various stages and convert them into comprehensible knowledge. Our future research endeavors are focused on expanding the dataset to encompass a wider array of rip current scenarios, exploring advanced model frameworks, and developing real-time rip current detection systems designed for coastal CCTV cameras. Furthermore, adding additional information such as wave patterns, lighting conditions, and bathymetrical details might also be beneficial in lowering false-negative detections.

References

[1] N. N. W. S. US Department of Commerce, "Surf Zone Fatalities in the United States in 2023: 87".

[2] "AI-Hub." Accessed: Sep. 11, 2023. [Online]. Available: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71297

[3] N. O. and A. A. US Department of Commerce, "Rip Currents - Currents: NOAA's National Ocean Service Education".

[4] "Haeundae riptide sweeps more than 70 people out to sea - The Korea Times." Accessed: Oct. 06, 2023. [Online]. Available: https://www.koreatimes.co.kr/www/nation/2023/10/113_233993.html

[5] M. C. Haller, D. Honegger, and P. A. Catalan, "Rip Current Observations via Marine Radar," *J Waterw Port Coast Ocean Eng*, vol. 140, no. 2, pp. 115–124, Mar. 2014, doi: 10.1061/(ASCE)WW.1943-5460.0000229.

[6] H. D. Kim, K.-H. Kim, H. D.; Kim, K.-H. Kim, H. Kowalewska-Kalkowska, and A. Cedro, "Analysis of Rip Current Characteristics Using Dye Tracking Method," *Atmosphere 2021, Vol. 12, Page 719*, vol. 12, no. 6, p. 719, Jun. 2021, doi: 10.3390/ATMOS12060719.

[7] B. K. Haus, "Remote sensing applied to rip current forecasts and identification," *Chapter*, vol. 8, pp. 133–145, 2011.

[8] Y. Liu and C. H. Wu, "Lifeguarding Operational Camera Kiosk System (LOCKS) for flash rip warning: Development and application," *Coastal Engineering*, vol. 152, p. 103537, Oct. 2019, doi: 10.1016/J.COASTALENG.2019.103537.

[9] J. MacMahan *et al.*, "An Introduction to Rip Currents Based on Field Observations," *https://doi.org/10.2112/JCOASTRES-D-11-00024.1*, vol. 27, no. 4, Jul. 2011, doi: 10.2112/JCOASTRES-D-11-00024.1.

[10] C. Maryan, M. T. Hoque, C. Michael, E. Ioup, and M. Abdelguerfi, "Machine learning applications in detecting rip channels from images," *Appl Soft Comput*, vol. 78, pp. 84–93, May 2019, doi: 10.1016/J.ASOC.2019.02.017.

[11] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *Int J Comput Vis*, vol. 57, no. 2, pp. 137–154, May 2004, doi: 10.1023/B:VISI.0000013087.49260.FB/METRICS.

[12] A. H. Rashid, I. Razzak, M. Tanveer, and A. Robles-Kelly, "RipNet: A Lightweight One-Class Deep Neural Network for the Identification of RIP Currents," *Communications in Computer and Information Science*, vol. 1333, pp. 172–179, 2020, doi: 10.1007/978-3-030-63823-8_21/COVER.

[13] A. H. Rashid, I. Razzak, M. Tanveer, and A. Robles-Kelly, "RipDet: A Fast and Lightweight Deep Neural Network for Rip Currents Detection," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2021-July, Jul. 2021, doi: 10.1109/IJCNN52387.2021.9533849.

[14] A. H. Rashid, I. Razzak, M. Tanveer, and M. Hobbs, "Reducing rip current drowning: An improved residual based lightweight deep architecture for rip detection," *ISA Trans*, vol. 132, pp. 199–207, Jan. 2023, doi: 10.1016/J.ISATRA.2022.05.015.

[15] A. de Silva, I. Mori, G. Dusek, J. Davis, and A. Pang, "Automated rip current detection with region based convolutional neural networks," *Coastal Engineering*, vol. 166, p. 103859, Jun. 2021, doi: 10.1016/J.COASTALENG.2021.103859.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Adv Neural Inf Process Syst*, vol. 28, 2015, Accessed: Sep. 08, 2023. [Online]. Available: https://github.com/

[17] A. de Silva *et al.*, "RipViz: Finding Rip Currents by Learning Pathline Behavior," *IEEE Trans Vis Comput Graph*, 2023, doi: 10.1109/TVCG.2023.3243834.

[18] D. Zhu, R. Qi, P. Hu, Q. Su, X. Qin, and Z. Li, "YOLO-Rip: A modified lightweight network for Rip currents detection," *Front Mar Sci*, vol. 9, p. 930478, Aug. 2022, doi: 10.3389/FMARS.2022.930478/BIBTEX.

[19] G. Jocher *et al.*, "ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation," *Zenodo*, 2022.

[20] A. Dumitriu, F. Tatui, F. Miron, R. T. Ionescu, and R. Timofte, "Rip Current Segmentation: A Novel Benchmark and YOLOv8 Baseline Results," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1261–1271, Jun. 2023, doi: 10.1109/CVPRW59228.2023.00133.

[21] R. Girshick, "Fast R-CNN," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, 2015.

[22] M. Durve *et al.*, "Benchmarking YOLOv5 and YOLOv7 models with DeepSORT for droplet tracking applications," *The European Physical Journal E 2023 46:5*, vol. 46, no. 5, pp. 1–7, May 2023, doi: 10.1140/EPJE/S10189-023-00290-X.

[23] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," Nov. 2019.

[24] "YOLOv8 - Ultralytics | Revolutionizing the World of Vision AI." Accessed: Sep. 20, 2023. [Online]. Available: https://ultralytics.com/yolov8

[25] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class Activation Map using Principal Components," *Proceedings of the International Joint Conference on Neural Networks*, Jul. 2020, doi: 10.1109/IJCNN48605.2020.9206626.

[26] N. N. W. S. US Department of Commerce, "Rip Current Photos".

A Test Method for the Convergence of the Metaverse and Blockchain

Tae-gyu Lee*

*Department of Smart Contents, Division of ICT Convergence, Pyeongtaek University, 3825 Sedong-daero Pyeongtaek-si, Gyeonggi,17869, Korea

tglee@ptu.ac.kr

***Abstract*— This paper proposes an effective testing method combining blockchain and metaverse technologies. This research analyses the interaction between blockchain networks and virtual reality worlds, developing a secure and efficient testing process. The paper evaluates the interaction between metaverse and blockchain in various scenarios, offering insights to enhance reliability and security. This study makes a significant contribution to strengthening safety in the digital environment through the fusion of metaverse and blockchain technologies. This study proposes key evaluation factors and formulas to objectively assess the integration of blockchain and the metaverse services. These evaluation metrics can be valuable in the development of assessment tools and simulation instruments.**

***Keywords*—— Blockchain, Metaverse, Cryptocurrency, Testing, Evaluation**

I. Introduction

In recent times, despite the ongoing economic and social issues and concerns associated with cryptocurrencies and cryptocurrency exchanges, the role of cryptocurrencies as a digital economy and an alternative means of currency continues to expand steadily [1, 2]. Especially, following Blockchain 2.0, various emerging technologies like smart contracts, NFTs, and metaverse platforms have extended their reach, forming new markets and enhancing utility and stability [3, 4]. Nevertheless, there remain limitations in ensuring mutual trust among stakeholders, including developers, operators, and users, within the cryptocurrency and exchange ecosystems, and the fundamental security and reliability of virtual trading systems are under threat [5, 6, 7].

Currently, the convergence of metaverse and blockchain technologies is presenting new possibilities in the digital domain [8]. Metaverse is realized through virtual reality, while blockchain offers security, transparency, and reliability in a distributed manner. This convergence provides opportunities and challenges in terms of testing and ensuring the reliability of system operations within a dynamic ecosystem.

A robust testing process tailored to metaverse-supported blockchain applications becomes essential. Conventional testing methodologies may often fail to address the complexities arising from the immersive and interconnected virtual environments of the metaverse. Addressing these challenges necessitates new approaches to guarantee uninterrupted functionality and security of blockchain networks in such environments.

This paper proposes a comprehensive framework for designing a testing process for metaverse-supported blockchain applications. Our research aims to provide a framework that allows thorough testing of blockchain-based metaverse applications, taking into consideration the unique features and challenges resulting from this convergence. By conducting a systematic investigation into metaverse interactions, blockchain integration, and security considerations, we contribute to the development of a reliable testing methodology. Our goal is to support the growth of both metaverse and blockchain technologies.

This paper is described as follows: Section 2 presents the related works, while Section 3 proposes the metaverse and blockchain interaction and convergence testing process. In Section 4, we analyse key issues related to the proposed testing process. Finally, Section 5 provides conclusions.

II. Related Works

Previous studies in this context have primarily focused on the security and safety of blockchain and cryptocurrencies. Additionally, research on the combination of the metaverse and blockchain is increasing, but there is still a lack of focus on the testing process [5, 9, 10]. These studies emphasize the need for the development of effective testing tools and methodologies in the context of metaverse and blockchain environments [11, 12].

The fusion of the metaverse and blockchain is reshaping the digital economy, fostering novel business models and value creation. However, ensuring success requires a robust testing process to safeguard stability and security, enabling trust and unlocking opportunities in finance, gaming, art, and beyond.

Existing research has shown the absence of new methodologies and testing processes for the convergence of the metaverse and blockchain [10, 13, 14]. Effective testing methods that consider the immersive and interconnected virtual environments of the metaverse and the distributed nature of blockchain are lacking, potentially leading to security and stability issues. Therefore, this study aims to propose a new testing process that considers the characteristics of the metaverse and blockchain.

Through this, we aim to design and propose an effective testing process for the metaverse and blockchain convergence environment. Figure 1 illustrates the design model for testing the integration of the metaverse and blockchain. This model is proposed to concretize the testing and authentication procedures while enhancing practicality [15].

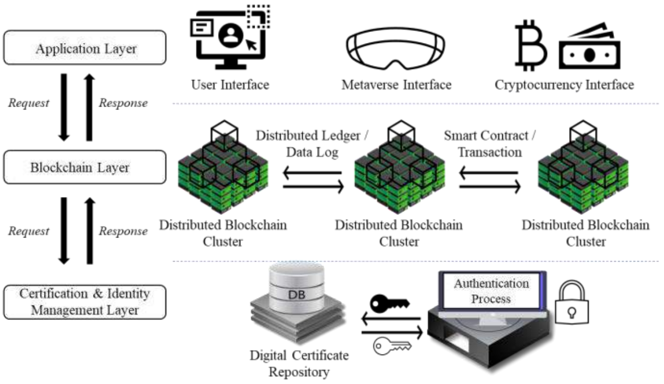

Application Layer
User Interface
Metaverse Interface
Cryptocurrency Interface
Request
Response
Blockchain Layer
Distributed Ledger / Data Log
Smart Contract / Transaction
Distributed Blockchain Cluster
Distributed Blockchain Cluster
Distributed Blockchain Cluster
Request
Response
Certification & Identity Management Layer
DB
Digital Certificate Repository
Authentication Process

Figure 1. Metaverse and Blockchain Convergence Testing Operating Environment

For conducting metaverse and blockchain integration tests, a hierarchical structure of converged application services can be organized into three layers. First, the top-tier Application Layer encompasses applications and smart contracts running in the metaverse environment. It operates user interfaces, virtual environments, and services. Second, the Blockchain Layer is based on blockchain technology and houses the distributed ledger and smart contract platform. Transaction processing and data recording occur within the blockchain. Nodes of the blockchain network and the smart contract execution environment belong to this layer. Third, the Certification & Identity Management Layer is a core component of the certified structure, responsible for identity verification and digital certificate management. It manages user identities and provides authentication and permissions needed by smart contracts and applications. This layer includes digital certificate repositories and secure authentication processes.

Through this layered model, when conducting blockchain integration tests in a metaverse distributed environment, smart contracts and applications running within the metaverse environment are implemented in the Application Layer, transaction processing and data storage occur in the Blockchain Layer, and user identities are managed and protected in the Certification & Identity Management Layer. This layered structure enables effective testing and security verification.

Furthermore, the metaverse and blockchain integration testing process model is used to identify the key components and issues in each testing phase and establish input and output elements. This research aims to provide solutions for testing issues related to stable smart contract service expansion for the integration of metaverse services and blockchain networks.

III. Metaverse and Blockchain Integration Testing Process

In this section, we aim to propose an integrated testing process to effectively test this interaction. It outlines the proposed testing process, taking into consideration the intricate interaction between the metaverse and blockchain. This process offers a step-by-step approach to effectively test metaverse-supported blockchain applications, accommodating various testing scenarios and interactions. The testing process, designed to address the complex interplay between the metaverse and blockchain, provides a systematic framework for testing metaverse-supported blockchain applications. Figure 2 illustrates the stepwise approach of this testing process, which can be utilized effectively.

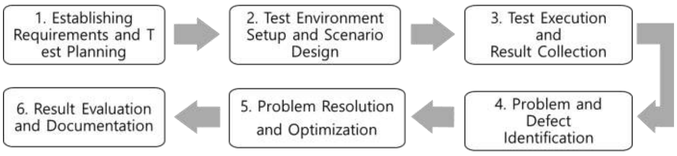

1. Establishing Requirements and T est Planning
2. Test Environment Setup and Scenario Design
3. Test Execution and Result Collection
4. Problem and Defect Identification
5. Problem Resolution and Optimization
6. Result Evaluation and Documentation

Figure 2. Convergence Testing Process

A. *Requirement Definition and Test Planning*

The initial phase of the testing process involves defining the requirements for the metaverse and blockchain application and establishing test objectives.

- **Define Test Objectives and Scope**: Clearly define the objectives for testing the metaverse-supported blockchain application and specify the aspects to be tested.
- **Set Up Test Environment:** Establish the test environment by integrating the metaverse and blockchain, and configuring the necessary hardware and software.
- **Allocate Test Resources:** Assign resources, including the testing team, testers, developers, etc., and define their roles.
- **Develop Test Plans:** Document comprehensive test plans, including schedules, task allocation, test case design, execution plans, and risk management.
- **Design Test Cases:** Create test cases based on the defined objectives and scope, specifying the expected behavior under different scenarios.
- **Generate Test Data:** Generate or import the required test data to simulate real-world situations.
- **Prepare Test Environment and Data:** Configure the test environment, load test data, initialize metaverse and blockchain networks, and apply necessary settings.
- **Review, Approve, and Execute Tests**: Review and gain approval for the test plans from the testing team and stakeholders, then execute the tests according to the established schedule and process.

These steps allow for the clear definition of requirements for metaverse-supported blockchain applications and the establishment of efficient test plans. This systematic approach enables the comprehensive testing of the interaction between the metaverse and blockchain, as well as the identification and improvement of any issues.

B. Test Environment Setup and Scenario Design

In this phase, the test environment for the metaverse and blockchain integration testing is set up, and various test scenarios are designed. The environment encompasses the integration of the metaverse environment and the blockchain network.

- **Environment Setup:** Construct and configure the testing environment, which includes integrating the metaverse and blockchain, along with setting up necessary hardware and software components. Ensure that the environment aligns with testing objectives.
- **Scenario Definition:** Define various test scenarios representing specific situations or use cases. Each scenario is intended to test the interaction between the metaverse and blockchain.
- **Test Data Preparation**: Prepare the required test data for the defined scenarios, simulating specific states and events in each scenario.
- **Test Case Design:** Design test cases based on the defined scenarios, specifying in detail the behavior to be verified.
- **Environment Validation:** Verify that the configured test environment functions as expected, addressing any issues that arise, and loading and preparing the test data.
- **Test Execution and Result Analysis:** Execute tests based on defined scenarios, record the results, identify and document any problems or errors, and analyze the results. Summarize the findings in a report, which may include proposed solutions for addressing issues and improvements.

Through the setup of the test environment and scenario design, various situations are created to test the metaverse-supported blockchain application.

C. Test Execution and Result Collection

In this phase, tests are executed in the previously established environment, and results are collected. This stage involves simulating actions such as transactions within the metaverse, smart contract execution, NFT issuance, and trading. The success or failure of the executed test cases is recorded, and data related to performance is collected.

- **Test Execution:** Execute tests based on predefined scenarios, simulating interactions between the metaverse and blockchain.
- **Real-time Monitoring:** Continuously monitor the test environment and scenarios during testing, checking for operational issues.
- **Result Collection and Performance Measurement:** Collect results after each scenario, document state changes, and measure system performance, including latency, response time, and throughput.
- **Error Documentation and Log Analysis:** Document errors and issues encountered during testing, including error types, timestamps, and their impact. Analyze log files to identify root causes.
- **Evaluation and Reporting:** Evaluate collected results against test objectives, identify issues, and document test outcomes in a report, including test success/failure, identified problems, and proposed solutions. Complete testing and clean up the test environment and resources once all tests are finished.

Through these steps, the metaverse and blockchain integration application can be effectively tested, and results can be collected to address and improve identified issues.

D. Identifying Issues and Defects

This process involves identifying and analysing issues and defects that occurred during test execution. During this stage, all errors and unexpected situations related to metaverse and blockchain interactions are recorded, and the root causes of issues are determined.

- **Error Identification:** Identify errors and deviations from expected system behavior during test execution.
- **Defect Classification and Severity Assessment:** Categorize and evaluate defects by type and severity, considering their impact on system functionality.
- **Root Cause Analysis and Issue Reporting**: Analyze the causes of defects, document them in a report, including descriptions, root causes, and proposed fixes.
- **Priority Assignment and Issue Tracking:** Prioritize defects for resolution, track their status, and manage them until they are addressed.
- **Test Re-execution:** After fixing defects, re-run relevant test cases to confirm issue resolution.

Through these processes, issues, and defects in metaverse-integrated blockchain applications can be identified and appropriate corrective actions can be taken to address them.

E. Issue Resolution and Optimization

In this phase, identified issues and defects are addressed, and the system is optimized. This involves fixing errors and implementing measures to enhance stability and reliability.

- **Issue Resolution:** Address identified issues and defects, including root cause analysis and corrective actions.
- **Re-Testing and System Testing:** Verify the modifications and assess their impact on system behavior, ensuring that previously identified problems are resolved without introducing new defects.
- **Performance and Security Enhancement:** Optimize system performance and strengthen security measures as needed, including improvements in latency, response times, scalability, and security features.
- **Performance Evaluation:** Evaluate the performance of the optimized system to confirm improvements are effective.
- **Issue Tracking and Documentation:** Manage and track resolved issues and defects to prevent recurrence, and update issue reports to reflect the status and outcomes of resolution and optimization efforts.

Through these activities, metaverse-integrated blockchain applications can be maintained in a stable and optimized state, with issues continually addressed and improvements made over time.

F. Evaluation and Documentation of Results

In this final phase, the test process's results are evaluated and documented. The aim is to summarize the outcomes of the testing process, confirm the reliability and safety of the metaverse and blockchain application, and document the scenarios and procedures used for future quality management and reporting.

- **Result Evaluation and Performance Measurement:** Evaluate the resolution of identified issues and measure performance improvements achieved through testing.
- **Result Report and Sharing:** Prepare a comprehensive result report that includes testing objectives, executed test cases, issues, resolution methods, optimization outcomes, and performance measurements. Share this report with stakeholders and teams involved in testing, gathering feedback.
- **Documentation and Quality Assurance:** Archive the result report in project records and archives for future reference. Assess overall system quality based on the results and develop strategies for improvement, aiming to prevent similar issues in future phases.
- **Process Improvement:** Evaluate the test process and identify areas for enhancement to ensure more efficient testing in future projects.

Through these evaluation and documentation activities, the testing outcomes and conclusions for metaverse-integrated blockchain applications are clearly summarized. They provide information for future improvement and maintenance tasks.

IV. Analysis of Key Issues in the Proposed Testing Process

In this section, we analyse the strengths and limitations of the proposed testing process and aim to enhance our understanding of the interaction between the metaverse and blockchain.

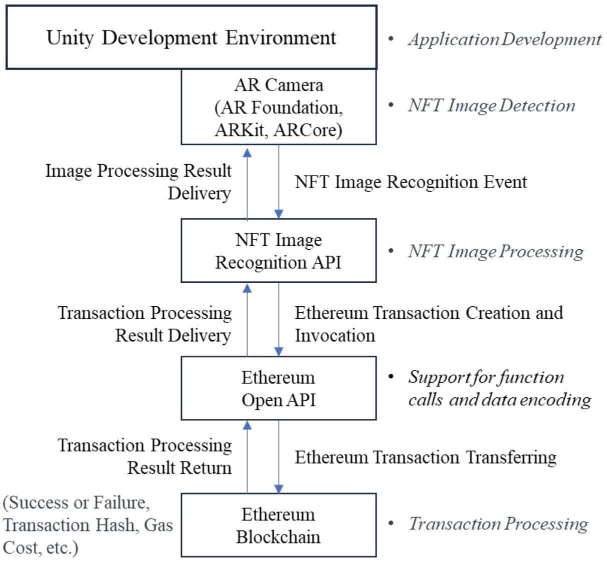

Unity Development Environment
• Application Development
AR Camera (AR Foundation, ARKit, ARCore)
• NFT Image Detection
Image Processing Result Delivery
NFT Image Recognition Event
NFT Image Recognition API
• NFT Image Processing
Transaction Processing Result Delivery
Ethereum Transaction Creation and Invocation
Ethereum Open API
• Support for function calls and data encoding
Transaction Processing Result Return
Ethereum Transaction Transferring
(Success or Failure, Transaction Hash, Gas Cost, etc.)
Ethereum Blockchain
• Transaction Processing

Figure 3. Metaverse and Blockchain Convergence Deveopment Environment and Routines

The following issues can be considered as key evaluation factors for conducting tests on the integration of blockchain and the metaverse.

TABLE 1. Evaluation Metrics for Convergence Test Process

Metric for Element	Explanation	Unit
Interaction Score	This metric can assign a high score when smooth data and asset transfer between blockchain and metaverse technologies is possible.	relative scores
Smart Contract Score	It is an evaluation metric for assessing smart contracts within the metaverse.	relative scores
Security and Safety Score	It is a function used to evaluate security and safety in the metaverse and blockchain integration testing.	relative scores
Improvement Score	It is a function used to evaluate the improvement and optimization of the testing process.	relative scores

A. *Interaction Issues Between the Metaverse and Blockchain*

The interaction between the metaverse and blockchain presents various complexities and challenges. This section analyses the key issues that can arise due to the interaction between these two technologies. For instance, difficulties in reconciling blockchain transaction processing with real-time interactions within the metaverse and concerns related to security and personal data protection are addressed. To analyse the interaction between the metaverse and blockchain, we define an evaluation function based on various weighted factors, including interoperability weight ($W_{interoperability}$), security level weight ($W_{security}$), performance and processing speed weight ($W_{performance}$), cost efficiency weight ($W_{cost\_efficiency}$), and community and developer support weight ($W_{community\_support}$) as seen in Equation (1).

$$\begin{aligned} Interaction\_Score = {} & W_{interoperability} * Interoperability \\ & + W_{security} * Security\_Level \\ & + W_{performance} * Performance \\ & + W_{cost\_efficiency} * Cost\_Efficiency \\ & + W_{community\_support} \\ & * Community\_and\_Developer\_Support \end{aligned} \tag{1}$$

Using this evaluation formula, we can assess the interaction between the metaverse and blockchain from various perspectives, with each weight adjusting the importance of its respective aspect.

B. *Smart Contract Issues Within the Metaverse*

The execution of smart contracts within the metaverse environment can pose complex issues. Errors or vulnerabilities in smart contracts can have a significant impact on interactions within the metaverse and integration with blockchain. Key variables related to smart contracts, such as metaverse integration weight ($W_{metaverse\_integration}$), smart contract security weight ($W_{security}$), smart contract functionality weight ($W_{functionality}$), smart contract efficiency weight ($W_{efficiency}$), and smart contract scalability weight ($W_{scalability}$), can be considered in the metaverse. The evaluation function for analysing smart contracts within the metaverse, taking these weight variables into account, is as seen in Equation (2).

$$\begin{aligned} Smart\_Contract\_Score \ = \ & W_{metaverse\_integration} * \ Metaverse\ Integration \\ & + W_{security} * \ Security \ + \ W_{functionality} * \ Functionality \\ & + W_{efficiency} * \ Efficiency \ + \ W_{scalability} * \ Scalability \end{aligned} \tag{2}$$

This formula allows for the evaluation of various aspects of smart contracts, with each weight representing the relative importance of the respective aspect.

C. Security and Safety Issues

The combination of the metaverse and blockchain introduces new considerations in terms of security and safety. Key concerns include user data protection, vulnerabilities in smart contracts, the application of encryption technologies, and more. As variables for analysing security and safety in metaverse and blockchain integration testing, key weights can be considered, such as metaverse security weight ($W_{metaverse\_security}$), blockchain security weight ($W_{blockchain\_security}$), data privacy weight ($W_{data\_privacy}$), and compliance weight ($W_{compliance}$). The security and safety evaluation function can be defined as seen in Equation (3).

$$\begin{aligned} Security\_and\_Safety\_Score = \ & W_{metaverse\_security} * \ Metaverse\_Security \\ & + W_{blockchain\_security} * Blockchain\ Security \\ & + \ W_{data\_privacy} * \ Data\ Privacy \\ & + \ W_{compliance} * \ Compliance \end{aligned} \tag{3}$$

To calculate the security and safety evaluation score (security_safety_score) for the Metaverse convergence application service, the following steps are taken: Firstly, assign weights to key components: *W_metaverse_security* = 0.3, *W_blockchain_security* = 0.3, *W_data_privacy* = 0.2, *W_compliance* = 0.2. Secondly. consider actual measurement values: *Metaverse_security* = 8.5, *Blockchain_security* = 9.0, *Data_privacy* = 7.5, *Compliance* = 8.8. Thirdly, multiply each component's measurement value by its weight, then sum the results. Finally, calculate the weighted average to obtain the security and safety evaluation score. The calculated score is 0.851, representing the security and safety of the Metaverse convergence application service.

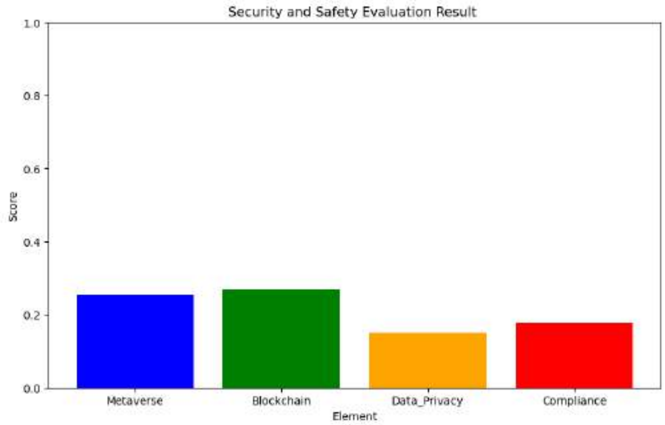


Figure 4. Security and Safety Evaluation Result on Metaverse and Blockchain Convergence Applications

Figure 4 illustrates the relative evaluation results, normalized to a range from 0 to 1, considering the weighting variables for each element. This indicates that data privacy and compliance are relatively more vulnerable compared to Metaverse and blockchain security.

D. Test Process Improvement and Optimization

After analysing the key issues, research is conducted on how to improve and optimize the test process. Consideration is given to improvements in error and defect handling, enhancing safety, performance optimization, and efficient test scenario design. The evaluation function for improving and optimizing the fusion test process for blockchain and the metaverse is as seen in Equation (4).

$$\begin{aligned} Evaluation\_Score = \ & \big(W1 \ * \ \big(1 - Test\ Duration \ / \ 100\big)\big) \\ & + \big(W2 \ * \ \big(Resource\ Utilization \ / \ 10\big)\big) \\ & + \big(W3 \ * \ \big(Test\ Coverage \ / \ 100\big)\big) \\ & + \big(W4 \ * \ \big(Error\ Handling \ / \ 10\big)\big) \end{aligned} \tag{4}$$

Each weight variable, *W1* represents the weight for *Test Duration*, *W2* represents the weight for *Resource Utilization*, *W3* represents the weight for *Test Coverage*, and *W4* represents the weight for *Error Handling*. These weights can be adjusted based on the project's characteristics, and they may vary depending on the project's goals and priorities for improving the testing process.

V. Conclusion

In conclusion, this paper has presented a comprehensive framework for testing the fusion of the metaverse and blockchain. It highlights the significance of a testing process that considers the intricate interactions between these technologies and underscores the advantages and outcomes it can yield. In Section 4, various evaluation function models applicable to the metaverse and blockchain integration testing process were presented. Based on these diverse evaluation models, improved simulations can be conducted to obtain accurate prediction data. In this study, the development and application of the testing process are expected to play a crucial role in the reliability and security of metaverse and blockchain applications.

In terms of future research directions, there is a need for the development of improved testing methodologies and tools to adapt to these changes. It is possible to develop assessment tools and simulation instruments. Furthermore, by analyzing the evaluation results, the objectivity of the assessment method can be validated.

References

[1] M. Fauzi and N. Paiman, “Bitcoin and Cryptocurrency: Challenges, Opportunities and Future Works,” Journal of Asian Finance Economics and Business, Vol.7 695-704. 2020. DOI: http://doi.org/10.13106/jafeb.2020.vol7.no8.695.

[2] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," Available: https://git.dhimmel.com/bitcoinwhitepaper/, 2008.

[3] Vitalik Buterin, “A Next Generation Smart Contract & Decentralized Application Platform,” Ethereum White Paper, Vol.3 Issue37, 2-1, 2014.

[4] Andrea Bonaceto, “Non-Fungible Tokens (NFTs)-The New Paradigm,” Eterna Capital, Mar. 30, 2021, https://eternacapital.medium.com/non-fungible-tokens-nfts-the-new-paradigm-6da78a85e73f

[5] J. Leng, M. Zhou, L.J. Zhao, Y. Huang and Y. Bian, "Blockchain Security: A Survey of Techniques and Research Directions," IEEE Transactions on Services Computing, 2020. DOI: http://doi.org/10.1109/TSC.2020.3038641

[6] J. Bonneau, A. Miller, J. Clark, A. Narayanan, J. A. Krol, E. W. Felten, “SoK: Research Perspectives and Challenges for Bitcoin and Cryptocurrencies,” 2015 IEEE Symposium on Security and Privacy, 104-121, 2015, DOI: http://doi.org/10.1109/SP.2015.14

[7] Z. Chen, J. Wu, W. Gan and Z. Qi, “Metaverse Security and Privacy: An Overview,” IEEE BigData, Nov. 2022, DOI: https://doi.org/10.48550/arXiv.2211.14948

[8] Q. Yang, Y. Zhao, H. Huang, Z. Xiong, J. Kang and Z. Zheng, “Fusing Blockchain and AI With Metaverse: A Survey,” IEEE Open Journal of the Computer Society, Vol.3, 122-136, 2022, DOI: http://doi.org/10.1109/OJCS.2022.3188249

[9] T.T. Huynh, T.D. Nguyen, and H. Tan, “A Survey on Security and Privacy Issues of Blockchain Technology,” 2019 International Conference on System Science and Engineering (ICSSE), 362-367, 2019, DOI: https://doi.org/10.1109/ICSSE.2019.8823094

[10] C. Lal and D. Marijan, “Blockchain Testing: Challenges, Techniques, and Research Directions,” Mar. 2021, DOI: https://doi.org/10.48550/arXiv.2103.10074

[11] T. H. The, T. R. Gadekallu, W. Wang, G. Yenduri, P. Ranaweera, Q.V. Pham, D.B. Costa and M. Liyanage, “Blockchain for the metaverse: A Review,” Future Generation Computer Systems, Vol.143, 2023, 401-419, DOI: https://doi.org/10.1016/j.future.2023.02.008.

[12] Z. Zheng, S. Xie, H.N. Dai, W. Chen, X. Chen, J. Weng and M. Imran, “An Overview on Smart Contracts: Challenges, Advances and Platforms,” Future Generation Computer Systems, Vol.105, Apr. 2020, 475-491, DOI: https://doi.org/10.1016/j.future.2019.12.019

[13] M. Krichen, M. Ammi, A. Mihoub and M. Almutiq, “Blockchain for Modern Applications: A Survey,” Sensors 2022, 22, 5274, DOI: https://doi.org/10.3390/s22145274

[14] T. H. Im, "Trends in Standards and Testing and Certification Technology - Current Status of International Standards for Software Testing (ISO/IEC/IEEE 29119)," TTA Journal, No. 167, Telecommunications technology Association, 96-101, Sep 2016.

[15] T. Lee, “Blockchain and Cryptocurrency Distributed Testing Methods,” International Journal of Internet, Broadcasting and Communication, Vol.14 No.1, Feb 2022, 1-9, Available: https://www.earticle.net/Article/A409169

Tae-Gyu Lee (BSc’92–MSc’96–PhD’06) received the B.Sc. degree from Kunsan National University, Kunsan, Korea in 1992, the M.Sc. degree from Soongsil University, Seoul, Korea in 1996, and the Ph.D. degree from Korea University in 2006. He is currently a Professor in the Dept. of Smart Contents, Division of ICT Convergence, Pyeongtaek University, Gyeonggi, Korea from 2018. He has been a Professor in the Support center for Field Practice Education, WonKwang University, Jeonbuk, Korea for 2014-2018. He has been a Professional Researcher in Advanced Convergent Technology R&D Group, Korea Institute of Industrial Technology (KITECH), Ansan, Korea for 2009-2013. He has also been a President in the JIGUNET Corporation, Seoul, Korea, from 1999. His research interests are in distributed systems, ubiquitous computing, middleware, networks, wearable and robot computing. Prof. Lee is an honorary member of the Korea Information Processing Society and actively serves as a reviewer for ICACT. Furthermore, he has fulfilled the role of a judge at the 10th International Abilympics.

Time-frequency Analysis for Validating Prognostics Algorithms of Rolling Element Bearings

Guanhua Zhu[*], Xiaoling Xu[**], Qing Zhong [**], Bing-Yuh Lu[**], Yushen Lu [**], Guangming Xu [**], Yumeng Zhou [**], Ziyi Jiang [**], Kai Sun [**], Minhao Wang [**]

[*]Guangdong Provincial Key Lab.of Petrochemical Equipment and Fault Diagnosis, Guangdong University of Petrochemical Technology, Maoming City, Guangdong, China
[**]Faculty of Automation, Guangdong University of Petrochemical Technology, Maoming City, Guangdong, China

FranklinLu888@outlook.com

Abstract—**This study employed the time-frequency analysis to compute some of the XJTU-SY bearing datasets and aimed at the investigation of spectrogram of the raw data of the datasets. The methods of this study are divided into 4 parts: (1) spectrogram, (2) XJTU-SY bearing datasets, (3) equipment and, (4) 2D correlation. The results show the dominant reasons of malfunction of the machine occur in the duration of the 75th to 100th minutes. Both 2D correlation coefficients of the spectrograms of horizontal and vertical vibrations in the 100th and 123th minutes are larger than 0.8 because rotation of the roller entered a distinguished state of malfunction in the 100th minute. The inner damage is enhanced step by step. The interpretation of VEs and HEs is helpful to detect the fault diagnosis of the roller. The further studies will test more data, and add more algorithms for the accurate diagnosis.**

Keywords—**XJTU-SY bearing datasets, spectrogram, correlation coefficient, episode, roller**

I. Introduction

This study employed the time-frequency analysis to compute some of the XJTU-SY bearing datasets which are supported by the Institute of Design Science and Basic Component at Xi'an Jiaotong University (XJTU), Shaanxi, P.R. China and the Changxing Sumyoung Technology Co., Ltd. (SY), Zhejiang, P.R. China [1]. Time-frequency analysis owns the benefits of obvious presentation of the fundamental and harmonic components of a segment of a signal. Therefore, some problems of the rolling elements can be detected by this analysis because of the periodic touches in the disorder place of the rotation of the roller. Therefore, we propose the method of time-frequency analysis to be the pre-processing for validating prognostics algorithms of rolling element bearings.

Prognostics and health management (PHM) is very helpful to (1) ensure the safe operation of machinery, (2) improve the productivity, and (3) increase economic benefits [2]. Many studies employed the XJTU-SY bearing datasets to explore the feasibilities of the aforementioned 3 categories of performances. For example, Wu et al. have developed a hybrid deep-learning model based on CNN and gcForest [3]. Yao *et al.* proposed a method using the ability of the 1D-CNN to extract signal features of the signals in the XJTU-SY bearing datasets [4]. Zhao *et al.* [5] proposed a method of the signal-to-signal translation into the field of data-driven fault diagnosis of bearings and gears. The performance and excellences of the proposed method are displayed in the accuracy of detections. The signal-to-signal translation is a feasible solution. Furthermore, the DL-based fault classification models in Zhao *et al.s'* study present the better performance on the four public datasets. Yang *et al.* proposed a classical fuzzy C-means method was used to classify the fused signals from the model [6]. Gao *et al.* employed many statistical approaches to explore the dataset. They improved the accuracy because time-domain features were extracted from the bearing vibration signal as well as Bayesian model of state parameters and bearing life is established [7]. Ma *et al.* used the method of multistep dynamic slow feature analysis [8]. The results shows that the interpretability of the fault information based on five monitoring indicators and the detection rate of unsteady process are improved.

However, the applications of time-frequency analysis were widely used in many fields, such as lung sound diagnosis [9–11] and speech recognition [12-14]. Furthermore, spectrogram is a very crucial tool to display the results of time-frequency analysis. Therefore, some studies related to the XJTU-SY bearing datasets were inserted spectrograms as a method for analysis. Zhang et al. employed the data of spectrogram as an input of improved ConvNext[15]. Cheng *et al.* compared the spectrograms between trained and untrained data {16}. The function of spectrogram supported an efficient method to thr studies.

Our study aimed at the investigation of spectrogram of the raw data of the XJTU-SY bearing datasets. Because of the physical factors of the rotation of the roller, a direct observation of spectrograms must show some findings to explain the fault diagnosis by the computation through the XJTU-SY bearing datasets.

II. Methods

The methods of this study are divided into 4 parts: (1) spectrogram, (2) XJTU-SY bearing datasets, (3) equipment and, (4) 2D correlation. The 3 parts are described as follows:

A. Spectrogram

The element x[n] at the point of time n in a spectrogram is defined as (Haykin, Van Veen, 1998),

$$|X_{\tau}(j\Omega)|^2 = \left|\sum_{-\infty}^{\infty} x(n)w^*(n-\eta)e^{-j\Omega n}\right|^2 \qquad (1)$$

where w*(n-η) is a Kaiser window function which is defined as:

$$\mathrm{w(n)} = \begin{cases} \dfrac{I_o\{\beta[1-(\frac{n-\eta}{n})^2]\}}{I_o(\beta)} & n = 0,1,2,\cdots,L \\ 0 & otherwise \end{cases} \qquad (2)$$

where η= (L−1)/ 2 and I_0 is the zeroth-order modified Bessel function of the first kind. When β = 0, the Kaiser window becomes a rectangular window, as well as, when β ≥ 0 parameter allows for a trade-off between side lobe amplitude and main lobe width. In this case, L=500 and β = 5. The operator $e^{(-j\Omega n)}$ is the same to discrete Fourier transform (FFT) as a phasor operation. The spectrogram presents the relationship of the magnitude (in dB) of time-independent discrete Fourier transform v.s. time. The number of a frame for short FFT is 512, and the number of overlapped pints is 475. The sampling frequency of the explored signals is 25600 Hz. The computation compiled in MATLAB 7 (MathWorks, USA) development environment.

B. XJTU-SY bearing datasets

In this study, we focused on Bearing1_1 whose conditions are 123 files, 2 hours and 3 minutes lifetime, and the fault element of outer race in the XJTU-SY bearing datasets. In a file, there are individually 32768 points in horizontal and vertical sensing, and the sampling frequency is 25600 Hz. Therefore, the length of the record in a file is 1.28 seconds. The data record occurred at the beginning of 1.28 seconds in every minute.

C. Equipment

The type of tested bearings is LDK UER204 which has been described in reference [1] in details.

D. 2D correlation coefficient (r)

The equation of 2D correlation coefficient of the two arrays, A_{mn} and B_{mn}, can be written as:

$$\mathbf{r} = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{(\sum_m \sum_n (A_{mn} - \bar{A}^2)(B_{mn} - \bar{B}^2)}} \qquad (3)$$

where the average of the elements in A_{mn} is denoted by $\bar{A}$, and the average of the elements in B_{mn} is denoted by $\bar{B}$.

III. Results

A. Spectrograms

The recorded signals and corresponding spectrograms in the 1st, 25th, 50th, 75th, 100th, 123th minutes are shown in Fig.s 1 to 6. The left sides are the horizontal vibrations, and right sides are the vertical vibrations.

In Fig. 1, there are some unclear vertical episodes (VE). It means some little pulses of pressure detected. Then, the number of VE increases in Fig. 2. In addition, the VE is much ably visual in the range of higher frequency in the spectrograms. The looks of Fig. 2 and Fig. 3 are very similar. However, the looks between Fig. 3 and Fig. 4 are different in vertical vibrations. Sequentially, the looks of Fig. 4, 5, and 6 are much different one another. Especially, in Fig. 6, the VEs are obviously increasing in the range of the low frequency.

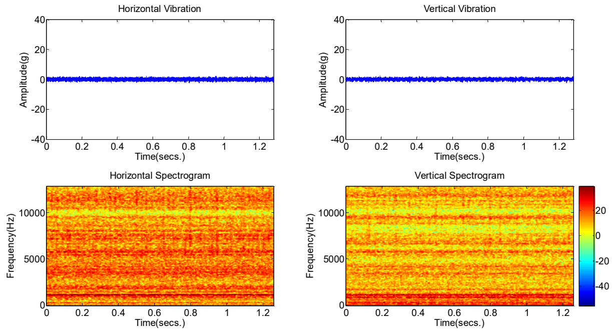


Figure 1. The recorded signals and corresponding spectrograms int the 1st minute.

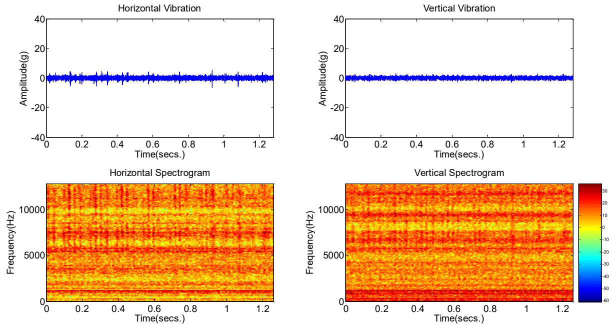


Figure 2. In the 25th min.

B. 2D correlation coefficients

To check the similarities of the spectrograms in Fig.s 1 to 6, we compute the r of spectrograms one another in horizontal and vertical vibrations, respectively. The results are shown in Tables 1 and 2. The r of horizontal and vertical vibrations are presented in Table 1 and Table 2 as well as denoted by r_h and r_v, respectively. In the 2 tables, A is the index of the column and B is the index of the raw. The indexes are the time point in minutes of 1st, 25th, 50th, 75th, 100th, and 123th, i.e. the corresponding time points in Fig.s 1 to 6. For example, $r_h(1,25) = 0.4039$, and $r_h(25,50) = 0.3875$. Consequently, $r_v(1,25) = 0.5992$, and $r_v(100,123) = 0.8007$ in Table 2.

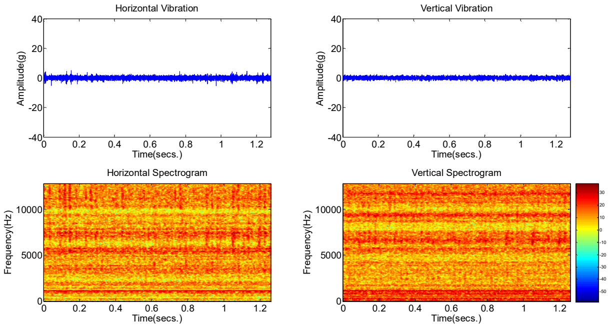


Figure 3. In the 50th min.

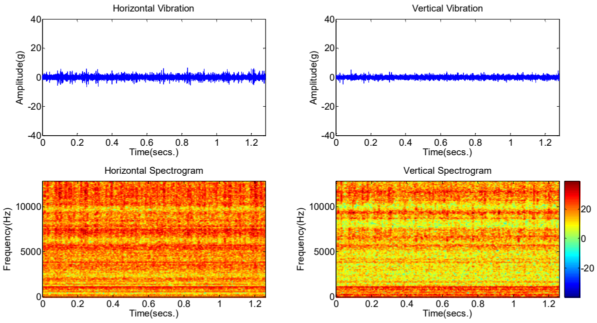


Figure 4. In the 75th min.

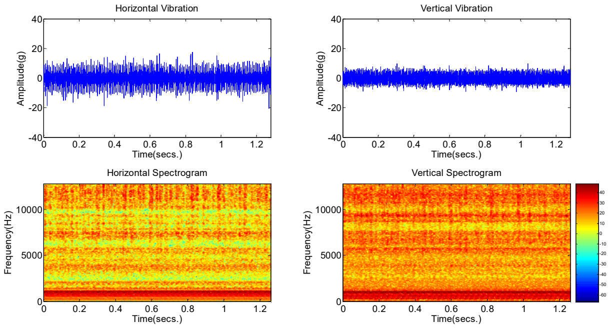


Figure 5. In the 100th min.

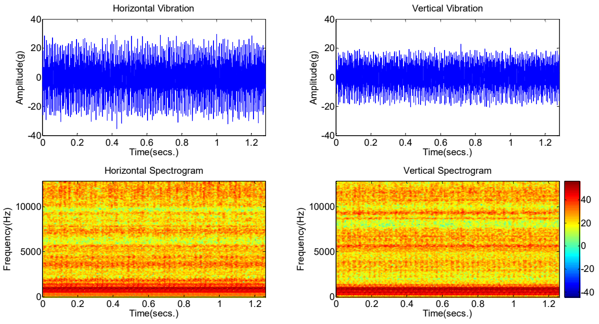


Figure 6. In the 123th min.

IV. Discussion

In fact, Fig. 1 indicates that some weak pulse-like pressures have been detected, because the VE means the frequency component of pulse waveform. The fundamental frequency in the spectrograms is near 250 Hz. The fundamental frequency can be ideally computed as:

$$f_o = f_r * N \qquad (4)$$

Table 1 $r_h(A,B)$

$r_h(A,B)$	*1*	*25*	*50*	*75*	*100*	*123*
1	1.0000	0.4039	0.3504	0.2282	0.3115	0.2909
25	0.4039	1.0000	0.3879	0.3397	0.2136	0.1574
50	0.3504	0.3879	1.0000	0.3402	0.1923	0.1403
75	0.2282	0.3397	0.3402	1.0000	0.1633	0.0736
100	0.3115	0.2136	0.1923	0.1633	1.0000	0.8401
123	0.2909	0.1574	0.1403	0.0736	0.8401	1.0000

Table 2 $r_v(A,B)$

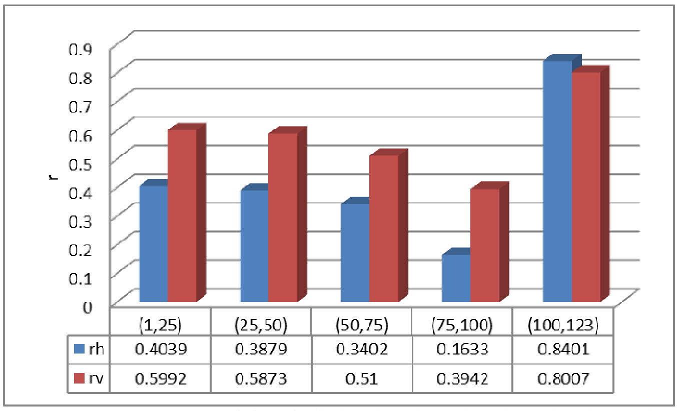
$r_v(A,B)$	*1*	*25*	*50*	*75*	*100*	*123*
1	1.0000	0.5992	0.5389	0.4346	0.4469	0.4765
25	0.5992	1.0000	0.5873	0.5155	0.3914	0.4266
50	0.5389	0.5873	1.0000	0.5100	0.3785	0.4070
75	0.4346	0.5155	0.5100	1.0000	0.3942	0.3756
100	0.4469	0.3914	0.3785	0.3942	1.0000	0.8007
123	0.4765	0.4266	0.4070	0.3756	0.8007	1.0000

where f_o and f_r denoted for fundamental frequency and rotating frequency in the operating condition, respectively. The number of ball is denoted by N. In this case f_r=35 Hz, N = 8. Therefore, f_o = 280 Hz. The frequencies of harmonics are kf_o where k >1 and k are integers.

In Fig. 2, the VEs are stronger; it indicates the collision in a pulse-like manner occurs. Therefore, the prediction of disorder can be made based on the phenomena at this time point. Fig.s 2 and 3 are similar so the rotation of the roller is stable in this duration, i.e. the 25th to 50th minutes. Possibly, it may span the stable duration to the 1st to 75th minutes. However, in the 75th minute, the spectrogram of the vertical vibration changes very obviously. The horizontal episodes (HE) reduced in the range of 1.5 to 5 KHz, i.e. the harmonics are weakening in the range in Fig. 4.. Therefore, the status of rotation of the roller is not smooth as before in vertical direction. The damage in vertical direction occurs at first.

Sequentially, the investigation moves to Fig. 5. Compared with Fig. 4, all figures in Fig. 5 have huge changes. Both amplitudes of horizontal and vertical vibration signals are increased. In addition, the spectrogram of horizontal vibration shows the HEs become discontinuous, and VEs become much obvious. That is the pulses are strong, and the harmonics are not continuous. This releases a message to users that some impact forces occur in the machine. It is very possible to make the machine break in a short time. In Fig. 6, the impact forces are much strong. The VEs are almost become vertical lines to be pure pulses in time domain in both horizontal and vertical vibrations. The inner damages of the machine disclose.

In Tables 1 and 2, the r of spectrograms one another in horizontal and vertical vibrations are displayed. To follow the sequential changes of spectrograms in Fig.s 1 to 6, two sequential sets are displayed in Fig. 7. One sequential set is $\{r_h(1,25), r_h(25,50), r_h(50,75), r_h(75,100), r_h(100,123)\}$ and the other is $\{r_v(1,25), r_v(25,50), r_v(50,75), r_v(75,100), r_v(100,123)\}$. Both $r_h(75,100)$ and $r_v(75,100)$ present the smallest values in the sets. Therefore, the dominant reasons of malfunction of the machine occur in the duration of the 75th to 100th minutes. Furthermore, $r_h(100,123)$ and $r_v(100,123)$ are very high. Most textbooks deliver the concept of "If r>0.8, the two data set are high correlation." In Fig. 7, $r_h(100,123)$ and $r_v(100,123)$ are larger than 0.8 because rotation of the roller entered a distinguished state of malfunction in the 100th minute. The inner damage is enhanced step by step.



	(1,25)	(25,50)	(50,75)	(75,100)	(100,123)
🟦 rh	0.4039	0.3879	0.3402	0.1633	0.8401
🟥 rv	0.5992	0.5873	0.51	0.3942	0.8007

Figure 7. One sequential set is {$r_h(1,25)$, $r_h(25,50)$, $r_h(50,75)$, $r_h(75,100)$, $r_h(100,123)$}and the other is {$r_v(1,25)$, $r_v(25,50)$, $r_v(50,75)$, $r_v(75,100)$, $r_v(100,123)$}.

V. Conclusion

The time-frequency analysis of Bearing1_1 of the XJTU-SY bearing datasets has been tried initially. The results present that the spectrogram-aided recognition is feasible. The interpretation of VEs and HEs is helpful to detect the fault diagnosis of the roller. The further studies will test more data, and add more algorithms for the accurate diagnosis.

Acknowledgments

The authors thank for the support from Guangdong University of Petrochemical Technology, Guangdong, China (Project number: 702/72100003142) and the funding from Maoming Green Chemical Industry Research Institute (Project number: MMGCIRI-2022YFJH-Y-042), Guangdong, China.

References

[1] National Laboratory of Railway Control and Safety, Beijing Jiaotong University, China (https://biaowang.tech/publication/)

[2] Lei, Y., Li, N., Guo, L., Li, N., Yan, T. and Lin, J., 2018. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. Mechanical systems and signal processing, 104, pp.799-834.

[3] Xu, Y., Li, Z., Wang, S., Li, W., Sarkodie-Gyan, T. and Feng, S., 2021. A hybrid deep-learning model for fault diagnosis of rolling bearings. Measurement, 169, p.108502.

[4] Yao, D., Li, B., Liu, H., Yang, J. and Jia, L., 2021. Remaining useful life prediction of roller bearings based on improved 1D-CNN and simple recurrent unit. Measurement, 175, p.109166.

[5] Zhao, B., Niu, Z., Liang, Q., Xin, Y., Qian, T., Tang, W. and Wu, Q., 2021. Signal-to-signal translation for fault diagnosis of bearings and gears with few fault samples. IEEE Transactions on Instrumentation and Measurement, 70, pp.1-10.

[6] Yang, J., Bai, Y., Wang, J. and Zhao, Y., 2019. Tri-axial vibration information fusion model and its application to gear fault diagnosis in variable working conditions. Measurement Science and Technology, 30(9), p.095009.

[7] Gao, T., Li, Y., Huang, X. and Wang, C., 2020. Data-driven method for predicting remaining useful life of bearing based on Bayesian theory. Sensors, 21(1), p.182.

[8] Ma, X., Si, Y., Yuan, Z., Qin, Y. and Wang, Y., 2020. Multistep dynamic slow feature analysis for industrial process monitoring. IEEE Transactions on Instrumentation and Measurement, 69(12), pp.9535-9548.

[9] Kaushal, B., Raveendran, S., Patil, M.D. and Birajdar, G.K., 2022. Spectrogram image textural descriptors for lung sound classification. Machine learning and deep learning in efficacy improvement of healthcare systems. CRC Press, pp.109-136.

[10] Jácome, C., Ravn, J., Holsbø, E., Aviles-Solis, J.C., Melbye, H. and Ailo Bongo, L., 2019. Convolutional neural network for breathing phase detection in lung sounds. Sensors, 19(8), p.1798.

[11] Lu, B.Y. and Wu, H.D., 2015. Auscultation using modern mobile communication. Acoustics Australia, 43, pp.303-309.

[12] Kingsbury, B.E., Morgan, N. and Greenberg, S., 1998. Robust speech recognition using the modulation spectrogram. Speech communication, 25(1-3), pp.117-132.

[13] Pinkowski, B., 1997. Principal component analysis of speech spectrogram images. Pattern recognition, 30(5), pp.777-787.

[14] Plante, F., Meyer, G. and Ainsworth, W.A., 1998. Improvement of speech spectrogram accuracy by the method of reassignment. IEEE Transactions on Speech and Audio Processing, 6(3), pp.282-287.

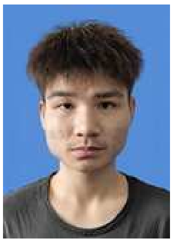
[15] Zhang, C., Qin, F., Zhao, W., Li, J. and Liu, T., 2023. Research on Rolling Bearing Fault Diagnosis Based on Digital Twin Data and Improved ConvNext. Sensors, 23(11), p.5334.

[16] Cheng, G., Lau, S., Tam, N., Wu, Z., Hu, A., Law, Y.N., Lai, E. and Ge, M., 2023, February. Unsupervised Remaining Useful Life Prediction for Bearings with Virtual Health Index. In 2023 13th International Conference on Power, Energy and Electrical Engineering (CPEEE) (pp. 126-131). IEEE.

[17] Haykin S., Van Veen B. (1998), *Signals and systems*, Wiley, New York, pp. 72–87.

Guanhua Zhu received his BS in Mechanical manufacturing and automation major from Guangdong Ocean University in 2003, MS in software engineering from Huazhong University of Science and Technology in 2006. He is currently a senior engineer with Guangdong Provincial Key Lab.of Petrochemical Equipment and Fault Diagnosis, Guangdong University of Petrochemical Technology, Maoming City, Guangdong, China. His academic interests focus on fault diagnosis, industrial internet and artificial intelligence.

Xiaoling Xu received the BS and MS degrees from Henan Polytechnic University, Henan, China, in 2006 and 2008, respectively，and received the PhD degree in in control science and engineering from Guangdong University of Technology. She is currently a lecturer with Faculty of Automation, GDUPT, Maoming, Guangdong, China. Her current research interests include distributed optimization, multiagent coverage control.



Qing Zhong is now student at Faculty of Automation, Guangdong University of Petrochemical Technology, Guangdong, China. He won the third prize of the 5th National College Students Embedded Chip and System Design Competition - Application Track South Division Rematch and the third prize of 8th National College Students Biomedical Engineering Innovation Design Competition.He Published a paper titled "Research on Data Structure and Algorithm of YOLO V8"

Bing-Yuh Lu received his BS in electrical engineering from National Central University in 1988, MS in electrical engineering from National Taiwan University in 1993, and PhD in electrical engineering from National Taiwan University in 2000. He is currently a professor with Faculty of Automation, Guangdong University of Petrochemical Technology (GDUPT), Maoming City, Guangdong, China. He has been an instructor (1993 to 2000), an associate professor (2000 to 2016), and a full professor (2016 to 2019) with the Department of Electronic Engineering, Tungnan University, New Taipei City, Taiwan, from 1993 to 2019. He is a member of IEEE, has been a member of the Technical Committee IEEE International Conference on Advanced Communication since 2015, and served as a reviewer for some international journals. His academic interests include electronic circuits and systems, medical engineering, acoustics, modeling, and signal measurement and processing.

Yushen Lu is a student with Faculty: of Automation, Guangdong University of Petrochemical Technology, Guangdong, China. He is proficient in C language. He is familiar with Arduino, 51, STM32 and other microcontrollers and can independently complete PCB hardware design according to project requirements. His research interests include Internet of things and MCU development.

Guangming Xu is a student with Faculty of Automation, Guangdong University of Petrochemical Technology, Guangdong, China.At present, he is participat in the solar photovoltaic project in the measurement and control system development laboratory of our school.

Kai Sun is an automation student at Guangdong University of Petrochemical Technology. Participate in Maoming Green Chemical Research Institute "sail Plan" 2022 "application innovation" project. His research interests include electronic circuit hardware design and embedded systems.

Yumeng Zhou is a student of Measurement and Control Technology and Instrumentation at the School of Automation, Guangdong University of Petroleum and Chemical Technology. She has won the Internet+School level Bronze Award, and she work diligently, study hard, solve professional technical problems, and coordinate work.

Ziyi Jiang is a student with Faculty: of Automation, Guangdong University of Petrochemical Technology, Guangdong, China.She majored in measurement and control Technology and instrumentation, and won the third-class scholarship for two consecutive years.Have a wide range of knowledge and skills. She is good at solving complex problems and working well with others. She has good communication skills and team spirit, and she can quickly adapt to the new environment and finish the work efficiently. she likes to challenge herself and keep learning and growing.

.

Minhao Wang is a student with Faculty:Electrical Engineering and Automation,Guangdong University of Petrochemical Technology, Guangdong, China.He won The National third prize of The 8th National College Student Biomedical Engineering Innovation Design Competition and The university-level plan project for college students' innovation and entrepreneurship was approved of Guangdong University of Petrochemical Technology.He is serious and motivated in his studies.He won the school's first-class scholarship and the school's top three students in 2023.His research interests include coding and PCB hardware design.

Evaluation System for Dancing Enlightenment Posture Training Using the Skeleton Tracking of Microsoft Common Objects in Context

Ruilong Huang*, Huifang Deng*, Ruei-Yuan Wang**, Bing-Yuh Lu*,
Hongwei Ren*, Yiheng Chen*, Jianwen Ye*, Jinhui Chen*, Yingbo Jia*, Leyang Lang*

*School of Automation, Guangdong University of Petrochemical Technology, Maoming City, Guangdong, China
**School of Science, Guangdong University of Petrochemical Technology, Maoming City, Guangdong, China

FranklinLu888@outlook.com

Abstract— **Dance enlightenment education is of great significance to children's physical health. Therefore, we developed an evaluation system to correct the posture of the beginners for their dancing training. The pre-train network is based on the Open Neural Network Exchange. The pattern of human pose is the skeleton in the Microsoft Common Objects in Context dataset. We designed a quantitative presentation to calculate the similarity of the postures between the skeletons of the dancing beginner and target image of the training, and obtained objective evaluation indicators based on the recorded of the angle differences of limbs to calculate the score. 8 angles of the joints have been computed and presented in the evaluation system. The results show that the dancing beginner can correct her postures to approach the target image of the training. She improved the score from 94 to 96. Now, parents pay more and more attention to the quality education of their children. The AI-aided dancing training will make the beginners to learn the performances of the postures much easier in any time and any location. Therefore, the learning of dancing will become more interesting for the beginners.**

Keywords—**Dance, Deep Learning, Evaluation System, MatLab, Microsoft Common Objects in Context, Posture, Skeleton, Training.**

I. Introduction

Dance enlightenment education is of great significance to children's physical and mental health. It can not only regulate children's behavior but also improve children's aesthetics. Dance enlightenment education requires scientific and effective learning methods. In this diversified society, movement practice, form, dance posture and other training in dance enlightenment can avoid the bad forms of children's bodies during growth and development, forge children's good appearance, and lay a good foundation for children's healthy development. The application of human posture recognition technology in computer vision and deep learning in dance enlightenment teaching can help teachers better understand the learner's movement, correct wrong movement postures in time, and improve teaching.

Wu et al. [1] has proposed multi-stage pose network (MSPN) and optimized in three aspects of the stacked hourglass network, i.e. network structure, feature information transfer and heat map generation. Li et al. [2] deducted the downsampling and upsampling on the Resnet structure, and featured out that the different stages are connected to enhance the ability of feature expression. Scholars generated different heatmaps at their respective stages to gradually improve the estimation accuracy of human pose. As the accuracy of human posture estimation networks continues to improve, the complexity of the network structure also gradually increases, which makes it difficult to compare and analyze between networks. Xiao et al. [3] proposed the simple baseline network (SBN) in 2018. It mainly proposed a relatively simplified baseline for human posture estimation and posture tracking. This paper mentioned believes that the current popular human posture estimation methods are too complex. The various models proposed on the Internet seem to be quite different in structure, but they are very similar in performance. Compared with the complex network structures of traditional networks such as Hourglass [4], the basic SBN structure appears intuitive and simple. This network structure is constructed by adding several layers of denaturation after Resnet to directly generate heatmaps. Compared to other models, this is done by replacing the oversampling structure through deconvolution.

Recently, there were many applications of the posture correction using the AI technology, especially, extracting the human skeleton to evaluate the posture. Carey et al. [9] compared the effectiveness of two different skeletal pose models for a near real-time, multi-stage classifier and find no significant difference between the 2 pose models. Therefore, we select Microsoft Common Objects in Context (MS COCO) dataset for this study. Liu et al. [10] proposed a mechanism for estimating and correcting fitness posture based on deep learning. The 14 keypoints of the human body can be obtained after correction. Jangade et al. [11] pointed out that Human Pose Estimation (HPE) will be a wide range of applications and enter human daily living step by step. In fact, many papers mentioned the potential applications of HPE [12-15].

However, little researches mentioned about the pose correction of training dancing. Intuitively, dancing is a kind of art. Therefore, little scientists pay attention to the dancing. The number of the studies related to AI-aided correction of dancing is less. Consequently, this study tried to develop a system to implement AI-aided correction of dancing. The operations of the evaluation system in this study are as follows:(1) Construct a deep learning human posture detection algorithm network to realize human posture recognition, extract and compare the coordinates of human body key points among learners and standards; (2) Based on the coordinate information of the key points of the human body of the standard person, the system determine the correlation and angle difference with the key points of the learner's human body, and finally obtain the similarity of dance movements between the learner and the standard person; and (3) Build a dance enlightenment evaluation system GUI platform on MATLAB software to achieve the purpose of assisting children's teaching and enlightenment education.

II. Method

A. Estimate Body Pose Using Deep Learning Using MatLab

This study follows an example of MatLabTM (The MathWorks Inc., USA) [5]. Therefore, the pre-train network is based on the Open Neural Network Exchange (ONNX) which is an open ecosystem for interoperable AI models [6] at first. The pattern of human pose is the MS COCO dataset which is a large-scale object detection, segmentation, key-point detection, and captioning dataset [7]. Then, the tested image was predicted the heatmaps and part affinity fields (PAFs) , which are output from the 2-D output convolutional layers. The post-processing part of the algorithm identifies the individual poses of the people in the image using the heatmaps and PAFs returned by the neural network.

B. MS COCO human skeleton

The MS COCO data set was proposed in 2014 and is the mainstream in the field of human posture recognition. It focuses on solving large datasets such as object detection, key point detection, object segmentation and subtitle generation in natural environments by using computer vision technology. In 2016, we added a new task to the MS COCO dataset using 2D human skeleton key point detection. The data set contains human samples labeled with coordinates of key points on the human body. These samples are labeled with 17 key points on the human skeleton. Below we combine pictures and specific tables to learn more. Figure 1 shows the 17 key points of the MS COCO human skeleton.

C. Indicators of human posture estimation

For the COCO data set, the officially designated joint point similarity measurement method is OKS object keypoint similarity (OKS), which is used to calculate the similarity between the true value and the predicted key point. This data set is inspired by the target detection metric. Different from other evaluation indicators, it uses the average accuracy mean based on the OKS as a more scientific evaluation indicator. The value of OKS is between 0 and 1. The closer it is to 1, the closer the predicted human joint points are to the real values of the data and annotations, and the better the prediction effect.

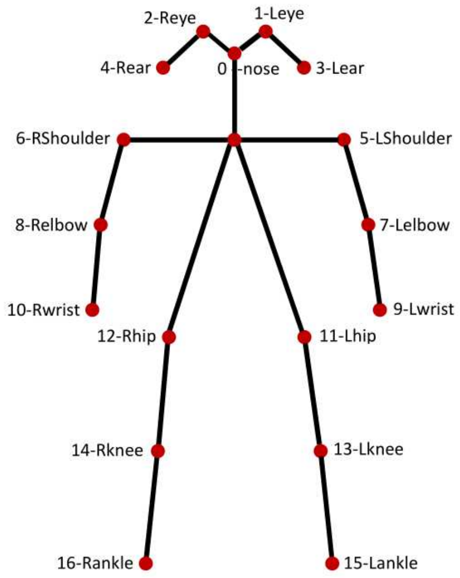


Figure 1. The 17 key points of the MS COCO human skeleton [8].

The calculation of OKS is based on a non-standardized Gaussian distribution.

Mathematically, the keypoint similarity (KS) for keypoint i is written as:

$$KS_i = \exp\left(-\frac{1}{2}\left(\frac{d_{pi}}{S_p\sigma_i}\right)^2\right) \qquad (1)$$

where the Euclidean distance between the ground truth and predicted keypoint i is denoted by di, the constant for keypoint i is denoted by k; the scale of the ground truth object is denoted by s; thus s2 becomes the object's segmented area. For each keypoint, the KS lies between 0 and 1.

The OKS is expressed by

$$OKS_i = \frac{\sum_i KS_i \cdot \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \qquad (2)$$

where the keypoint similarity for keypoint i is denoted by OKSi; vi is the ground truth visibility flag for keypoint i. δ (vi > 0) is the Dirac-delta function which computes as 1 if the keypoint i is labeled, otherwise 0.

D. Dance Enlightenment Evaluation System Based on Posture Estimation

A new method for the dance enlightenment evaluation system is introducing in this section. This method follows the steps:

- Use the pre-trained COCO model to extract human skeleton information from standard dance pose parameters and student dance pose parameters.
- Obtain position information of human body key points and calculate based on the obtained coordinate information of human body key points.
- Compute the angle differences between each limb, and infer the angle difference between the limb angle of the image of student's dance pose and that of standard dance pose.
- Design quantitative presentation to calculate pose similarity, and obtain objective evaluation indicators based on the recorded of the angle differences of limbs.
- Select the largest error to advice the corrections of the student's poses.

In summary, the computing processes of the dance enlightenment evaluation system based on posture estimation are presented in Fig. 2.

E. Equipment

1. Hardware

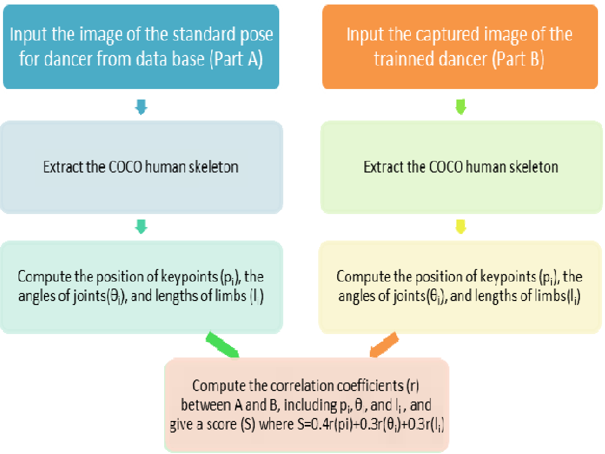
The experiment is performed by a laptop computer with Intel® CoreTM i5-8250U CPU @ 1.60GHz, 4 cores, 8 logical processors and 8GB memory. The operating system is Windows 10.

2. Software

The program is run in MatLab 2020a integral developing environment. The MatLab is a high-level computer programming language and interactive environment for developing algorithms, data analysis, and visualization as well as provides a large number of toolboxes and functions to help users to efficiently handle various complex mathematical operations and data analysis.

III. Results

The initial page of the evaluation system is presented in Fig. 3. There are 3 frames in the figure. The left frame shows the picture of the target pose of this train which can be selected by the items of the list under the picture. The middle frame is null in the figure. The frame will be filled with a picture of the learner when the train begins. The rightmost frame is designed for showing the differences of the joint angles between the target and learner's MS COCO skeletons which is denoted by Δθi.



Input the image of the standard pose for dancer from data base (Part A)
Input the captured image of the trainned dancer (Part B)
Extract the COCO human skeleton
Extract the COCO human skeleton
Compute the position of keypoints (p_i), the angles of joints(θ_i), and lengths of limbs (l_i)
Compute the position of keypoints (p_i), the angles of joints(θ_i), and lengths of limbs(l_i)
Compute the correlation coefficients (r) between A and B, including p_i, θ, and l_i, and give a score (S) where S=0.4r(pi)+0.3r(θ_i)+0.3r(l_i)

Figure 2. The computing processes of the dance enlightenment evaluation system based on posture estimation

In Fig. 4, the null space in Fig. 3 has been filled. The image of the learner and the corresponding skeleton is uploaded and displayed. Furthermore, the Δθi are measured and shown in the blocks in the rightmost frame. They are 17, 12, 2, 6, 9, 0, 5, and 6 degrees, respectively. The approaching time is 0.15 sec. and the score of the training is 94 points.

In Fig. 5, the learner corrects her postures to improve the dancing based on the Δθi of the Fig. 4. Therefore, the Δθi improve to 5, 3, 3, 0, 9, 0, 5, and 6 degrees. Concurrently, the score goes ahead to 96.

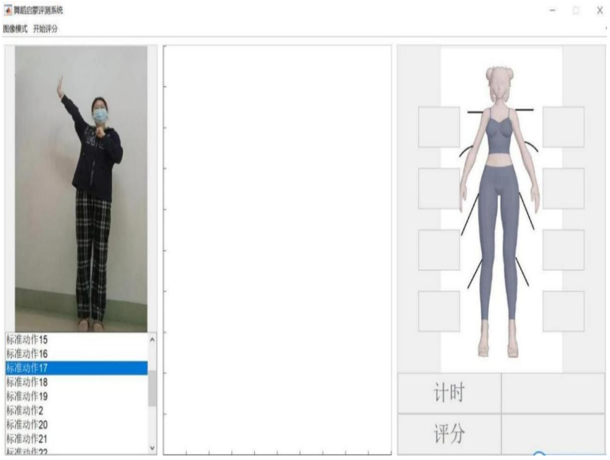
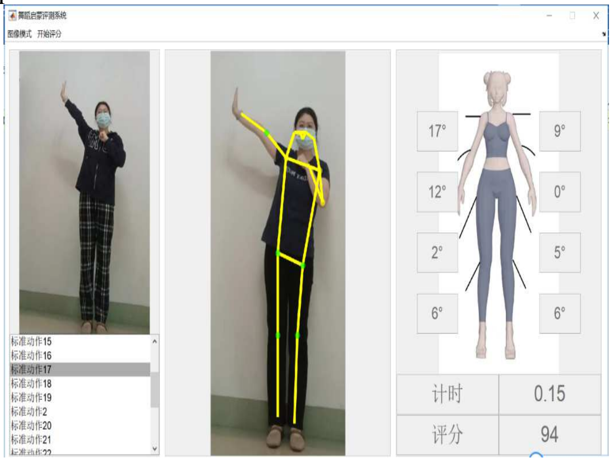


Figure 3. The initial page of the evaluation system

IV. Discussion

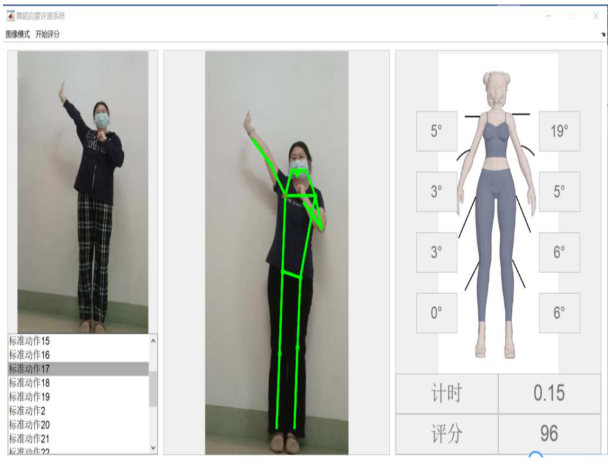
In Fig. 4, the difference angle of the right elbow (8-Relbow) between target and learner's images is 17 degree. Therefore, the learner raises her arms slightly and stands slightly straighter. Then, the corrected learner's dance movements and postures are photographed and uploaded to the dance enlightenment evaluation system. Figure 5 shows the corrected human posture model. From the picture, we can know that after raising the right arm, the left arm also needs to be further adjusted, such as relaxing the arm slightly and raising the left arm slightly, so as to meet the pose of the target picture as much as possible, so that the similarity will also be increased. Therefore, the posture and score have been improved.



计时 0.15

评分 94

Figure 4. The training phase of the evaluation system



计时 0.15

评分 96

Figure 5. The correcting phase of the evaluation system

The computation of the dance posture similarity is an open question. However, we proposed a method to compute the score. Furthermore, the evaluation of the pose focuses on the limbs, because we think the poses of limbs are the main postures to influence the touch of the dancers' expressions. However, the evaluation of total skeletons might be tested in the near future.

If the color of the learner's clothes is similar to their skin color, the background is cluttered or the learner wear the loose-fitting clothes, the recognitions of the human bones and joints become less accurate. Therefore, it is necessary to improve the training data systematically and improve the human posture estimation algorithm.

In this test, the learner danced with the face mask. Therefore, some errors of the keypoints positions on the face are obvious. However, the target image of the training is in the same condition, i.e. the dancer also posed with the face mask. Therefore, the "noise", the face mask, has been cancelled. The computations of Δθi keep accurate. In fact, the computation of the keypoints on the body and limbs are independent to those on the face. Therefore, this study verified the computing accuracy when the face mask is necessary to be with the learner

V. Conclusion

Now, parents pay more and more attention to the quality education of their children. Now, parents pay more and more attention to the quality education of their children. The AI-aided dancing training will make the beginners to learn the performances of the postures much easier in any time and any location. Therefore, the learning of dancing will become more interesting for the beginners.

Acknowledgments

The authors thank for the support from the Guangdong University of Petrochemical Technology, Guangdong, China (Project number: 702/72100003142), the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515010168, Grant 2019A1515010830, and Maoming Green Chemical Industry Research Institute (Project number: MMGCIRI-2022YFJH-Y-042), Guangdong, China.

References

[1] Wu, H., Lu, X., Lei, B. and Wen, Z., 2021. Automated left ventricular segmentation from cardiac magnetic resonance images via adversarial learning with multi-stage pose estimation network and co-discriminator. Medical Image Analysis, 68, p.101891.

[2] Li, B. and He, Y., 2018. An improved ResNet based on the adjustable shortcut connections. Ieee Access, 6, pp.18967-18974.

[3] Xiao, B., Wu, H. and Wei, Y., 2018. Simple baselines for human pose estimation and tracking. In Proceedings of the European conference on computer vision (ECCV) (pp. 466-481).

[4] Newell, A., Yang, K. and Deng, J., 2016. Stacked hourglass networks for human pose estimation. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14 (pp. 483-499). Springer International Publishing.

[5] Estimate Body Pose Using Deep Learning Using MatLab (https://www.mathworks.com/help/deeplearning/ug/estimate-body-pose-using-deep-learning.html#EstimateBodyPoseUsingDeepLearningExample-3)

[6] Open Neural Network Exchange (https://onnx.ai/)

[7] Microsoft Common Objects in Context (https://paperswithcode.com/dataset/coco)

[8] The 17 key points of the MS COCO human skeleton. (https://www.stubbornhuang.com/525/)

[9] Carey, K., Abruzzo, B., Lowrance, C., Sturzinger, E., Arnold, R. and Korpela, C., 2020, April. Comparison of skeleton models and classification accuracy for posture-based threat assessment using deep-learning. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II (Vol. 11413, pp. 671-678). SPIE.

[10] Liu, X., Xiao, H. and Cheng, J., 2021, July. Human posture estimation and correction based on the CPM and the Pearson correlation coefficient. In International Conference on Sensors and Instruments (ICSI 2021) (Vol. 11887, pp. 385-390). SPIE.

[11] Jangade, J. and Babulal, K.S., 2023, March. Study on Deep Learning Models for Human Pose Estimation and its Real Time Application. In 2023 6th International Conference on Information Systems and Computer Networks (ISCON) (pp. 1-6). IEEE.

[12] Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B., 2014. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (pp. 3686-3693).

[13] Toshev, A. and Szegedy, C., 2014. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1653-1660).

[14] Zhang, F., Zhu, X. and Ye, M., 2019. Fast human pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3517-3526).

[15] Munea, T.L., Jembre, Y.Z., Weldegebriel, H.T., Chen, L., Huang, C. and Yang, C., 2020. The progress of human pose estimation: A survey and taxonomy of models applied in 2D human pose estimation. IEEE Access, 8, pp.133330-133348.

Ruilong Huang received his B.Eng in industry automation from Northeast Petroleum University in 2002, M.Eng in control theory and control engineering from South China University of Technology in 2005. He is currently a lecturer with Faculty of Automation, GDUPT Maoming City, Guangdong, China. His academic interests focus on intelligent detection and control technology.

Huifang Deng received her B.S. from the Faculty of Automation, Guangdong University of Petrochemical Technology (GDUPT), Guangdong, China in June, 2023. She got the scholarships of excellent learning at GDUPT many times. She is currently a junior engineer at Chinese Huadian Corporation, Guangdong, China. Her research interests include AI and image processing technologies.

Ruei-Yuan Wang received his PhD in Science from the Institute of Geosciences at the Chinese Culture University in Taiwan in 2010; he has been served as a postdoctoral researcher at the Spatial Information Research Center (SIRC) of Taiwan University (NTU) and the Center for Space and Remote Sensing Research (CSRSR) of Central University (NCU). He is currently serving as an associate professor in the Department of Geographic Sciences at the Faculty of Science, Guangdong University of Petroleum and Chemical Technology (GDUPT), Maoming City, Guangdong, China. He primarily engaged in teaching courses in geographic science, remote sensing (RS), and geographic information systems (GIS). His academic interests focus on geographic science related fields, environmental ecology, remote sensing carbon monitoring, tourism geography, etc. His application technology mainly focuses on decision support (DSS), knowledge management (KM), artificial neural networks (ANN), big data, remote sensing technology, and geographic information system technology (GIS). He previously served as expert peer reviewers of various journals for Sustainability; Land; Forests; Expert Systems with Application; Business Management and Economics, Journal of Coastal Research (JCR), etc.

Bing-Yuh Lu received his BS in electrical engineering from National Central University in 1988, MS in electrical engineering from National Taiwan University in 1993, and PhD in electrical engineering from National Taiwan University in 2000. He is currently a professor with Faculty of Automation, Guangdong University of Petrochemical Technology (GDUPT), Maoming City, Guangdong, China. He has been an instructor (1993 to 2000), an associate professor (2000 to 2016), and a full professor (2016 to 2019) with the Department of Electronic Engineering, Tungnan University, New Taipei City, Taiwan, from 1993 to 2019. He is a member of IEEE, has been a member of the Technical Committee IEEE International Conference on Advanced Communication since 2015, and served as a reviewer for some international journals. His academic interests include electronic circuits and systems, medical engineering, acoustics, modeling, and signal measurement and processing.

Hongwei Ren received her B.S. degree in IndustryAutomation from Northeast Electric Power Univer-sity, Jilin, China, in 1998, her M.S. degree andPh.D. degree in Systems Engineering from SouthChina University of Technology, Guangzhou, Chinain 2002 and 2017,respectively.She is currently an Associate Professor in theSchool of Automation, Guangdong University ofPetrochemical Technology, Maoming, China. Herresearch interests include synchronization of com-plex networks, consensus of multi-agent systems andstochastic dynamic system analysis and control.

Yiheng Chen is a student with Faculty: of Automation, Guangdong University of Petrochemical Technology, Guangdong, China.He won the third prize of 2022 "Cossberg Cup" Engineering College Students' Comprehensive Experimental Skills Competition of Guangdong University of Petrochemical Technology, and participated in the 2022 "Application Innovation" project of "Sail Plan" of Maoming Green Chemical Research Institute. His research interests include PCB hardware design and fuzzy control.

Jianwen Ye,is a student majoring in Electrical Engineering and Automation at Guangdong University of Petrochemical Technology, has won university-level scholarship and the second prize of Guangdong Provincial Engineering College Students' Comprehensive Experimental Skills Competition.

Jinhui Chen is a student majoring in Electrical Engineering and Automation at Guangdong University of Petrochemical Technology, has won university-level scholarship and the second prize of university-level innovation and Entrepreneurship Competition.

Yingbo Jia is a student majoring in measurement and control technology and instrumentation at Guangdong University of Petrochemical Technology, has won first-class scholarships for many times

Leyang Lang is a student with Faculty of Automation, Guangdong University of Petrochemical Technology, Guangdong，China. He has won the third prize of the Guangdong Provincial Research Cup. His main research interests are geographic remote sensing monitoring, machine learning, and deep learning.

Computer Vision-Based Structural Deformation Monitoring System on Android Smartphones: Design and Implementation

Xiang DONG*, Maokai LAI*, Hui LIANG**, ***, Peng WU*, Chaoxia WANG****, Ting PENG*

* Chang'an University, Xi'an 710064

** Shaanxi Huashan Road and Bridge Group Co., LTD. Xi'an 710016

*** Shaanxi Zhengcheng Road and Bridge Engineering Research Institute Co., LTD. Xi'an 712000

**** Xi’an University of Posts and Telecommunications, Xi'an 710061

1290740787@qq.com, 3412303785@qq.com, 5936001270@qq.com,1169886830@qq.com, 2102332664@qq.com, t.peng@ieee.org

Abstract— **Computer vision displacement monitoring techniques offer a promising alternative to traditional displacement sensors, but most current approaches or systems use high-cost cameras and require limited measurement sites. To facilitate widespread implementation in real-world engineering, simpler systems or approaches are needed. This paper introduces a computer vision-based structural deformation monitoring system, written and developed using OpenCV and Kotlin. The system uses an Android smartphone camera and telescope as the acquisition device and adopts template matching technology based on digital image correlation to realize target displacement monitoring. A real-time display of the offset curve is realized with MPAndroidChart, and socket TCP communication is used to transmit the monitoring data. The detailed operation process of the system is shown as an example of dynamic detection. The test results show that the system runs smoothly, is easy to operate, and has the following advantages: 1) Low cost of hardware composition and simple construction process; 2) High monitoring accuracy (accurate to 0.01 mm); 3) High monitoring frequency (up to 10 frames per second in dynamic detection mode); 4) Visualization of monitoring data, with the monitoring structural offset presented in real time as a curve for easy user comprehension. Overall, this system can provide technical and auxiliary decision support for structural deformation monitoring.**

Keywords— **Computer Vision, Kotlin, OpenCV, Structural Deformation Monitoring, Socket TCP**

I. INTRODUCTION

Structural deformation monitoring plays a very important role in ensuring the long-term safety of transportation infrastructure such as highways, bridges, and tunnels, as well as in their early warning and assessment. The commonly used displacement measurement methods include contact and non-contact measurements. The literature systematically discusses the advantages and disadvantages of the two displacement measurement methods [1-3] . Among them, contact displacement measurement methods usually use accelerometers, linear variable differential pressure transducers (LVDT), and tie-wire displacement sensors. Accelerometers can directly measure the acceleration response of a structure, but the measurement effect is insufficient in the low-frequency range. Linear variable differential pressure transducers (LVDT) have higher measurement accuracy, but they are not easy to install in practical engineering because they usually need to be mounted between the target point and a fixed reference point. Non-contact displacement measurement methods usually include the Global Positioning System (GPS), Total Station, Laser Doppler Vibrometer (LDV), Computer Vision Displacement Measurement, etc. GPS sensors are relatively easy to install, but their measurement accuracy is limited. Total stations have high accuracy, but the measurement speed is low, and the collection of data is time-consuming, which limits the overall efficiency. The Laser Doppler Vibrometer (LDV) can realize high-resolution, high-accuracy remote displacement measurements, but it needs to be installed close to the object and is limited to a few measurement points and expensive equipment.

In recent years, computer vision-based methods have received extensive research attention due to their advantages of non-contact, long-distance, high accuracy, and multi-point monitoring, and have been applied in civil engineering structural health monitoring, including bridge deflection testing and damage identification. However, most of the current monitoring programs or systems use high-cost cameras, and the measurement sites used to install hardware devices are often limited and affected by external environmental factors, which restricts the widespread application of this technology. To this end, many researchers have made contributions in reducing deployment costs, optimizing target tracking algorithms, and reducing system complexity. Chen [1] proposed an improved visual method for robust multi-point dynamic displacement monitoring of smartphones in interference environments and conducted a series of sinusoidal swept vibration experiments considering different interference factors to investigate the performance of the method. Bai [4] proposed a non-contact visual sensing system for monitoring structural displacements using the advanced Zernike sub-pixel edge detection technique, which provides the intrinsic frequency and vibration pattern of the target instantaneously and can be used to accurately localize damage. Khan [5] proposed a new low-cost, non-contact displacement measurement system called the high-tech computer-Vive tracking system (HTC-VTS) and achieved favourable results in laboratory-scale free vibration tests and shaker tests on cantilever girders, as well as actual bridge displacement measurements.

To facilitate widespread implementation in real-world engineering, this paper proposes a computer vision-based structural deformation monitoring system on Android smartphones. The computer vision-based structural deformation monitoring system's overall architecture is initially introduced in this study, after which each module's design parameters and implementation tactics are discussed. By using dynamic detection as an example, the basic processes of the computer vision-based structural deformation monitoring system are shown. To a certain extent, the system can offer technical and auxiliary decision support for structural deformation monitoring.

II. Design and Implementation of computer vision-based structural deformation monitoring system

A. *Overall System Design*

This paper is based on the Android Studio Arctic Fox |2020.3.1 Patch 2 development environment, which uses OpenCV and Kotlin to set up a structural deformation monitoring program and develop a computer vision-based structural deformation monitoring system. The system uses an Android smartphone camera plus a telescope as the acquisition device and adopts a template matching technique based on digital image correlation to realize target displacement monitoring. The overall architecture of the system is shown in **Figure 1**, which mainly includes four modules: image acquisition, image processing, data visualization, and data transmission. Among them, the image processing module contains two modes: static detection and dynamic detection, which can track the deformation of the structure in real time according to the change in relative offset of multiple markers on the structure and the change in absolute offset of the overall markers and provide an early warning so as to take timely measures to deal with the potentially dangerous structure.

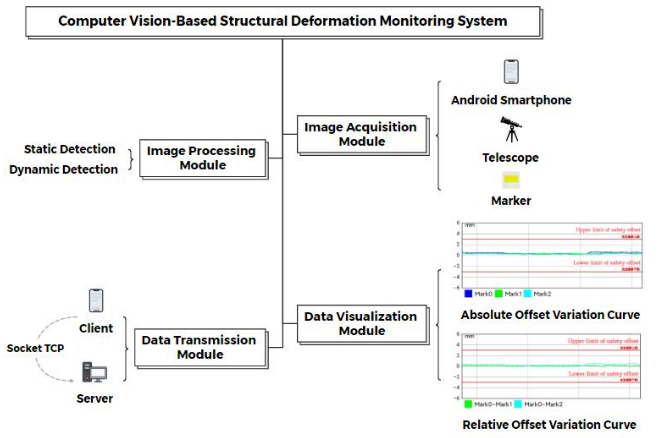


Figure 1. Computer vision-based structural deformation monitoring system architecture

1) ***Image Acquisition Module:*** The image acquisition module includes an Android smartphone, a telescope, and markers for the purpose of image acquisition of markers on the structure for template creation and to facilitate subsequent image processing of the markers.

Android smartphones equipped with high-resolution cameras offer the advantages of low cost, portability, and ease of operation, while being able to perform optical sensing under varying conditions [1]. In order to ensure the smooth operation of the system and the high clarity of the captured images, the smartphone used in this paper requires Android 5.0 or higher version.

This paper uses the SAGA Maksutov-Cassegrain II telescope and the accompanying metal mobile phone clip. The telescope's parameters are as shown in **Table 1**. The advantages of the telescope are mainly reflected in three aspects: first is the 70mm large caliber and brighter brightness; second is the combination of the advantages of reflection and refraction; the short-lens body achieves a long focal length, improved resolution, and an effective magnification multiplier; and third is the very short mirror body, which achieves portability and stability.

TABLE 1. Telescope Parameters

No.	Telescope Parameters	Value
1	Size	250×83×135mm
2	Eyepiece Size	20mm
3	Exit Pupil Diameter	2.8~0.93mm
4	Distance of Exit Pupil	15~13mm
5	Magnification Rate	25~75 times
6	Focal Length	780mm
7	Objective Lens Caliber	70mm
8	Near Focal Distance	4.5m

Markers include both artificial markers and natural markers. Among them, artificial markers are markers added to the measured structure to increase the differentiation between the tracking target and the surrounding environment and to improve the tracking accuracy, while natural markers utilize the image features and texture structure on the surface of the structure as markers [3]. As the accuracy of artificial markers is higher than that of natural markers, this study prioritizes the use of artificial markers to obtain stable measurement results when the installation conditions allow.

The acquisition device consisting of an Android smartphone and a telescope is shown in **Figure 2**.

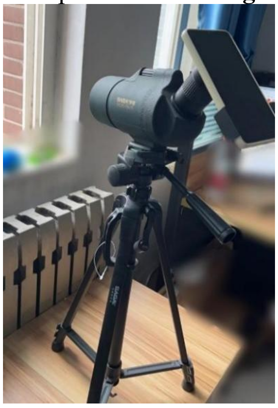


Figure 2. Acquisition device

2) Image Processing Module: The image processing module includes two modes, static detection and dynamic detection, both of which are realized by switching the monitoring mode through switch code. The image processing is achieved by cropping the acquired image, template creation, conversion of the actual length of the template to the pixel length, target tracking, displacement calculation, and providing a data source for the data visualization module.

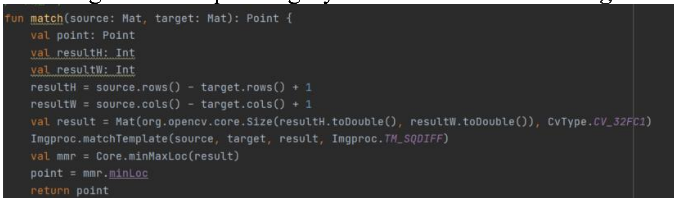
Currently the commonly used target tracking method in computer vision structure displacement measurements is based on template matching with digital image correlation. The template matching methods in OpenCV include six [3, 6], as listed in **Table 2**.

TABLE 2. MATCHTEMPLATE () FUNCTION TEMPLATE MATCHING METHOD TO SELECT THE MARK PARAMETERS

No.	Mark Parameters	Function
1	TM_SQDIFF	Sum of Squared Difference
2	TM_SQDIFF_NORMED	Normalized Sum of Squared Difference
3	TM_CORR	Cross Correlation
4	TM_CCORR_NORMED	Normalized Cross Correlation
5	TM_CCOEFF	Zero-mean Cross Correlation
6	TM_CCOEFF_NORMED	Zero-mean Normalized Cross Correlation

Of those indicators listed in **Table 2**, TM_SQDIFF uses the sum of squared differences to match, and TM_SQDIFF_NORMED normalizes TM_SQDIFF so that the result is scaled to a value between 0 and 1. Both methods compute a value of 0 when the template and the sliding window are perfectly matched, and the lower the match between the two, the greater the computed value. TM_CORR uses a multiplication operation between the template and the image; the larger the matching value, the better the matching effect, and 0 means the worst matching result. TM_CCORR_NORMED normalizes the TM_CORR so that the result is scaled to the range between 0 and 1. When the template matches perfectly with the sliding window, the calculated value is 1, and when the two don't match at all, the calculated result is 0. TM_CCOEFF uses TM_CORR to match the result of the template subtracted from the mean and the result of the original image subtracted from the mean, and this method can be a better solution to the effect between the template image and the original image due to the difference in brightness. In this method, the higher the match between the template and the sliding window, the larger the calculated value is; the lower the match, the smaller the calculated value is; and the calculated result of this method can be negative. TM_CCOEFF_NORMED normalizes the TM_CCOEFF so that the result is scaled down to a value between 1 and -1. When the template matches perfectly with the sliding window, the calculated value is 1, and when the two are not matched at all, the calculated result is -1.

Due to the fact that the sum of squared differences calculates the Euclidean distance between the template image and the overlapping image as a similarity measure, which is the most intuitive indicator and simple to calculate [6] , Therefore, the computer vision-based structural deformation monitoring system design implements target tracking by calling the TM_SQDIFF method in OpenCV for template matching. Its corresponding Python code is shown in **Figure 3**.



```
fun match(source: Mat, target: Mat): Point {
    val point: Point
    val resultH: Int
    val resultW: Int
    resultH = source.rows() - target.rows() + 1
    resultW = source.cols() - target.cols() + 1
    val result = Mat(org.opencv.core.Size(resultH.toDouble(), resultW.toDouble()), CvType.CV_32FC1)
    Imgproc.matchTemplate(source, target, result, Imgproc.TM_SQDIFF)
    val mmr = Core.minMaxLoc(result)
    point = mmr.minLoc
    return point
```

Figure 3. Template matching

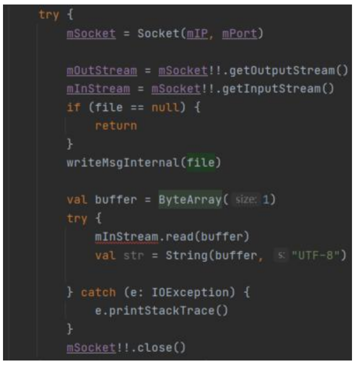
The matching algorithm calculation formula is as follows when using TM_SQDIFF from OpenCV:

$$R(x,y) = \sum_{x',y'} \left[T(x',y') - I(x+x',y+y')\right]^2 \tag{1}$$

In the formula, T refers to a template image, and I stands for an image to match. x and y represent the coordinates of the element in the upper left corner of the current search box in the I matrix. x', y' represent the element coordinates of the matrix I of the T and the search box out.

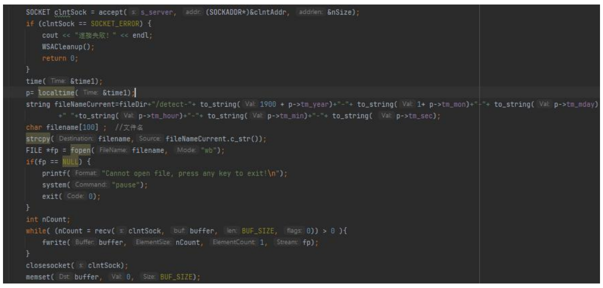
3) ***Data Visualization Module:*** Based on MPAndroidChart, the visual monitoring system of structural deformation is visualized. The module can record the absolute and relative offsets of each marker in real time and display them in the form of curves, and the visualization function of the system can also be used as an auxiliary means of manual monitoring to ensure the reliability of the data.

4) Data Transmission Module: In order to facilitate further calibration and verification of the data, ExcelUtil is called during the development of the system to realize the function of exporting the monitoring data in the form of an Excel table. When exporting data, you can choose to export the data to smartphone or PC. If the monitoring data is exported to the smartphone, the system directly stores the monitored data in the path "android\data\com.android.detect" in the file manager by default. If the monitoring data is exported to a PC, it is necessary to transfer the monitoring data through socket communication based on the Transmission Control Protocol (TCP). Zhang [7] mentioned stream sockets use the TCP protocol, and due to their internal flow control settings, they can transmit large amounts of data streams when used. At the same time, it provides a reliable service to ensure that no packet errors occur during transmission. Therefore, when designing the data transmission module, this method is used to achieve data transmission, and the implementation code is shown in **Figure 4** and **Figure 5**.



```
try {
    mSocket = Socket(mIP, mPort)

    mOutStream = mSocket!!.getOutputStream()
    mInStream = mSocket!!.getInputStream()
    if (file == null) {
        return
    }
    writeMsgInternal(file)

    val buffer = ByteArray( size: 1)
    try {
        mInStream.read(buffer)
        val str = String(buffer,  s: "UTF-8")

    } catch (e: IOException) {
        e.printStackTrace()
    }
    mSocket!!.close()
```

Figure 4. Client implementation



```
SOCKET clntSock = accept( s: s_server,  addr: (SOCKADDR*)&clntAddr,  addrlen: &nSize);
if (clntSock == SOCKET_ERROR) {
    cout << "连接失败! " << endl;
    WSACleanup();
    return 0;
}
time( Time: &time1);
p= localtime( Time: &time1);
string fileNameCurrent=fileDir+"/detect-"+ to_string( Val: 1900 + p->tm_year)+"-"+ to_string( Val: 1+ p->tm_mon)+"-"+ to_string( Val: p->tm_mday)
        +" "+to_string( Val: p->tm_hour)+"-"+ to_string( Val: p->tm_min)+"-"+ to_string(  Val: p->tm_sec);
char filename[100] ;  //文件名
strcpy( Destination: filename, Source: fileNameCurrent.c_str());
FILE *fp = fopen( FileName: filename,  Mode: "wb");
if(fp == NULL) {
    printf( Format: "Cannot open file, press any key to exit!\n");
    system( Command: "pause");
    exit( Code: 0);
}
int nCount;
while( (nCount = recv( s: clntSock,  buf: buffer,  len: BUF_SIZE,  flags: 0)) > 0 ){
    fwrite( Buffer: buffer,  ElementSize: nCount,  ElementCount: 1,  Stream: fp);
}
closesocket( s: clntSock);
memset( Dst: buffer,  Val: 0,  Size: BUF_SIZE);
```

Figure 5. Server implementation

B. *Implementation*

When applying the computer vision-based structural deformation monitoring system for displacement monitoring, the static detection is the same as the dynamic detection except that it is not necessary to slide the static detection to the dynamic detection and fill in the time of the dynamic detection after the image acquisition is completed. Take dynamic detection as an example, the process is roughly divided into the following 4 steps:

1) ***Test Preparation:*** Attach the telescope to a tripod and secure the Android phone through the phone clip, then aim at artificial targets.

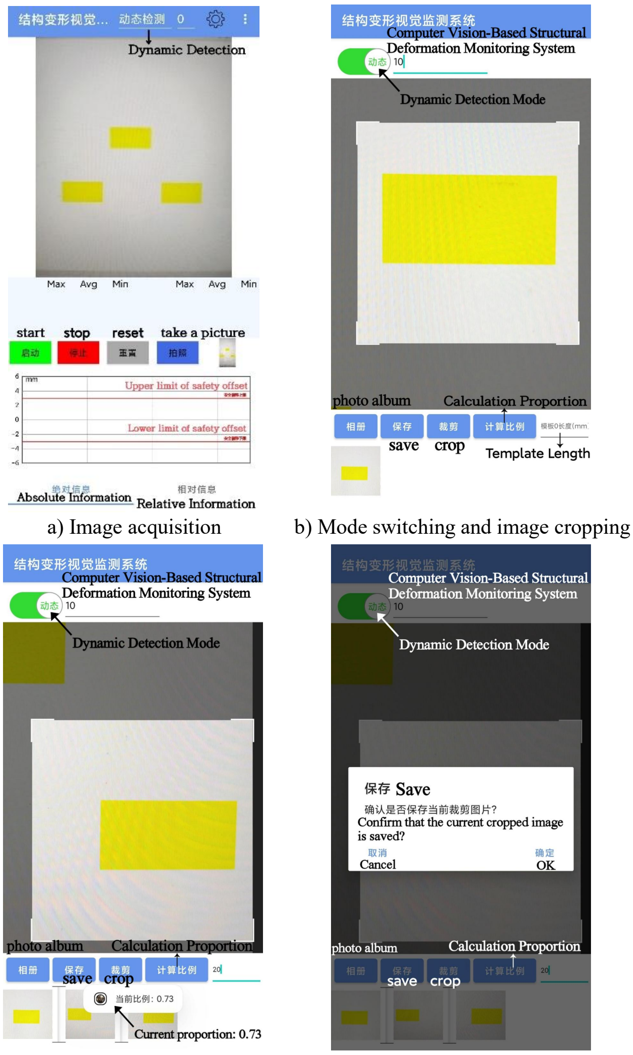
2) ***Image Acquisition and Processing:*** The operation procedure of dynamic detection is shown in **Figure 6**, and the specific operation of image acquisition and processing includes the following 4 steps:

First, click the take a picture button to take a picture of the target marker, and then click the photo next to the take a picture button to jump to the image processing interface.

Second, after jumping to the image processing module, slide the static detection to dynamic detection and fill in the dynamic detection time. Then, zoom the image in the image preview interface so that the individual markers are in the template cropping window, click the crop button, and then select the next marker to repeat the above operation until the completion of the cropping of all the markers. The purpose of this step of the operation is to select a sliding window in the image to be matched with the same size as the template and to extract the features from the selected target region for target tracking.

Third, fill in the actual length of the template and click the Calculate Proportion button, then click the Save button to save the cropped image.

Fourth, click OK in the pop-up interface, and the system will jump to the main interface. Click on the start button to monitor the absolute and relative offsets of the target marker in real time.



a) Image acquisition

b) Mode switching and image cropping

c) Calculation proportion

d) Save

Figure 6. Image acquisition and processing

3) ***Real-time View of Monitoring Data:*** In the bottom area of the main interface of the system, the absolute and relative offset variation curves of each marker in the dynamic detection mode during the set time period will be displayed in real time, as shown in **Figure 7** and **Figure 8**.

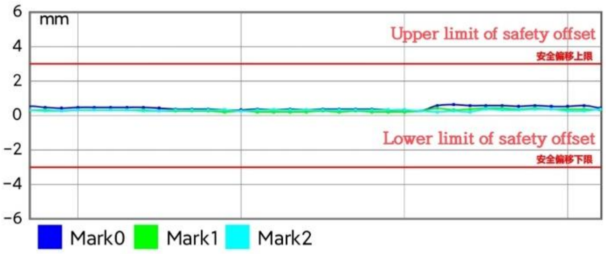


Figure 7. Absolute offset variation curve

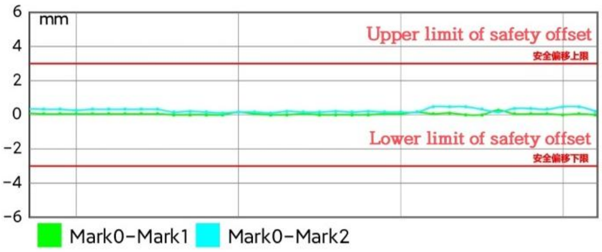


Figure 8. Relative offset variation curve

4) ***Data Transmission:*** When the data monitoring is completed, firstly, click on the More Options button located at the top of the main interface to select Export Data to PC; secondly, enter the server IP and port number in the corresponding position of the pop-up interface, whose IP address is in the same address segment as that of the server (PC); and at the same time, enter the server IP and port number on the PC and select the file storage path and click on Start; finally, click on the Export File button on the client (smartphone) to realize the data transfer. The implementation process is shown in **Figure 9**.

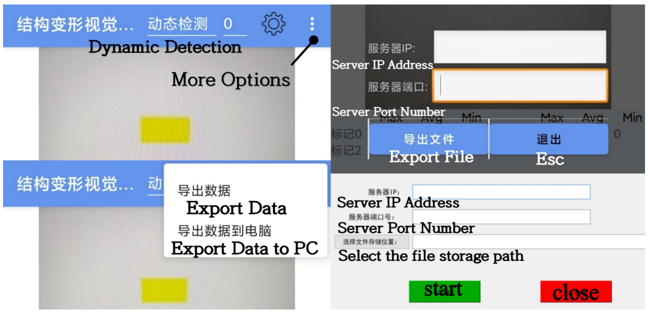


Figure 9. Data transmission

III. Conclusions

This paper illustrates in detail the overall architecture of the computer vision-based structural deformation monitoring system and the implementation method of each module, and takes dynamic detection as an example to show the operation steps when applying the system, and the test results show that the system has great applicability in the monitoring of structural displacement, and the conclusions are as follows:

- The hardware composition is low cost and simple to construct, which can ensure the accuracy of the data.
- High monitoring precision; for structural offsets, accurate to 0.01 mm.
- High monitoring frequency, the measurement of structural offset in the dynamic detection mode can be up to 10 frames per second.
- Monitoring data visualization, the monitoring of structural offset is presented in the form of curves, which is convenient for users to clearly perceive.

Acknowledgment

The study is funded by Shaanxi Huashan Road and Bridge Group Co., Ltd. The authors acknowledge Shaanxi Huashan Road and Bridge Group Co., Ltd. for supporting this work.

References

[1] T. Chen and Z. Zhou, "An improved vision method for robust monitoring of multi-point dynamic displacements with smartphones in an interference environment," *Sensors (Switzerland)*, vol. 20, pp. 1-26, 2020.

[2] C. Xiu, Y. Weng and W. Shi, "Vision and Vibration Data Fusion-Based Structural Dynamic Displacement Measurement with Test Validation," *Sensors*, vol. 23, 2023.

[3] Y. Zhuang, W. Chen, T. Jin, B. Chen, H. Zhang and W. Zhang, "A Review of Computer Vision-Based Structural Deformation Monitoring in Field Environments," *Sensors*, vol. 22, 2022.

[4] X. Bai, M. Yang and B. Ajmera, "An Advanced Edge-Detection Method for Noncontact Structural Displacement Monitoring," *Sensors*, vol. 20, September 2020.

[5] S. Khan, J.-W. Park and S. Ham, "Noncontact Structural Displacement Measurements Using a Low-Cost Motion Tracking System," *IEEE Sensors Journal*, vol. 23, pp. 11695-11703, 2023.

[6] T. Liu, Y. Lei and Y. Mao, "Computer Vision-Based Structural Displacement Monitoring and Modal Identification with Subpixel Localization Refinement," *ADVANCES IN CIVIL ENGINEERING*, vol. 2022, June 2022.

[7] H. Zhang, G. Guan, H. Zhao, X. Liu, L. Xue and C. Xiong, "Design of image transmission system based on TCP/IP protocol," *2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, Dali, China, 2022, pp. 617-622, doi: 10.1109/ICCASIT55263.2022.9987111.

Xiang DONG is a postgraduate student at the Highway School of Chang'an University. She received her B.S. degree in Transportation Engineering from the City College of Southwest University of Science and Technology in 2022. Her current research interests include intelligent detection technology for infrastructure and engineering data analysis.

Maokai LAI is an undergraduate student at Chang'an Dublin International College of Transportation at Chang'an University. Pursuing a dual undergraduate degree in Transportation Urban Planning and Environmental Policy. His current research interests are in intelligent transportation.

Hui LIANG is working as a researcher with Shaanxi Zhengcheng Road and Bridge Engineering Research Institute Co., Ltd., and Shaanxi Huashan Road & Bridge Group Ltd. She received a master's degree in Materials Science and Engineering from Chang'an University in 2013. Her current research interest is applied research in highway municipal engineering.

Peng WU is a postgraduate student at the Highway School of Chang'an University. He received his B.S. degree in Civil Engineering from Chongqing Jiaotong University in 2022. His current research interests include intelligent detection technology for infrastructure and engineering data analysis.

Chaoxia WANG is a postgraduate student at the Xi'an University of Posts and Telecommunications. She received the bachelor's degree in Software Engineering in Shandong Technology and Business University in 2022. The current research interest is deep learning, and the surface defect detection with industrial products.

Ting PENG is an Associate Professor in the Highway School of Chang'an University. He received his B.S. degree in Highway and Urban Street Engineering from Xi'an Highway University in 1999, his M.S. degree in Road and Railway Engineering from Chang'an University in 2004, and his Ph.D. degree in Computer Science from Xi'an Jiaotong University in 2010. His research interests include infrastructure monitoring, big data mining for engineering, highway assets management systems, and artificial intelligence applications.

Session 5A: Communication Network

Chair: Prof. Ammar Muthanna, Saint-Petersburg State University of Telecommunications, Russia

1 Paper ID: 20240199, 343~349

Utilizing Machine Learning for Sensor Fault Detection in Wireless Sensor Networks

Mr. Abubakar Abdulkarim, Mr. Israel Ehile, Prof. Refik Caglar Kizilirmak,

Nazarbayev University. Kazakhstan

2 Paper ID: 20240414, 350~353

Multicore Packet Distribution method using Multicore Network Interface Card based on Tile-gx72 Network Processor

Dr. Choi Won Seok, Mr. Lee Sang Ju , Mr. Kim Jong Oh, Prof. Choi Seong Gon,

Chungbuk National Univertisy. Korea(South)

3 Paper ID: 20240426, 354~359

Flexible Localization Method with Motion Estimation for Underwater Wireless Sensor Networks

Dr. Abdelrahman Samy, Prof. Ammar Hawbani, Prof. Xingfu Wang, Dr. Samah Abdel Aziz, Prof. Liang Zhao, Prof. Nasir Saeed,

University of Science and Technology of China. China

4 Paper ID: 20240378, 360~364

A reliable routing method for remote entanglement distribution under limited resources

Ms. Tianzhu Hu,

USTC. China

5 Paper ID: 20240067, 365~368

Energy Efficiency Analysis of novel Index Modulation-based Non-Orthogonal Multiple Access (IMNOMA) system for 5G Networks

Ms. Shwetha H M, Prof. Anuradha Sundru,

Department of Electronics & communication Engineer. India

Utilizing Machine Learning for Sensor Fault Detection in Wireless Sensor Networks

Abubakar Abdulkarim, Israel Ehile Ehile, Refik Caglar Kizilirmak

Dept. of Electrical and Computer Engineering, Nazarbayev University, Astana, Kazakhstan

***Abstract*—This paper discusses the utilization of machine learning for sensor fault detection in Wireless Sensor Networks (WSNs). The WSN comprises many wireless devices with limited processing power, battery life, and memory capacity. Successfully detecting faulty sensors within a WSN can lead to increased efficiency in the fault detection system, reduced network traffic, and lower power consumption. To enhance the network management, researchers sought a technique for detecting sensor defects. In this paper, machine learning techniques are employed to address the issue of failure detection in WSNs. The utilization of machine learning techniques, specifically Kernel Support Vector Machine (SVM) and Artificial Neural Network (ANN), are demonstrated. The paper further compares the performances of the chosen machine learning algorithms in classifying sensor data as faulty or fault-free. The problem is treated as a binary classification problem. The findings of this study contribute to the development of effective fault detection systems in WSNs.**

***Index Terms*—fault detection, WSN, SVM, ANN, machine learning**

I. Introduction

The Wireless Sensor Network (WSN) system comprises several wireless devices with limited processing power, battery, and memory space. Each device in the network has sensing and limited processing capacities. This comprises sensors and actuators that monitor changes or occurrences in the environment and transmit information to the base station for analysis and subsequent action. The events include temperature, humidity, motion, air density, wind velocity, etc. WSNs have found extensive application in diverse sectors for control and monitoring purposes, such as the military sector, health sector, transportation, smart homes, etc. This includes traffic monitoring, environmental surveillance, patient monitoring, etc.

Sensor nodes are occasionally deployed in hazardous or unmonitored environments such as rain, forests, snow, volcanos, thunder, and wind [1]. Moreover, sensors can experience malfunctions and failures due to their electronic nature. These failures can be broadly classified into three types: communication failure, software failure, and hardware failure [2].

Faulty sensors in WSN may lead to a decrease in the efficiency of the fault detection mechanism, an increase in network traffic, and an increase in power consumption. Sensors devices have limited resources, and they are designed to operate for a long period, which can be from hours, days, months, or years. Therefore, there is a need to conserve sensor power. In addition, it may not be practical to replace their batteries in case they are deployed in hazardous locations.

Since faulty sensors may increase network congestion by forwarding misleading and inaccurate data, a sensor fault detection mechanism is required for better management of the network. The detection mechanisms may enhance sensor data reliability and network bandwidth. However, a complex detection mechanism may increase the sensor's power consumption. Hence, there is a tradeoff between maintaining the service quality of the network and conservation of the sensor's energy [3]. Fault detection mechanisms detect and record faulty sensors in WSN; the record can be used to replace a faulty sensor, fault recovery process, or isolate the faulty sensor from the WSNs.

In this work, our aim is to utilize machine learning methods for fault detection of sensors in the WSNs. The problem is viewed as a binary classification problem; this is because the dataset used is labeled data of sensor readings that are either faulty or fault-free. The paper aims to assess the performance of different machine learning techniques in classifying faulty and fault-free sensor data.

II. Previous works

Several research efforts have been conducted in the literature to deal with the issue of sensor fault detection in WSNs. The techniques proposed in the literature can either be a distributed approach, a centralized approach, or a combination of both which is known as a hybrid approach [4]–[6]. Some of those proposed methods are based on machine learning. An approach based on machine learning was introduced for the identification of sensor faults within WSN [7]. In this method, all sensors within the network transmit their data to a central node referred to as the sink node. For each sensor, the timestamps associated with its data are organized in descending order.

Furthermore, [8] proposed a fault detection mechanism in WSNs using an SVM classifier. The SVM classifier provides a decision function that separates faulty sensor readings from the correct ones. An experimental study is provided to show the performance of their machine learning-based approach. Detection accuracy and the false positive rate were considered as the accuracy metric to be used to evaluate the accuracy of the proposed machine learning-based detection method. The

979-8-3503-4693-0/23/$31.00 ©2023 IEEE

accuracy of the proposed approach was studied using data collected from sensors deployed in a lab environment.

A distributed mechanism for isolating faulty sensors in WSNs has been proposed in [9]. This isolation mechanism allows for faulty sensors to function as routers only, but logically, they are disconnected from the network by the fault detection mechanism. In addition, the mechanism compares sensor data with their neighbors to improve the accuracy of the sensed data and isolate the faulty sensors from the network.

In contrast, a recurrent neural network-based distributed fault detection technique has been proposed in [10]. The technique requires collaborative work between the sensors in the network to detect faults. Authors in [11] focus on the classification and identification of system and data faults. A Markov model is utilized to capture the dynamics of the faulty data and the fault-free data. Moreover, a structural analysis of the Markov model is used to determine the kind of system fault or data fault that affects the measurements of the sensor devices.

Subsequently, a group of K-Nearest Neighbors is created based on the values of the sensor data. Euclidean distances are then computed for each data from the sensor, and any data points that deviate significantly from the norm are identified as potential faults. An artificial neural network-based method for detecting outliers is proposed in [12], [13]. The method uses temperature data captured by sensors in WSN. In the method, a temperature value is considered an outlier when there is a large difference between the predicted temperature value and the expected temperature value. This difference is considered an error value. The proposed method is an anomaly prediction with good performance in terms of prediction accuracy.

In [14], a hybrid prediction technique utilizing a Kalman filter and an Extreme learning machine is presented. The sink node was trained using faulty sensor data using the filter and the extreme learning machine, which served as the predictive classifier. The extreme learning machine has a low communication overhead and can make highly accurate predictions. Utilizing WSN data and artificial random anomalies introduced into the data, the performance of the suggested technique has been assessed. The performance metric used is the computational time and the detection accuracy.

In contrast, our paper utilizes the approach of two machine learning techniques to detect faults in WSNs, as opposed to other authors who used single machine learning approach in detecting faults. Our study assesses and compares the performances of two machine learning techniques in accurately classifying sensor data as either faulty or fault-free. This approach allows us to offer a comprehensive evaluation of the suitability and effectiveness of different machine learning approaches in the context of sensor fault detection, enabling researchers and practitioners to make informed choices based on their specific needs and constraints.

A. Sensor Faults

This section presents types of sensor faults. There are various types and causes of faults in WSNs. Sensor faults can be categorized broadly into two primary groups: faults based on location and those based on time. In time-based faults, the time of faults indicates the duration taken by the fault. Sometimes faults occur temporarily and it only takes a limited duration of time. Moreover, the time-based faults are further divided into transient faults and persistent faults. Transient faults do not stay permanently; they normally disappear quickly. Transient faults detection mechanisms have been proposed in [15]–[18]. On the other hand, persistent faults are the types of faults that occur permanently, they exist till a fault recovery is conducted. In most cases, persistent faults occur to a specific component in the network.

In general, the whole network is not faulty, faults affect only specific components in the network. Therefore, rather than the entire network, fault detection techniques must focus on certain network components. Data-centric faults and system-centric faults are two general categories of faults that may be distinguished based on where they occur. Assume that the data generated by a specific sensor can be described as $d(i_{\mathrm{d}}, t, f(t))$, where i_{d} represents the sensor identity, t denotes the timestamp, and $f(t)$ is a function representing the sensor's value at the given timestamp t. Furthermore, $f(t)$ can be expressed as $\alpha + \beta x + \eta$, where α represents an additive offset, β signifies the multiplicative constant referred to as gain, x stands for the sensor's actual non-faulty value at the timestamp t, and η accounts for noise originating from external factors.

Several offset fault techniques have been discussed in the literature. An offset fault refers to a minor discrepancy in the original data, where it deviates by a specific constant additive value from the anticipated sensor readings. This type of fault can arise due to bad sensor calibration. The model of this type of fault can be expressed as $d' = \alpha + d + \eta$ where d is the original sensor data. Another type of sensor fault is called Gain Fault. Gain Fault occurs when there is a deviation in the rate of change of sensor data from the expected behavior over a certain period. In this type of fault, non-faulty sensor data get multiplied by some constant value to produce faulty sensor data. This fault is modeled as $x' = \beta x + \eta$ where x is the original non-faulty sensor measurement, while β is a constant multiplied by non-faulty sensor measurement due to reasons such as improper calibration of the sensor device. Other sensor faults include stuck-at, data loss, out-of-bounds and random faults [19].

III. Methodology

In this paper, we have compared the performance of two machine learning-based algorithms for sensor fault detection in WSN. The machine learning algorithms are SVM and ANN (a single-layered). SVM is originally designed for binary classification which is the problem targeted in this study. Binary classification deals with the classification between only two classes. Moreover, in this paper we have only two classes of sensor data, the data is either faulty data or fault-free data. In SVM, the predictions only depend on a part of the training data which are close to the decision boundary. These subsets are known as support vectors, and they are the most important data points in SVM classification. Modifying the loss function and combining it with a kernel trick is termed a support vector machine. Support vector machine approaches binary classification problems using the concept of margin [20]. A margin in SVM is defined as the distance between the support vectors of the two classes that are separated by the decision boundary. Therefore, the aim is to choose a decision boundary that provides a maximum margin between the two classes. This is termed as the maximum margin classification.

For a two-class classification problem, the decision boundary is defined by the linear model, where $w^T\phi(x)$ represents a feature vector resulting from applying a feature transformation and b represents the bias term.

$$y(x) = w^T\phi(x) + b \tag{1}$$

where the distance from point x to the hyperplane defined by $y(x) = 0$ in a perpendicular direction is expressed as |y(x)|/||w||. For correct classification, the distance of a point x_n to the decision is determined as

$$\frac{t_n y(x_n)}{||w||} = \frac{t_n(w^T\phi\left(x_n\right) + b)}{||w||}. \tag{2}$$

The margin is determined by the perpendicular distance to the nearest point x_n within the dataset, and to maximize this distance we need to optimize the parameters w and b by solving the optimization equation given as

$$\underset{w,b}{argmax} = \{\frac{1}{||w||}\underset{n}{min}\left[t_n\left(w^T\phi\left(x_n\right) + b\right)\right]\}. \tag{3}$$

However, solving the above equation directly is complex, an alternative problem that is more manageable should be derived. The kernel trick and the concept of Lagrange multipliers are employed to solve this problem.

To utilize the previously trained model for classifying new data points, the prediction is expressed in terms of the Lagrange multipliers instead of weights as

$$y(x) = \sum a_n k(x, x_n) + b \tag{4}$$

where a_n is the Lagrange multiplier and $k(x, x_n)$ is the kernel function. The algorithm for SVM shown in Algorithm 1 [21] iteratively trains a Support Vector Machine (SVM) by initializing support vectors, generating subsample spaces, evaluating Lagrange multipliers, determining optimal directions and biases, and ultimately obtaining the decision boundary.

A broad, useful approach to understanding discrete-valued, real-valued, or vector-valued functions from a collection of training samples is provided by artificial neural networks. ANN learning is an ideal option for issues where the training data correlates to noisy, complex sensor data because of its robustness to errors in the training data. Multilayer perceptron employs algorithms such as the backpropagation algorithm. Gradient descent is a technique used by backpropagation algorithms to fine-tune network parameters to best suit a training set of input-output pairs [22].

The relevant cost function based on sum squared training errors is stated as follows for a network with many output units

$$E\left(\overrightarrow{w}\right) = \frac{1}{2}\sum_{d\in D}\ \sum_{k\in outputs}\left(t_{kd} - O_{kd}\right)^{\,2} \tag{5}$$

where O_{kd} is the network output linked to network output k and training example d, and t_{kd} is the target value. The weight values under consideration are determined using the gradient descent rule to minimize E. The backpropagation algorithm used to discover these weights is shown in Algorthim 2 [23]. In a typical application, the weight-update loop in backpropagation may be iterated thousands of times. To end the process, several termination criteria might be used.

Algorithm 1 SVM Algorithm [21]

```
Initialize support vectors and $k(x, x_n)$
for $s \leftarrow 1$ to $S$ do
    Generate subsample space
    Evaluate Lagrange Multipliers $(a_n)$
    Obtain the optimal direction and bias
end for
Obtain the decision boundary
```

Algorithm 2 (Backpropagation Algorithm) [23]

```
Establish a feed-forward neural network
Initialize all the network weights
for each training example $(x, t)$ do
    Forward propagate the input through the network
    a. Input the instance $x$ into the network and calculate
    $\delta_h$ for each unit $u$ in the network
    Backward propagate the errors through the network
    for each network output unit $k$, the error term is
    computed as: do
        $\delta_h \leftarrow -o_k(1 - o_k)(t_k - o_k)$
    end for
    for each hidden unit $h$, the error term is calculated
    as: do
        $\delta_h \leftarrow o_h(1 - o_h)\sum_{k\in\text{outputs}} w_{kd}\delta_k$
    end for
    Update each network weight $w_{jk}$ as follows:
    $w_{ji} \leftarrow w_{ji} + \Delta w_{ji}$
    where;
    $\Delta w_{ji} = \alpha\delta_j x_{ji}$
end for
```

IV. Results and Discussion

In this section, we present the effectiveness of the proposed sensor failure detection systems. We consider two different WSN deployment scenarios. Fig. 1 presents the single-hop scenario, where the sensors can directly communicate with the base station. Fig. 2 presents a multi-hop scenario sensor deployment where the sensors communicate with the base stations through a router node.

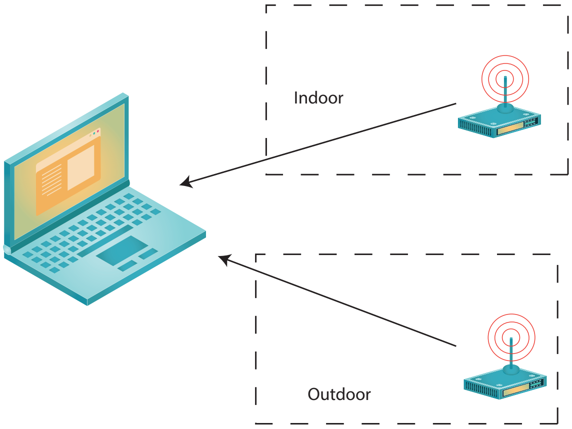


Fig. 1. Single-hop scenario sensor deployment [24].

A publicly accessible dataset of sensor measurements was trained using the single-layered NN and SVM machine learning methods with linear kernel function. Data from single-hop and multi-hop wireless sensor networks are included in the dataset, which is labeled wireless sensor networks data. Specifically, the dataset from four distinct sensors has been utilized to train the machine learning models to evaluate the performance of the sensor fault detections. The four different scenarios are the Indoor Single-hop scenario, Outdoor Single-hop scenario, Indoor multi-hop scenario, and Outdoor multi-hop scenario. In a single-hop scenario, there is direct communication between the sensor and the base station, sensors can directly send their sensed data to the base station. In a multi-hop scenario, the sensed data is routed through an intermediate node between the source and the destination. The dataset consists of two features which are the value of humidity and the value of temperature. These values are collected within six hours at an interval of 5 seconds. For both SVM and the single-layered perceptron, the data for all four sensors with different scenarios have been randomly shuffled and training utilizes 80% of the data, while the remaining 20% is allocated for testing. In perceptron, one hidden layer is used with a varying number of neurons. A varying number of neurons in the hidden layer will help in determining the best architecture with the smallest testing and training errors. Hence the architecture with the smallest testing and training errors will have the highest accuracy. Specifically, we have used 2, 4, 10, 20, 50, and 100 neurons in the perceptron hidden layer, and the average testing error and training error have been observed.

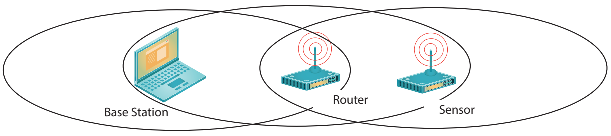


Fig. 2. Multi-hop scenario sensor deployment [24].

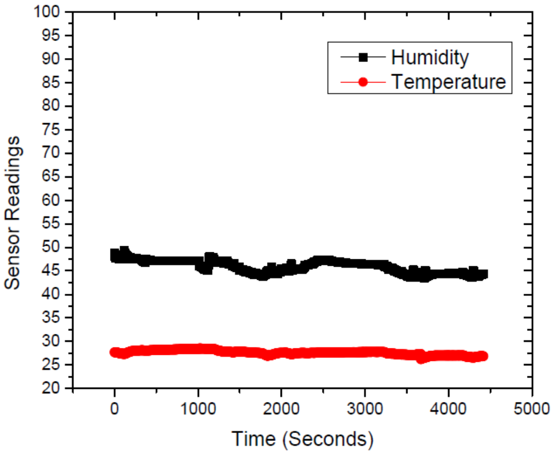


Fig. 3. Single hop scenario Indoor Sensor reading without anomalies

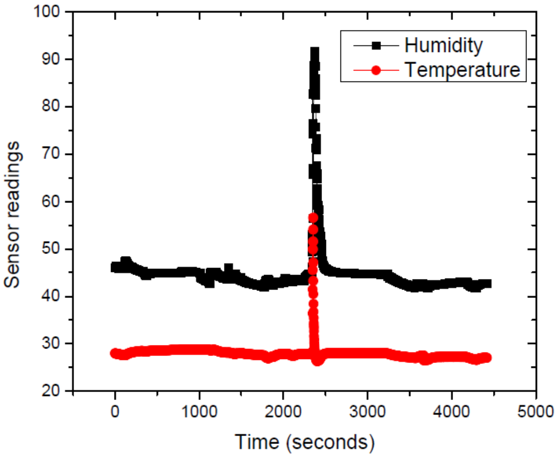


Fig. 4. Single hop scenario Indoor Sensor reading with anomalies

Fig. 3 presents the Indoor Sensor readings without faults for the Single hop scenario. It consists of humidity readings and temperature readings collected over six hours at an interval of every five seconds. Fig. 4 presents the Indoor Sensor reading with anomalies. From the Fig. 4, a spike can be observed at around halfway in the x-axis, this spike presents the anomalies in the sensor readings.

Figs. 5 to 8 present the results of the sensor faults detection using an SVM classifier. Measurements from four sensors have been used for the four different scenarios. Fig. 5 presents the SVM classification of the Indoor Sensor Single scenario. Fig. 6 presents the SVM classification of the Indoor Sensor Multi-hop scenario. Fig. 7 presents the SVM classification of the Outdoor single-hop scenario. Fig. 8 presents the SVM classification of the Outdoor multi-hop scenario. In the figures, the decision boundary of the SVM classification has separated the normal sensor data and the faulty sensor data. Normal sensor data is shown in red color and the faulty sensor data is shown in green color. Table 1 presents the Single-layered perceptron testing accuracy for sensor data of the four different scenarios.

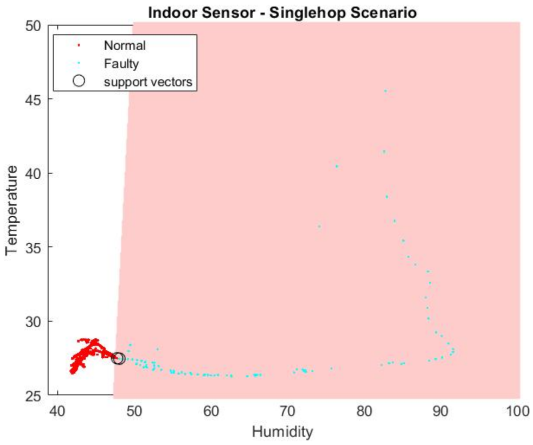


Fig. 5. SVM classification of Indoor Sensor Single scenario

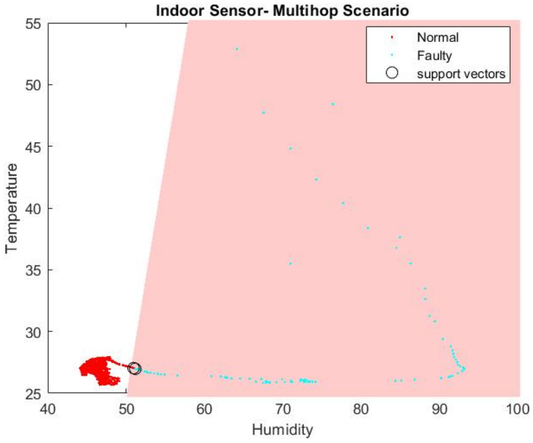


Fig. 6. SVM classification of Indoor Sensor Multi-hop scenario

Table 2 presents the performance comparison between the two machine learning methods used in sensor fault detection. In this study, testing the accuracy of the two machine learning methods is considered as the performance metric. The SVM classifier and the neural network, both two machine learning methods have achieved good classification accuracy, as shown in the figure the accuracy of all four scenarios is above 99%. Moreover, SVM has higher accuracy compared to the neural networks for the three scenarios. Specifically, SVM has an accuracy of 99.77% for the indoor single-hop scenario compared to 99.38% accuracy of the neural network, SVM has an accuracy of 100.00% for the Indoor Multi-hop scenario compared to 99.77% accuracy of the neural networks, SVM has an accuracy of 99.89% for the Outdoor Multi-hop scenario compared to 99.44% accuracy of the neural networks, but for the Outdoor Single-hop scenario, the neural networks have higher classification accuracy of 99.98% compared to the SVM accuracy of 99.80%.

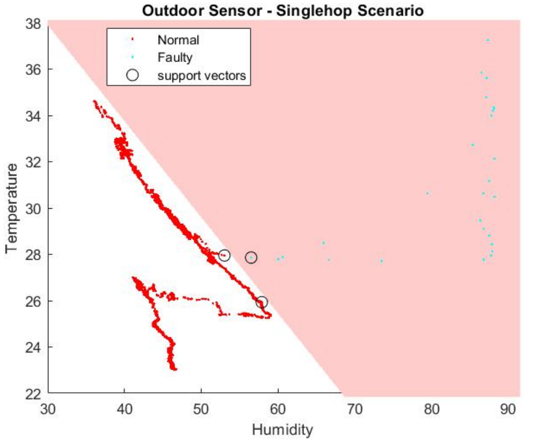


Fig. 7. SVM classification of Outdoor Sensor Single scenario

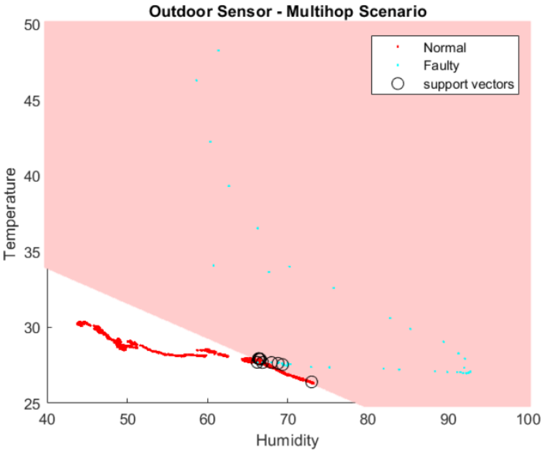


Fig. 8. SVM classification of Outdoor Sensor Multi-hop scenario

TABLE I. Testing Accuracy (%) for Different Hidden Layer Neurons

Hidden Layer Neurons	Indoor		Outdoor	
	Single-hop	**Multihop**	**Single-hop**	**Multihop**
2	99.98	99.98	99.99	99.99
4	99.98	99.99	99.99	99.98
10	99.99	99.99	99.99	99.99
20	99.99	100.00	99.99	99.99
50	99.88	99.76	99.99	99.99
100	96.48	98.90	99.95	96.69
Average	**99.38**	**99.77**	**99.98**	**99.44**

TABLE II. Scenarios Comparison: SVM vs. Single-layered Perceptron

Scenarios	SVM	Single-layered Perceptron
Indoor Single-hop	99.77%	99.38%
Indoor Multi-hop	100.00%	99.77%
Outdoor Single-hop	99.80%	99.98%
Outdoor Multi-hop	99.89%	99.44%

V. Conclusion and future work

In this paper, machine learning-based sensor fault detection in WSN has been studied. Specifically, a performance comparison of SVM and Single-layered perceptron in sensor fault detection has been presented. The dataset used in this study is labeled wireless sensor networks data collected from a single hope scenario and multi-hop scenario from both indoor and outdoor sensor deployment. Specifically, four scenarios have been considered. The scenarios are indoor single-hop scenario, indoor multi-hop scenario, Outdoor single-hop scenario, and Outdoor multi-hop scenario. For all the aforementioned scenarios, the two machine learning methods have archived good classification accuracy of above 99%. From the simulation results, SVM has archived relatively higher accuracy than the Single-layered perceptron. In future work, we plan to use another sensor dataset that has more than two features and plan to use other machine learning algorithms and compare their performance so that we can observe the most suitable machine learning algorithm for the detection of sensor faults in WSN.

ACKNOWLEDGEMENT

This work was supported by the Faculty Development Competitive Research Grant Program of Nazarbayev University (grant no. 20122022FD4125).

REFERENCES

[1] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Computer networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[2] T. Muhammed and R. A. Shaikh, "An analysis of fault detection strategies in wireless sensor networks," *Journal of Network and Computer Applications*, vol. 78, pp. 267–287, 2017.

[3] M. Yu, H. Mokhtar, and M. Merabti, "Fault management in wireless sensor networks," *IEEE Wireless Communications*, vol. 14, no. 6, pp. 13–19, 2007.

[4] Z. Feng, J. Q. Fu, and Y. Wang, "Weighted distributed fault detection for wireless sensor networks based on the distance," in *Proceedings of the 33rd Chinese Control Conference*. IEEE, 2014, pp. 322–326.

[5] R. R. Panda, B. S. Gouda, and T. Panigrahi, "Efficient fault node detection algorithm for wireless sensor networks," in *2014 International Conference on High Performance Computing and Applications (ICHPCA)*. IEEE, 2014, pp. 1–5.

[6] C. Titouna, M. Aliouat, and M. Gueroui, "Outlier detection approach using bayes classifiers in wireless sensor networks," *Wireless Personal Communications*, vol. 85, pp. 1009–1023, 2015.

[7] A. Abid, A. Kachouri, A. B. F. Guiloufi, A. Mahfoudhi, N. Nasri, and M. Abid, "Centralized knn anomaly detector for wsn," in *2015 IEEE 12th International Multi-Conference on Systems, Signals & Devices (SSD15)*. IEEE, 2015, pp. 1–4.

[8] S. Zidi, T. Moulahi, and B. Alaya, "Fault detection in wireless sensor networks through svm classifier," *IEEE Sensors Journal*, vol. 18, no. 1, pp. 340–347, 2017.

[9] M.-H. Lee and Y.-H. Choi, "Fault detection of wireless sensor networks," *Computer Communications*, vol. 31, no. 14, pp. 3469–3475, 2008.

[10] O. Obst, "Distributed fault detection in sensor networks using a recurrent neural network," *Neural processing letters*, vol. 40, pp. 261–273, 2014.

[11] E. U. Warriach and K. Tei, "Fault detection in wireless sensor networks: A machine learning approach," in *2013 IEEE 16th International Conference on Computational Science and Engineering*. IEEE, 2013, pp. 758–765.

[12] K. Zhang, K. Yang, S. Li, D. Jing, and H.-B. Chen, "Ann-based outlier detection for wireless sensor networks in smart buildings," *IEEE Access*, vol. 7, pp. 95 987–95 997, 2019.

[13] F. Aliyu, S. Umar, and H. Al-Duwaish, "A survey of applications of artificial neural networks in wireless sensor networks," in *2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO)*, 2019, pp. 1–5.

[14] P. Biswas, R. Charitha, S. Gavel, and A. S. Raghuvanshi, "Fault detection using hybrid of kf-elm for wireless sensor networks," in *2019 3rd international conference on trends in electronics and informatics (ICOEI)*. IEEE, 2019, pp. 746–750.

[15] K. P. Sharma and T. P. Sharma, "rdfd: reactive distributed fault detection in wireless sensor networks," *Wireless Networks*, vol. 23, pp. 1145–1160, 2017.

[16] A. Mahapatro and A. K. Panda, "Choice of detection parameters on fault detection in wireless sensor networks: A multiobjective optimization approach," *Wireless personal communications*, vol. 78, no. 1, pp. 649–669, 2014.

[17] M. N. Sahoo and P. M. Khilar, "Diagnosis of wireless sensor networks in presence of permanent and intermittent faults," *Wireless personal communications*, vol. 78, no. 2, pp. 1571–1591, 2014.

[18] ——, "Distributed diagnosis of permanent and intermittent faults in wireless sensor networks," in *Advanced Computing, Networking and Informatics-Volume 2: Wireless Networks and Security Proceedings of the Second International Conference on Advanced Computing, Networking and Informatics (ICACNI-2014)*. Springer, 2014, pp. 133–141.

[19] S. Gnanavel, M. Sreekrishna, V. Mani, G. Kumaran, R. Amshavalli, S. Alharbi, M. Maashi, O. I. Khalaf, G. M. Abdulsahib, A. D. Alghamdi *et al.*, "Analysis of fault classifiers to detect the faults and node failures in a wireless sensor network," *Electronics*, vol. 11, no. 10, p. 1609, 2022.

[20] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[21] B. Hou, B. Zhou, X. Li, L. Yi, Q. Wei, and R. Zhang, "Nonlinear error compensation of capacitive angular encoders based on improved particle swarm optimization support vector machines," *IEEE Access*, vol. 8, pp. 124 265–124 274, 2020.

[22] T. M. Mitchell, *Machine learning*. McGraw-hill New York, 1997, vol. 1, no. 9.

[23] A. Darii, M. Moll, M. S. Nistor, S. Pickl, O. C. Novac, C. M. Novac, I. M. Gordan, and C. E. Gordan, "Analysis, combination and integration of neuroevolution and backpropagation algorithms for gaming environment," in *2023 15th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. IEEE, 2023, pp. 1–5.

[24] S. Suthaharan, M. Alzahrani, S. Rajasegarar, C. Leckie, and M. Palaniswami, "Labelled data collection for anomaly detection in wireless sensor networks," in *2010 sixth international conference on intelligent sensors, sensor networks and information processing*. IEEE, 2010, pp. 269–274.

Abubakar Abdulkarim received his bachelor's degree of Science in electrical and electronic engineering at Eastern Mediterranean University, Turkish Republic of Northern Cyprus in 2017. He joined Binyaminu Usman Polytechnic Hadejia, Nigeria as an Assistant lecturer in 2019. He is currently a Master student in Electrical and Computer Engineering at Nazarbayev University, Kazakhstan. His research interests include Application of UAVs in wireless communication, Application of EV in logistic, Smart grid.

Israel Ehile Ehile received the B.Eng. in electrical and electronics engineering from the University of Agriculture, Makurdi, Nigeria in 2018. He was a Research Assistant with the Department of Electrical and Computer Engineering, Nazarbayev University, Astana, Kazakhstan in 2023. He is currently a Master student in Electrical and Computer Engineering at Nazarbayev University, Astana, Kazakhstan. His research interests include Communication in Inter-Satellite, Routing in Deep Space Network, Computer Networks and Cloud Security.

Refik Caglar Kizilirmak (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 2004 and 2006, respectively, and the Ph.D. degree from Keio University, Yokohama, Japan, in 2010. He was with the Communications and Spectrum Management Research Center, Turkey, on several telecommunication and defense industry projects. He is currently with the Department of Electrical and Computer Engineering, Nazarbayev University, Astana, Kazakhstan. He has contributed to the technical requirements document of IEEE 802.15.7r1 standardization, which will enable visible light communication. He has authored several articles in the field of wireless communications and has filed three patent applications with the patent offices of USA and Japan.

Multicore Packet Distribution method using Multicore Network Interface Card based on Tile-gx72 Network Processor

Won Seok Choi*, Sang Ju Lee**, Jong Oh Kim**, Seong Gon Choi*

* Information & Communication Engineering, Chungbuk University, Cheongju-si, Chungcheongbuk-do, Korea

***Fisys Inc., 168, Gajeong-ro, Yuseong-gu, Daejeon, Korea*

wschoi@chungbuk.ac.kr, angelet86@fisys.co.kr, jokim@fisys.co.kr, choisg@chungbuk.ac.kr

Abstract— **We propose a data plane acceleration technology to deliver data from the network to the host system in a high-performance computing environment. In the fourth industrial revolution, server systems are developing into high-performance computing systems through convergence with major technologies such as IoT, cloud, AI, and self-driving cars. The 4th industrial revolution is the convergence of various technologies and IT, requiring various flows and large amounts of data to be processed on servers. When transferring packets from the network interface card to the host server, packet processing in kernel space has a large overhead. Additionally, for fast packet processing by the host server, packets must be processed according to core affinity. Therefore, we propose a load balancing data transmission method to 48 cores based on Tile-Gx72 network processor to transfer data from the network interface card to the host CPU by kernel bypass in a multi-core-based high-performance server system. In addition, the performance of the 48 cores-based load balancing data transmission system based on the Tile-Gx72 network processor is confirmed through implementation.**

***Keywords*— Multicore, Network Interface Card, Network Processor, Packet distribution, Tile-gx72**

I. Introduction

The 4th Industrial Revolution is developing through the convergence of various technologies such as IoT, cloud, AI, and self-driving vehicle, and the use and amount of data required to provide services using technology is increasing. [1].

Internet Data Center (IDC) predicts that the amount of data generated will reach 73 trillion bytes by 2025 due to the development of IoT technology, and the generated data is expected to be transmitted to servers through the network

IDC predicts that the amount of data generated by IoT technology will reach 73 trillion gigabytes in 2025 [2]. This data will be transmitted to servers or devices through the network. This means that the amount of data transmitted through the network and the amount of data processed by the server CPU increases. However, the gap between network line speeds and the speed at which CPUs can produce and consume data is rapidly widening [3].

In particular, when transmitting data between a Network Interface Card (NIC) and a CPU, a kernel space is a bottleneck, and overhead due to packet processing causes performance degradation. In this case, the problem is usually solved by using the kernel bypass to directly deliver packets to the user space [4].

Solutions that use the kernel bypass include Data Plane Development Kit (DPDK), Single Root I/O Virtualization (SR-IOV), and Jae Woo Ahn et al.'s load distribution method using multicore based network interface card for high-performance computing System.

DPDK is an intel architecture-based packet processing optimization system that uses environment abstraction layer and poll mode driver as software to the bypass kernel space and transmit packets to user space [5]. However, in order to use the DPDK, a NIC driver that supports the DPDK is required, which limits NICs that can be used.

SR-IOV is a technology defined by PCI-SIG that shares I/O devices in a virtualization environment. In a virtualization server using a virtual machine, the SR-IOV maps the virtual NIC of the virtual machine to a Virtual Function (VF) of the physical NIC of the host server, bypassing the host server's kernel space and enabling fast packet transmission and reception [6]. As a result, SR-IOV reduces the overhead of kernel space on the host server and provides the virtual NIC functionality, but cannot reduce the overhead of kernel space on the virtual Machine.

Jae Woo Ahn et al. propose a load distribution method using a multicore based NIC for high-performance computing system that distributes packets of up to 20 Gbps to 28 cores of servers using a NIC based on Tile-gx36 network processor [4]. For fast packet processing in the host server CPU, packets are distributed according to core affinity according to flow. However, as the computing power and number of cores of servers are increasing, there are limitations in transmitting packets over 20 Gbps and distributing packets over 28 cores.

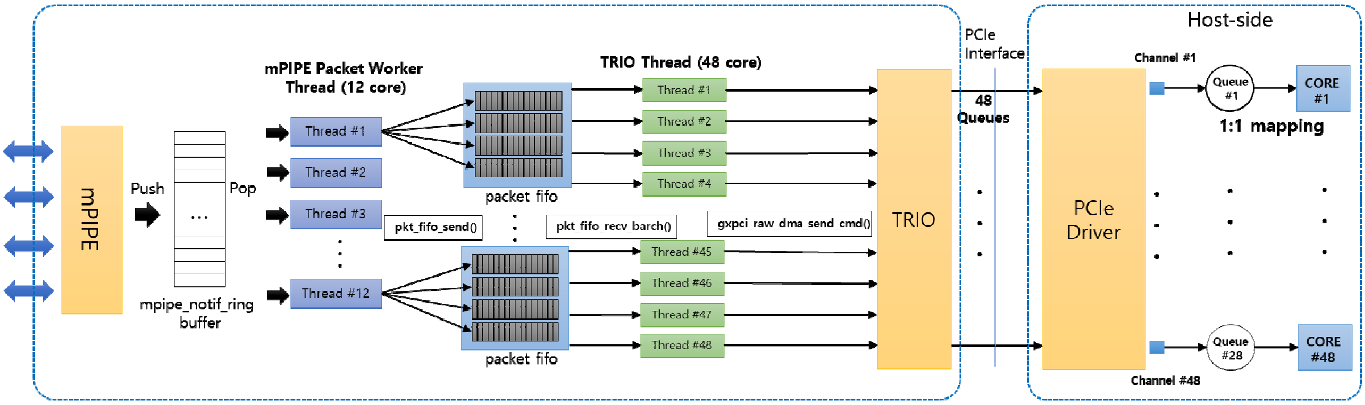


Figure 1. Architecture of the proposed system

Therefore, in order to receive 40 Gbps packets from a Multicore-based high-performance server system, we use Tile-gx72 that is a multicore-based network processor to accept 40 Gbps packet reception. Furthermore, the proposed system distributes packets of 40 Gbps to 48 host server cores by bypassing the server's kernel space. Additionally, performance tests were performed using laboratory-level instruments.

The remainder of this paper is organized as follows: we first introduce data plane acceleration method such as DPDK, SR-IOV in Section 2. In Section 3, we introduce the proposed method in detail. Section 4 reports the implementation and test results. Finally, conclusions and future studies are described in Section 5.

II. Related Work

In this session, we introduce DPDK and SR-IOV as data plane acceleration methods.

A. DPDK

DPDK is a set of data plane development tools provided by Intel that provides library functions and driver support for efficient packet processing in user space on Intel processor architectures. This is different from the purpose of the universal design of Linux system. The DPDK focuses on high-performance processing of packets in network applications. The best feature is that DPDK applications run on user space and use their own data plane libraries to send and receive data packets, bypassing the Linux kernel protocol stack for packet processing [5].

However, in order to use the DPDK, a NIC driver that supports the DPDK is required, which limits NICs that can be used.

B. SR-IOV

SR-IOV proposes a set of hardware improvements to PCIe devices that aim to eliminate key Virtual Machine Monitor (VMM) interventions in performance data movement, such as packet classification and address translation. The SR-IOV inherits Direct I/O technology, which uses I/O Memory Management Unit (IOMMU) to offload memory protection and address translation. The SR-IOV capable devices can create multiple "lightweight" instances of PCI function entities, called VFs. Each VF can be assigned to a guest for direct access [6].

As a result, the SR-IOV reduces the overhead of kernel space on the host server and provides the virtual NIC functionality, but cannot reduce the overhead of kernel space on the virtual machine because it still performs packet processing on kernel space of the virtual machine.

Therefore, we propose the load balancing data transmission method to 48 cores based on the Tile-Gx72 network processor to transfer data from the network interface card to the host CPU by kernel bypass in a multi-core-based high-performance server system.

III. Proposed Method

We propose a system that delivers packets to 48 cores of the host based on Tile-gx 72 network processor in order to distribute the data collected from the network interface card according to the flow with load balancing.

Figure 1 shows the structure of the proposed 48 cores-based data transmission system with load balancing. The proposed structure consists of tile-side, which consists of a network interface card based on the Tile-Gx72 process, and host-side, where the network interface card is installed. Tile-side and Host-side are connected through the PCIe interface.

Tile-side consists of multicore Programmable Intelligent Packet Engine (mPIPE), mPIPE packet worker thread, TRIO Thread, and Transactional I/O (TRIO).

The mPIPE is a hardware packet processing engine that performs packet collection, packet transmission, packet header parsing, packet distribution, packet buffer management, load balancing, and L4 checksum. It performs the same role as a network interface card in a typical PC or server [3]. Additionally, mPIPE hashes packets transmitted to the host and stores them in the mpipe_notif_ring buffer.

The TRIO handles Direct Memory Access (DMA) data movement and buffer management for data transfer through PCIe between the Tile memory system and the Host memory system [3].

The mPIPE packet worker thread is a thread that processes data collected from the mPIPE, and the transactional I/O (TRIO) thread is a thread that controls the TRIO and delivers packets to the host [3].

The mPIPE packet worker thread is responsible for processing packets collected from mPIPE. The mPIPE packet worker thread uses the packet information stored in the mpipe_notif_ring buffer to match the hashing-based flow affinity and transmits it to the TRIO thread as a packet fifo. The mPIPE packet worker thread uses a total of 12 tile cores, each transmitting with 4 TRIO threads [3].

The TRIO thread is responsible for transmitting data collected from the mPIPE packet worker thread to the host using direct memory access technology. There are 48 TRIO threads, each mapped 1:1 to transmit packets to 48 cores of host.

Host-side consists of a PCIe driver and host core to support the 48-core-based load balancing data transmission software of the network interface card.

Packets with flow affinity from mPIPE packet worker threads are delivered to 48 cores of host by matching core affinity.

IV. Implementation Results

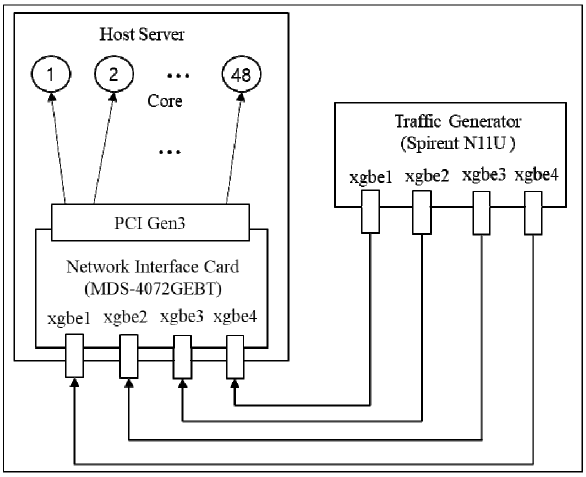


Figure 2. Ideal test environment

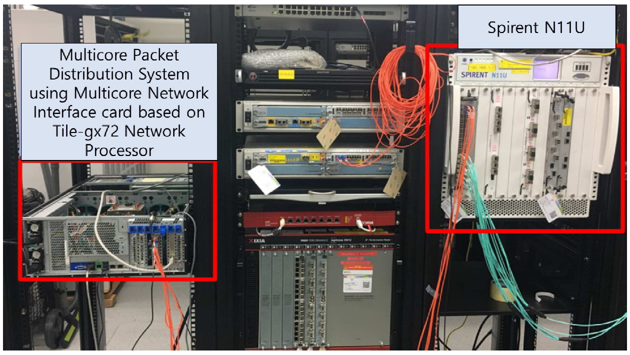

Multicore Packet Distribution System using Multicore Network Interface card based on Tile-gx72 Network Processor
Spirent N11U

Figure 3. Real test environment

We analyze the performance of the proposed multicore packet distribution method using MDS-4072GEBT and Spirent N11U.

Figure 2 shows ideal test environment and figure 3 shows real test environment to confirm performance of the proposed method. We performed performance tests using the MDS-4072GEBT that is a Tile-gx72 network processor-based network interface card manufactured by Fisys Inc. The MDS-4072GEBT network interface card was installed on the host server through the PCI Gen3 interface and the proposed 48-core data transmission system was operated. Traffic generation was performed using the Spirent N11U.

TABLE 1. MDS-4072GEBT Network Interface CarD

List	Specification
Network Processor	- Tile-Gx72 - Core: 72
Interface	- 10Gbps SFP+ X 4 Port
PCIe	- PCI Gen3
OS	- Tile OS

TABLE 2. Host System

List	Specification
CPU	- Intel(R) Xeon(R) CPU - E5-2695 v3 @ 2.30GHz - Core: 14 x 2 - Thread: 28 x 2
PCIe	- PCI Gen3
OS	- CentOS 6

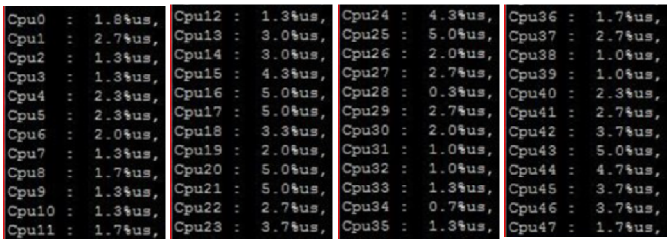
The specifications of MDS-4072GEBT are shown in table 1. The network processor is equipped with Tile-gx72 and can support 10 Gbps SFP+ 4 ports. Furthermore, it supports PCI Gen3 and PCIe Gen2.

Table 2 shows the system information of the host system. The host system uses Intel(R) Xeon(R) CPU E5-2695 v3 CPU, and a total of 56 logical cores can be used.

In this performance test, traffic was generated using 4 ports of the Spirent N11U, and the results of data transmission from the MDE-4072GEBT to 48 host cores are shown in figure 4 and figure 5. Traffic settings were set to 400 flows of 512 bytes and different IPs, and traffic was transmitted at 40Gbps.

Figure 4 shows utilization of host cores using top program that is provides a dynamic real-time view of a running system in Linux system. It was confirmed that packets were distributed to 48 host cores, resulting in a load of less than 5 percent.

Figure 5 shows the results of packet transmission from the Spirent N11U to the MDS-4072GEBT. The left side of figure 5 shows the amount of packets transmitted from four SFP+ ports on the Spirent N11U. The right side of figure 5 shows the amount of packets received by 48 host cores. The Spirent N11U transmitted 739,563,206 packets to the host server at 40 Gbps. As a result, it was confirmed that the MDS-4072GEBT installed on the host server received packets and transmitted 739,563,206 packets to 48 host cores. We used a self-developed host application to check received packets on host cores.



```
Cpu0  :  1.8%us,
Cpu1  :  2.7%us,
Cpu2  :  1.3%us,
Cpu3  :  1.3%us,
Cpu4  :  2.3%us,
Cpu5  :  2.3%us,
Cpu6  :  2.0%us,
Cpu7  :  1.3%us,
Cpu8  :  1.7%us,
Cpu9  :  1.3%us,
Cpu10 :  1.3%us,
Cpu11 :  1.7%us,
```

a. core 0~11

```
Cpu12 :  1.3%us,
Cpu13 :  3.0%us,
Cpu14 :  3.0%us,
Cpu15 :  4.3%us,
Cpu16 :  5.0%us,
Cpu17 :  5.0%us,
Cpu18 :  3.3%us,
Cpu19 :  2.0%us,
Cpu20 :  5.0%us,
Cpu21 :  5.0%us,
Cpu22 :  2.7%us,
Cpu23 :  3.7%us,
```

b. core 12~23

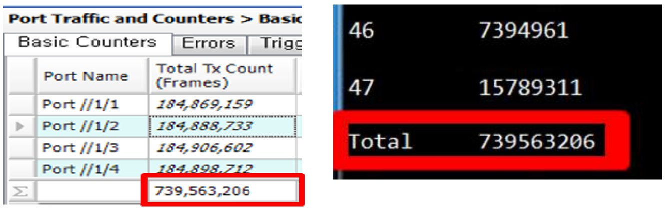
```
Cpu24 :  4.3%us,
Cpu25 :  5.0%us,
Cpu26 :  2.0%us,
Cpu27 :  2.7%us,
Cpu28 :  0.3%us,
Cpu29 :  2.7%us,
Cpu30 :  2.0%us,
Cpu31 :  1.0%us,
Cpu32 :  1.0%us,
Cpu33 :  1.3%us,
Cpu34 :  0.7%us,
Cpu35 :  1.3%us,
```

c. core 24~35

```
Cpu36 :  1.7%us,
Cpu37 :  2.7%us,
Cpu38 :  1.0%us,
Cpu39 :  1.0%us,
Cpu40 :  2.3%us,
Cpu41 :  2.7%us,
Cpu42 :  3.7%us,
Cpu43 :  5.0%us,
Cpu44 :  4.7%us,
Cpu45 :  3.7%us,
Cpu46 :  3.7%us,
Cpu47 :  1.7%us,
```

d. core 36~47

Figure 4. Result of top command



Port Traffic and Counters > Basic

	Port Name	Total Tx Count (Frames)
	Port //1/1	184,869,159
▶	Port //1/2	184,888,733
	Port //1/3	184,906,602
	Port //1/4	184,898,712
Σ		739,563,206

a. Spirent N11U

```
46      7394961

47      15789311

Total   739563206
```

b. received packets of host side

Figure 5. Send and received packets

V. Conclusions

We proposed a 48-core data transmission system based on Tile-gx72 network processor to distribute 48 cores of host by distributing the load with flow affinity. Furthermore, the proposed system was tested through empirical testing. As a result of the test, it was confirmed that the load was distributed across 48 host cores and packets were transmitted with keeping under 5% core load. The proposed system is a system that distributes packet load in a high-performance server system and transmits packets from the network interface to the host. It can be used in high-performance packet processing servers such as self-driving cars and internet of things that have various flow and high-capacity traffic characteristics.

In further study, we plan to conduct performance experiments and research by linking the load-balancing data transmission system with the host's service application.

Acknowledgment

This work was partly supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2024-2020-0-01462, 33%) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). This work was partly supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2020R1A6A1A12047945, 34%). This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C1006167, 33%).

*corresponding author is S.G. Choi(choisg@cbnu.ac.kr).

References

[1] Hansen, E. B., and Bøgh, S., "Artificial intelligence and internet of things in small and medium-sized enterprises: A survey," *Journal of Manufacturing Systems*, 58, pp. 362-372. 2021.

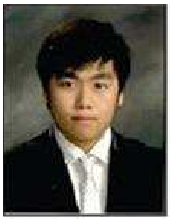
[2] Tomer, V., and Sharma, S., " Detecting IoT Attacks Using an Ensemble Machine Learning Model," *Future Internet*, 14(4), 2022.

[3] Lin, Jiaxin, et al. "{PANIC}: A {High-Performance} Programmable {NIC} for Multi-tenant Networks," *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, 2020.

[4] Ahn, J. W., Kim, J. B., Choi, W. S., Kim, J. O., and Choi, S. G., "Load distribution method using multicore based NIC for high-performance computing system," *In 2018 20th International Conference on Advanced Communication Technology (ICACT)*, pp. 90-93, Feb. 2018.

[5] Zhu, W., Li, P., Luo, B., Xu, H., and Zhang, Y. "Research and implementation of high performance traffic processing based on intel DPDK," *In 2018 9th international symposium on parallel architectures, algorithms and programming (PAAP)*, pp. 62-68, 2018.

[6] Dong, Y., Yang, X., Li, J., Liao, G., Tian, K., and Guan, H. "High performance network virtualization with SR-IOV," *Journal of Parallel and Distributed Computing*, 72(11), pp. 1471-1480, 2012.



Won Seok Choi received B.S. and Ph.D. degree in the College of Electrical and Computer Engineering, Chungbuk National University, Korea in 2008 and 2014 respectively. He is currently researcher in Research institute of Computer and Information Communication, Chungbuk National University. His research interests include vehicle network, energy saving network, SDN, NFV and NGN.

Sang Ju Lee recived B.S and M.S. degree in the College of Electrical and Computer Engineering, Chungbuk National University, Korea in 2014 and 2016 respectively, He is Senior Researcher in Fisys Inc. His research intersts include SDN, NFV and CCIX

Jong Oh Kim received B.S. and M.S. degree in Electronics Engineering from Kyeongbuk National University in 1990 and 1992 respectively. He is currently CEO in Fisys Inc. His research interests is SDN, NFV

Seong Gon Choi received B.S. degree in Electronics Engineering from Kyungpook National University in 1990, and M.S. and Ph.D. degree from KAIST in Korea in 1999 and 2004, respectively. He is currently a professor in College of Electrical & Computer Engineering, Chungbuk National University. His research interests include V2X, AI, smart grid, IoT, mobile communication, high-speed network architecture and protocol.

Flexible Localization Method with Motion Estimation for Underwater Wireless Sensor Networks

A. S. Ismail*§, Ammar Hawbani*,**±, Xingfu Wang*±, Samah Abdel Aziz*§, Liang Zhao**, Nasir Saeed***
*School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China
§Faculty of Science, Zagazig University, Zagazig 44519, Egypt
**School of Computer Science, Shenyang Aerospace University, Shenyang 110136, China
***Department of Electrical and Communication Engineering, United Arab Emirates University, Al Ain, 15551, UAE
Corresponding Author ±**anmande@ustc.edu.cn, wangxfu@ustc.edu.cn**
abdo_samy@mail.ustc.edu.cn, samahhabib10@yahoo.com, lzhao@sau.edu.cn, mr.nasir.saeed@ieee.org

Abstract—**Due to the challenging conditions of underwater environments, such as node mobility and large-scale networks, achieving localization in large-scale mobile underwater sensor networks (UWSN) is a difficult task. This paper introduces a scheme known as the Flexible Localization Method with Mobility Estimation (FLMME) for UWSNs by utilizing the expected mobility patterns of underwater objects. FLMME performs localization hierarchically by splitting the process into anchor and ordinary node localization. Each node estimates its next mobility pattern based on previous location information, enabling estimates about its next location. Anchor nodes, holding known locations, manage the localization process to balance accuracy and error trade-offs. Extensive simulations demonstrate that FLMME reduces localization errors and hence increases localization accuracy.**

Keywords— **Underwater Wireless Sensor Networks (UWSN), Localization, Movement Estimation, Tracking, Sensor Deployment**

I. Introduction

Approximately 70% of Earth's surface comprises ocean, river, and sea, that drawing global focus toward harnessing its resources. Underwater wireless sensor networks (UWSNs) have become essential to oceanic advancements. They are self-organizing networks of small-sensor nodes deployed in oceanic monitoring areas, enabling real-time and precise underwater environment monitoring [1]. UWSNs are widely applied in water environment monitoring, disaster prediction, marine navigation, military applications, and seabed exploration. Accurate knowledge of UWSN node locations is crucial for tasks like topology control, coverage, and routing decisions. Hence, UWSN localization research is of essential significance [2]–[4].

Localizing mobile nodes within UWSNs holds significant importance in various domains. In marine military defense, it's vital for identifying and tracking intrusion objects. For animal tracking, it aids in observing marine animals' locations and movements. Mobile node localization extends the monitoring area, proving valuable in marine environmental monitoring by identifying potentially polluted areas and covering extensive regions. Effective path planning and optimizing underwater node coverage depend on localizing mobile nodes, given the significance of location in managing data collection and transmission in sensor networks. As a result, localizing mobile sensor nodes has become a prominent research focus.

Overcoming major challenges in underwater localization, including substantial acoustic channel delays and high signal attenuation in RF/optical channels, is crucial. The primary challenges include (1) Node mobility concerns driven by water currents, turbulence, and winds that displace sensor nodes from their positions. While GPS can locate surface buoy-based anchor nodes, underwater node location measurement remains difficult [5]. (2) Sensor node deployment is critical in network establishment but challenging in harsh underwater conditions. (3) Achieving synchronization without GPS signals poses a challenge, leading to significant localization errors in time-of-arrival-based ranging due to lack of synchronization [6], [7].

In this paper, we introduce an approach for underwater sensor network localization, which we call a Flexible Localization Method with Motion Estimation (FLMME). FLMME employs a hierarchical localization process by dividing it into two stages: localizing anchor nodes and ordinary sensor nodes. Each node in the localization process makes predictions about its next movement patterns based on prior known location data, enabling it to estimate its next location. To balance the compromises between the localization error, coverage, and accuracy, the known location of anchor nodes within the network will manage the localization procedure. Simulation outcomes demonstrate that FLMME significantly reduces localization error while preserving high localization accuracy and coverage.

The rest of the paper is organized as follows: the classification of localization techniques for UWSN is given in Section II, then the network architecture and proposed FLMME algorithms are described in Section III. In Section IV, the simulation results are presented. Finally, the paper is included in Section V.

II. Classification of Localization Techniques for UWSN

Addressing the aforementioned challenges, researchers have developed diverse localization algorithms suitable for UWSNs. These algorithms can be categorized based on various data acquisition and processing approaches. In this section, we divided the localization schemes into five categories: range-based, range-free, network-based topology, centralized, and distributed localization schemes, as seen in Figure 1. Each category will be explained with some examples as follows:

1) **Range-based Localization:** The ranging-based algorithm employs varied techniques, such as Angle of Arrival (AOA), Time of Arrival (TOA), Received Signal Strength Indication (RSSI), and Time Difference of Arrival (TDOA), to measure node distances to anchor nodes and determine unknown node locations [8]–[10]. TOA, reliant on signal transmission time and speed, requires network-wide time synchronization, providing more accuracy in underwater localization due to faster wireless acoustic signal propagation. TDOA calculates node distances based on signal arrival time differences and respective velocities, exhibiting higher accuracy with demanding hardware requirements. RSSI translates signal loss into distance, cost-effectively, without the need for time synchronization, yet environmental variations affect its accuracy. The AOA algorithm mandates node-deployed antennas to gauge received signal angles for node coordination, entailing additional costs and posing scalability challenges [11]–[13].
2) **Range-free Localization:** The range-free algorithms estimate node distances based on hop count and directional data, converting this information into approximate node positions. However, these methods generally lack precise localization and typically offer only approximate node estimations [14]. Examples include DV-Hop, Center-of-mass (CoM), and APIT [15]–[17]. DV-Hop relies on hop count to calculate node distances. It calculates the minimum hops to the anchor node, estimates the average hop distance, and multiplies these to approximate the unknown node's position. CoM requires a node to select several anchors within its communication range, gathering their information. Upon reaching a set threshold, it calculates the center-of-mass coordinates based on the coordinates of the anchor nodes. Although straightforward, the CoM method lacks high localization accuracy.
3) **Network-based Topology Localization:** in which the localization algorithms are categorized into two primary types based on network topology and node connectivity: single-hop and multi-hop. Single-hop localization describes instances where an unknown node communicates directly with an anchor node within a single-hop distance. In contrast, multi-hop localization requires unknown nodes to pass through multiple nodes to reach the anchor node. While single-hop algorithms demand a high density of anchor nodes, resulting in substantial costs, multi-hop algorithms, requiring nodes to relay information through intermediaries, offer a more cost-effective and feasible solution for wider network coverage.

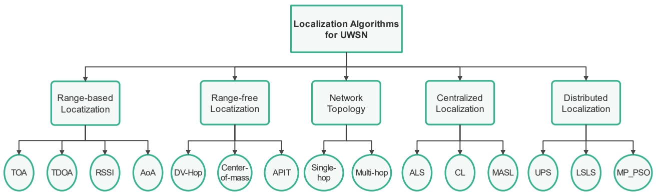


Figure 1: Localization techniques for UWSN

4) **Centralized Localization:** In a centralized localization approach, all nodes transmit collected data to a central sink node for processing. Equipped with significant computational power, the sink node calculates the positions of unknown nodes and relays this information back. However, due to the substantial computational load on the sink node, this approach is best suited for small-scale UWSNs, resulting in high energy consumption. Common methods include Area-Based Localization Scheme (ALS), Collaborative Localization (CL), and Motion-Aware Self-Localization (MASL) [18], [19].
5) **Distributed Localization:** The distributed localization method involves nodes processing their information autonomously and then interchanging this data for localization, making it ideal for large-scale UWSNs. This enables real-time node position retrieval while distributing computational tasks across the network. Well-known distributed localization approaches include the Underwater Positioning Scheme (UPS), Large-Scale Localization Scheme (LSLS), and Localization based on Mobility Prediction and Particle Swarm Optimization (MP-PSO) [20], [21].

III. Description of FLMME Algorithm

First, this section outlines the network architecture of large-scale UWSNs, followed by an overview of the FLMME algorithm, and then the anchor node mobility estimation and the ordinary node localization are described in detail.

A. *Network Architecture*

To perform node localization within large-scale UWSNs, a hierarchical network structure is introduced and illustrated in Figure 2. This architecture involves three node types as follows:

- Surface buoys: These are located on the ocean surface and achieve precise location data through GPS systems. They are responsible for aiding in the localization of anchor nodes.
- Anchor nodes: These are high-energy devices that communicate directly with buoy nodes to determine their positions. They move in concurrence with the water current. These high-energy anchor nodes tend to be costly. So, to

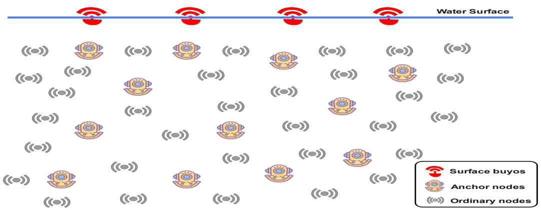


Figure 2: UWSN architecture with different node types

mitigate expenses, this network includes only a limited number of anchor nodes (10).

- Ordinary Sensor nodes: Prior to the localization process, all nodes, excluding buoys and anchor nodes, are categorized as ordinary nodes. Similar to anchor nodes, these ordinary nodes inertly drift with the water current. They are unable to connect with buoys directly and are not all within the coverage of anchor nodes. Only ordinary nodes within the communication range of an anchor node can ascertain their location via these anchors [22].

To summarize, anchor nodes acquire their locations directly through communication with surface buoys, while ordinary nodes self-localize by interacting with neighboring anchors or ordinary nodes. Each sensor node is anticipated to periodically obtain its location, denoted as the localization period T_1. This assumption is realistic in various applications, like environmental monitoring, where sensor nodes periodically transmit observed data to the sink. Location information at these specific time points proves to be particularly relevant.

B. Overview of FLMME

FLMME utilizes a hierarchical localization strategy. This approach separates the overall localization procedure into two parts: (1) localization of anchor node and (2) localization of ordinary node. Initially, a limited number of surface buoys possess their locations through GPS or alternative methods. In this network, at least four buoys are required to enable the localization of anchor nodes. As the anchor nodes possess higher capabilities, they directly determine their positions from the surface buoys within each localization period. This allows the implementation of intricate mobility estimation algorithms on these anchor nodes.

We introduce a distributed range-based approach for the localization of ordinary nodes, detailed in Section III-D. As ordinary nodes have constrained computational power and memory, intricate estimation algorithms are challenging to implement on them. However, owing to the collective movement characteristics of underwater entities, an ordinary node can infer its movement pattern by observing the movement patterns of neighboring nodes. The ordinary node actively listens for localization messages within the network. If it does not receive any messages beyond a predefined duration, it assumes it is out of range, marking itself as unlocalized. Conversely, upon receiving localization messages, the node initiates its localization and movement estimation algorithm to estimate its location and movement pattern. The estimation algorithm will be elaborated in Section III-D.

C. Mobility Estimation of Anchor Nodes

In the beginning, anchor nodes have the advantage of direct communication with surface buoys, enabling them to obtain their actual locations at any localization period. Additionally, these nodes can estimate their next mobility patterns based on previous measurements. In localization period T_1, the anchor node can calculate its actual location as $Anchor_{Actual}(T_1)$, then it can estimate its location as follows:

$$\text{Anchor}_{Estimated}(m) = Anchor_{Actual}(m-1) + T_1 * v(m-1) \quad (1)$$

where $Anchor_{Actual}(m-1)$ is the actual location of the anchor node at the time $(m-1)T_1$ and $v(m-1)$ is the estimated speed of the anchor node in the localization period $(m-1)$.

Then, according to the anchor node's actual and estimated locations, the localization error can be obtained as follows:

$$\text{Anchor}_{Error}(m) = \text{Anchor}_{Estimated}(m) - Anchor_{Actual}(m) \quad (2)$$

where $Anchor_{Actual}(m)$ is the actual location of anchor node at mT_1. After that, the anchor node assists in the localization of ordinary nodes by transferring its mobility information through synchronous transmission.

D. Localization of Ordinary Nodes

To estimate the accurate next location of ordinary nodes, we need the location and the movement speed of the anchor nodes as reference nodes, in addition to the previous locations of the ordinary nodes. In contrast to anchor nodes, the prior location of ordinary nodes is unidentified. Consequently, the initial location of each ordinary node i needs to be determined. To calculate the initial location of ordinary nodes, we need to calculate the distance of the ordinary node from three anchor nodes at time t as follows:

$$d_1 = \sqrt{(X_1 - x_i)^2 + (Y_1 - y_i)^2 + (Z_1 - z_i)^2} \quad (3)$$

$$d_2 = \sqrt{(X_2 - x_i)^2 + (Y_2 - y_i)^2 + (Z_2 - z_i)^2} \quad (4)$$

$$d_3 = \sqrt{(X_3 - x_i)^2 + (Y_3 - y_i)^2 + (Z_3 - z_i)^2} \quad (5)$$

where (X_1, Y_1, Z_1), (X_2, Y_2, Z_2), (X_3, Y_3, Z_3) are the anchor nodes real-time location that can be determined through surface buoys, and (x_i, y_i, z_i) is the location of ordinary node at time t. Then, by solving those equations to obtain the current location of the ordinary node. Then, we calculate the movement speed of the ordinary node by using the previous location $(x_{i-1}, y_{i-1}, z_{i-1})$ with the current location (x_i, y_i, z_i) as follows:

$$v = \frac{\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2}}{t - (t-1)} \quad (6)$$

According to the mobility of the ordinary node, then the estimated location of the ordinary node can be calculated as follows:

$$Ordinary_{Est}(m) = Ordinary_{Act}(m-1) + T_1 * v(m-1) \quad (7)$$

where $Ordinary_{Est}(m)$ is the estimated location of ordinary node at the time mT_1, $Ordinary_{Act}(m-1)$ is the location of the ordinary node at time $(m-1)T_1$, and $v(m-1)$ is the movement speed at time $(m-1)T_1$.

IV. Simulation and Results

This section conducts MATLAB simulations to validate the efficacy of the FLMME algorithm. Table I illustrates the additional parameters utilized in the simulation experiments.

In the subsequent simulation, a total of 100 ordinary nodes and 10 anchor nodes are deployed randomly in a 3D monitoring area of 500m * 500m * 500m. Figure 3 shows the deployment of the anchor nodes and the ordinary nodes at the beginning of the simulation. Node density is considered as the estimated number of sensor nodes within a node's proximity. As a result, the degree of each sensor node and its density are the same. To manage node density, we alter each sensor node's communication range while keeping a similar deployment area. We assume that the estimated distances between sensor nodes conform to normal distributions. This assumption is realistic and aligns with present technologies for measuring underwater distances.

To validate the localization accuracy of the FLMME algorithm, the evaluation metric employed is the localization error (LE), which is defined as the mean distance between the estimated locations and the actual locations of all nodes. The localization error of an individual node is expressed as follows:

$$\text{LE} = \sqrt{(X_{Act} - X_{Est})^2 + (Y_{Act} - Y_{Est})^2 + (Z_{Act} - Z_{Est})^2} \quad (8)$$

where $(X_{Act}, Y_{Act}, Z_{Act})$ is the actual location and $(X_{Est}, Y_{Est}, Z_{Est})$ is the estimated location.

Therefore, in UWSN, the average localization error (ALE) for the N localized ordinary nodes is given by:

$$\text{ALE} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(X(i)_{Act} - X(i)_{Est})^2 + (Y(i)_{Act} - Y(i)_{Est})^2 + (Z(i)_{Act} - Z(i)_{Est})^2} \quad (9)$$

TABLE I: SIMULATION PARAMTERS

Simulation Parameter	Value
Test area	500 m * 500 m * 500 m
Ordinary nodes	100 to 250
Anchor node number	10
Localization period	1 s
Communication range of anchor nodes	200 m
Communication range of ordinary nodes	100 m
Acoustic wave speed	1500 m/s

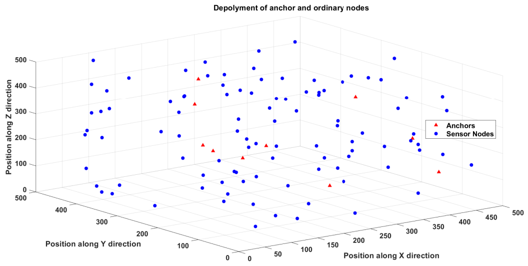


Figure 3: Deployment of 10 anchors and 100 ordinary nodes

A. *Impact of Node Density on Localization Results*

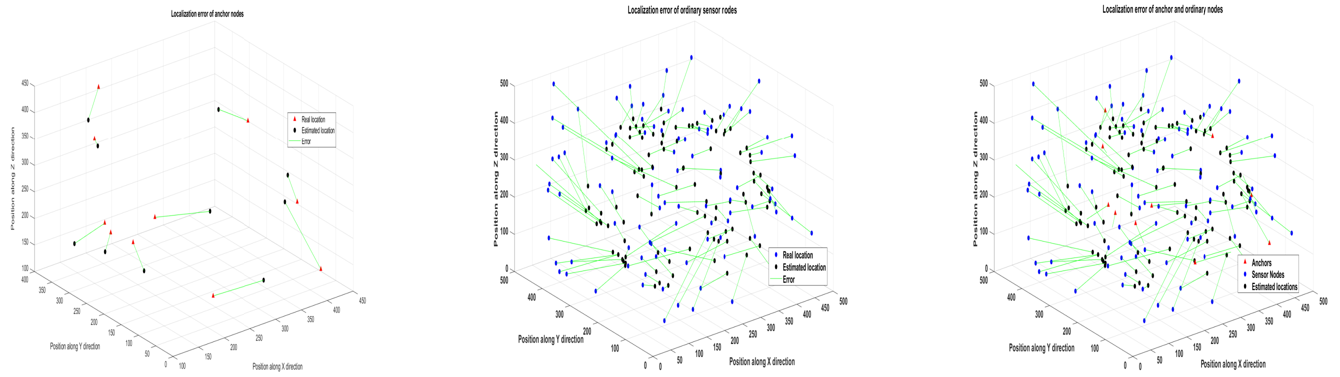
In this study, we analyze the impact of the node density (increasing the number of nodes) on the localization performance. The localization results of moving sensor nodes due to the water currents are shown in Figures 4 to 7. Hence, increasing the node density leads to decreasing the average localization error and, therefore, increasing the localization accuracy of our method.

Figure 4 presents the simulation for 100 ordinary nodes and 10 anchor nodes. Figure 4a presents the movement of the anchor nodes, and we utilized the range-based algorithm to estimate the next location of anchor nodes and then calculated the average localization error between the actual and the estimated locations and found that the accuracy of the anchor nodes is good which helping in increasing the localization accuracy of the ordinary nodes. Figure 4b shows the relation between the estimated and the actual locations of the ordinary nodes with an average localization error of 107.90 m. While Figure 4c presents the localization results for all anchor and ordinary nodes.

Figure 5 presents the simulation for 150 ordinary nodes and 10 anchor nodes. Figure 5a presents the movement of the anchor nodes, and we utilized the range-based algorithm to estimate the next location of anchor nodes and then calculated the average localization error between the actual and the estimated locations and found that the accuracy of the anchor nodes is good which helping in increasing the localization accuracy of the ordinary nodes. Figure 5b shows the relation between the estimated and the actual locations of the ordinary nodes with an average localization error of 97.86 m. While Figure 5c presents the localization results for all anchor and ordinary nodes.

Figure 6 presents the simulation for 200 ordinary nodes and 10 anchor nodes. Figure 6a presents the movement of the anchor nodes, and we utilized the range-based algorithm to estimate the next location of anchor nodes and then calculated the average localization error between the actual and the estimated locations and found that the accuracy of the anchor nodes is good which helping in increasing the localization accuracy of the ordinary nodes. Figure 6b shows the relation between the estimated and the actual locations of the ordinary nodes with an average localization error of 84.60 m. While Figure 6c presents the localization results for all anchor and ordinary nodes.

Figure 7 presents the simulation for 250 ordinary nodes and 10 anchor nodes. Figure 7a presents the movement of the anchor nodes, and we utilized the range-based algorithm to estimate the next location of anchor nodes and then calculated the average localization error between the actual and the estimated locations and found that the accuracy of the anchor nodes is good which helping in increasing the localization accuracy of the ordinary nodes. Figure 7b shows the relation between the estimated and the actual locations of the ordinary nodes with an average localization error of 69.43 m. While Figure 7c presents the localization results for all anchor and ordinary nodes.

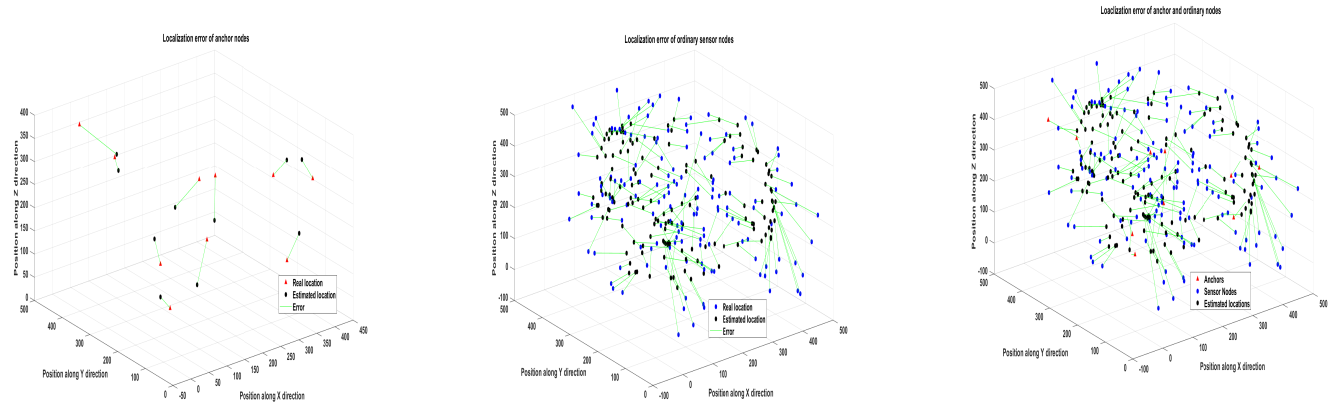


(a) Localization result of anchor nodes

(b) Localization result of ordinary nodes

(c) Localization result of anchor and ordinary nodes

Figure 4: Localization results when the number of nodes=100, anchor nodes=10, and the localization error is 107.90 m

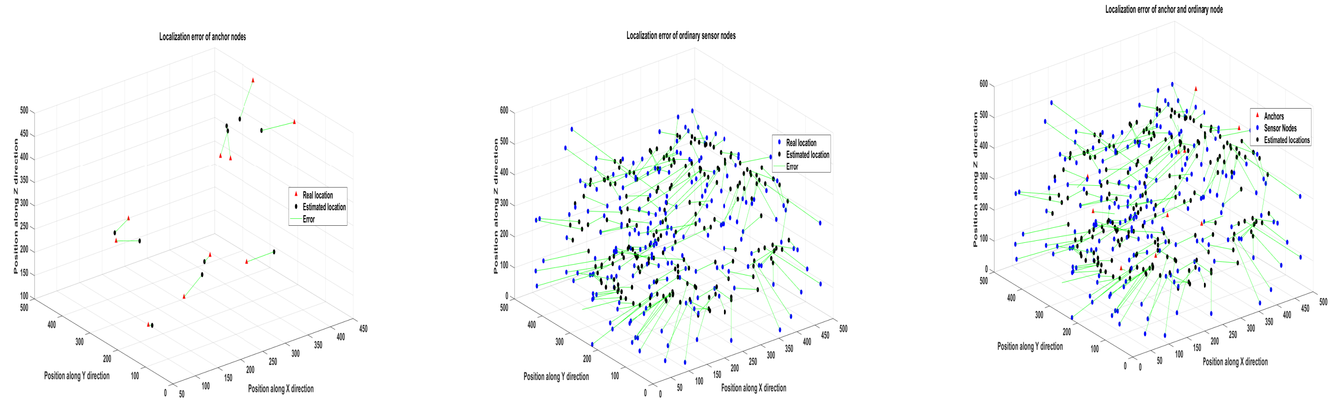


(a) Localization result of anchor nodes

(b) Localization result of ordinary nodes

(c) Localization result of anchor and ordinary nodes

Figure 5: Localization results when the number of nodes= 150, anchor nodes= 10, and the localization error is 97.86 m



(a) Localization result of anchor nodes

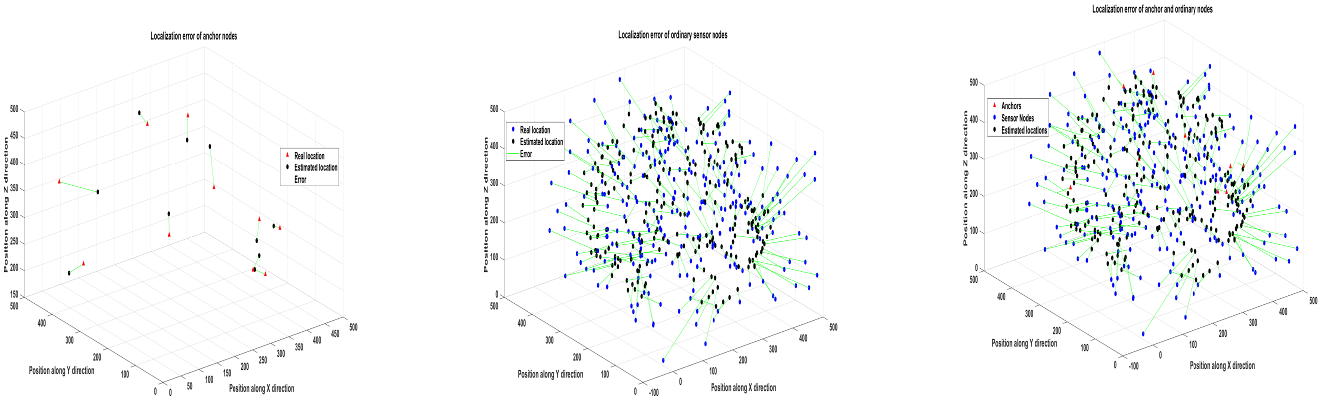
(b) Localization result of ordinary nodes

(c) Localization result of anchor and ordinary nodes

Figure 6: Localization results when the number of nodes= 200, anchor nodes= 10, and the localization error is 84.60 m

V. Conclusions

This paper presented FLMME, a localization scheme for UWSNs. FLMME utilized the mobility patterns in underwater environments and employed a hierarchical approach. FLMME divided the localization process into anchor and ordinary node



(a) Localization result of anchor nodes

(b) Localization result of ordinary nodes

(c) Localization result of anchor and ordinary nodes

Figure 7: Localization results when the number of nodes= 250, anchor nodes= 10, and the localization error is 69.43 m

localization, enabling each node to estimate its next mobility based on previous location data. Anchor nodes, with recognized locations, guide the process to balance accuracy and potential errors. Extensive simulations demonstrated FLMME's ability to significantly reduce localization errors, enhancing the overall accuracy of the localization process.

Acknowledgment

This paper is supported by the Innovation Team and Talents Cultivation Program of the National Administration of Traditional Chinese Medicine (No. ZYYCXTD-D-202208). A. S. Ismail is grateful for financial support from the CAS-TWAS Fellowship.

References

[1] Z. Zeng, S. Fu, H. Zhang, Y. Dong, and J. Cheng, "A survey of underwater optical wireless communications," *IEEE communications surveys & tutorials*, vol. 19, no. 1, pp. 204–238, 2016.

[2] H. Luo, K. Wu, R. Ruby, Y. Liang, Z. Guo, and L. M. Ni, "Software-defined architectures and technologies for underwater wireless sensor networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2855–2888, 2018.

[3] Y. Zhou, H. Yang, Y.-H. Hu, and S.-Y. Kung, "Cross-layer network lifetime maximization in underwater wireless sensor networks," *IEEE Systems Journal*, vol. 14, no. 1, pp. 220–231, 2019.

[4] R. Su, D. Zhang, C. Li, Z. Gong, R. Venkatesan, and F. Jiang, "Localization and data collection in auv-aided underwater sensor networks: Challenges and opportunities," *IEEE Network*, vol. 33, no. 6, pp. 86–93, 2019.

[5] A. Ismail, X. Wang, A. Hawbani, S. Alsamhi, and S. Abdel Aziz, "Routing protocols classification for underwater wireless sensor networks based on localization and mobility," *Wireless Networks*, vol. 28, no. 2, pp. 797–826, 2022.

[6] N. Saeed, A. Celik, T. Y. Al-Naffouri, and M.-S. Alouini, "Underwater optical wireless communications, networking, and localization: A survey," *Ad Hoc Networks*, vol. 94, p. 101935, 2019.

[7] W. Pu, "A survey of localization techniques for underwater wireless sensor networks," *Journal of Computing and Electronic Information Management*, vol. 11, no. 1, pp. 10–15, 2023.

[8] X. Su, I. Ullah, X. Liu, D. Choi *et al.*, "A review of underwater localization techniques, algorithms, and challenges," *Journal of Sensors*, vol. 2020, 2020.

[9] S. Xu, Y. Ou, and X. Wu, "Optimal sensor placement for 3-d time-of-arrival target localization," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 5018–5031, 2019.

[10] J. Fang and Z. He, "Robust modified newton algorithms using tikhonov regularization for tdoa source localization," *Circuits, Systems, and Signal Processing*, vol. 38, no. 11, pp. 5342–5359, 2019.

[11] L. Zhiyu, B. Bin, W. Jianhui, L. Wenchao, and W. Daming, "A direct position determination method with combined tdoa and fdoa based on particle filter," *Chinese Journal of Aeronautics*, vol. 31, no. 1, pp. 161–168, 2018.

[12] P. Marzioli, L. Frezza, F. Curiano, A. Pellegrino, A. Gianfermo, F. Angeletti, L. Arena, T. Cardona, M. Valdatta, F. Santoni *et al.*, "Experimental validation of vor (vhf omni range) navigation system for stratospheric flight," *Acta Astronautica*, vol. 178, pp. 423–431, 2021.

[13] X. Ding and S. Dong, "Improving positioning algorithm based on rssi," *Wireless Personal Communications*, vol. 110, pp. 1947–1961, 2020.

[14] I. Nemer, T. Sheltami, E. Shakshuki, A. A. Elkhail, and M. Adam, "Performance evaluation of range-free localization algorithms for wireless sensor networks," *Personal and Ubiquitous Computing*, vol. 25, pp. 177–203, 2021.

[15] V. Kanwar and A. Kumar, "Dv-hop localization methods for displaced sensor nodes in wireless sensor network using pso," *Wireless Networks*, vol. 27, pp. 91–102, 2021.

[16] Y. Shi, H. Liu, W. Zhang, Y. Wei, and J. Dong, "Research on three-dimensional localization algorithm for wsn based on rssi," in *Cyber Security Intelligence and Analytics*. Springer, 2020, pp. 1048–1055.

[17] Z. Aiqing, Y. Xinrong, and H. Haifeng, "Point in triangle testing based trilateration localization algorithm in wireless sensor networks [j]," *Internet and Information Systems*, vol. 6, pp. 2567–2586, 2012.

[18] V. Chandrasekhar and W. Seah, "An area localization scheme for underwater sensor networks," in *OCEANS 2006-Asia Pacific*. IEEE, 2006, pp. 1–8.

[19] D. Mirza and C. Schurgers, "Collaborative localization for fleets of underwater drifters," in *OCEANS 2007*. IEEE, 2007, pp. 1–6.

[20] X. Cheng, H. Shu, Q. Liang, and D. H.-C. Du, "Silent positioning in underwater acoustic sensor networks," *IEEE Transactions on vehicular technology*, vol. 57, no. 3, pp. 1756–1766, 2008.

[21] Y. Zhang, J. Liang, S. Jiang, and W. Chen, "A localization method for underwater wireless sensor networks based on mobility prediction and particle swarm optimization algorithms," *Sensors*, vol. 16, no. 2, p. 212, 2016.

[22] M. Dong, H. Li, R. Yin, Y. Qin, and Y. Hu, "Scalable asynchronous localization algorithm with mobility prediction for underwater wireless sensor networks," *Chaos, Solitons & Fractals*, vol. 143, p. 110588, 2021.

A Reliable Routing Method for Remote Entanglement Distribution under Limited Resources

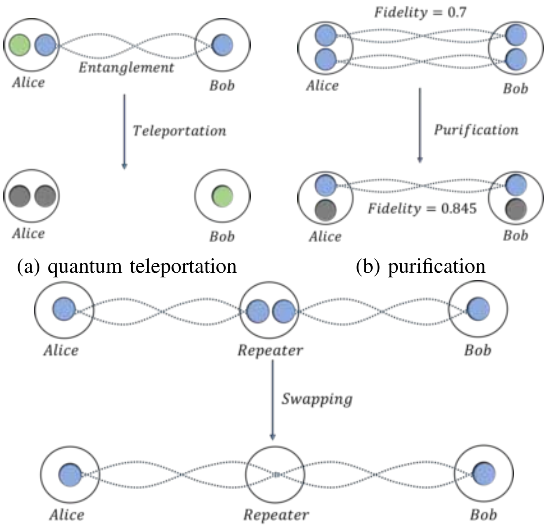
Tianzhu Hu[*], Xiaofeng Jiang[**], Tianze Zhu[**], Xin Sun[**], Haomin Chen[**], and Jian Yang[**]
[*]Institute of Advanced Technology, University of Science and Technology of China, Hefei, China
[**]the Department of Automation, University of Science and Technology of China, Hefei, China
htzustc@mail.ustc.edu.cn, jxf@ustc.edu.cn, tz_19@mail.ustc.edu.cn, sx21010097@mail.ustc.edu.cn, chen0215@mail.ustc.edu.cn, jianyang@ustc.edu.cn

***Abstract*—Generating and distributing entangled pairs between arbitrary nodes is essential to fully realize the network's capabilities, with the challenges of limited qubit resources, severe decoherence and stochastic physical mechanism. In this paper, we model the service process of quantum repeater nodes based on the concept of queuing theory to help characterize their availability. Further, we propose a link-disjoint multi-path routing algorithm, with repeaters' availability, nodes' qubit capacity, entangled links' fidelity and classical delay taken into consideration. The performance of our scheme has been evaluated with simulated environment and compared with other existing routing schemes.**

***Keywords*—Quantum Networks, Entanglement Distribution, Queuing Theory, Routing Algorithm, Quantum Repeaters**

I. Introduction

Quantum communication technology has unveiled the possibility of numerous ground-breaking applications, including highly secure communication, distributed quantum computing, and remote quantum clock synchronization [1]. The realization of these capabilities relies on the utilization of quantum entanglements and their unique features. Entanglement is a phenomenon in quantum mechanics where two or more particles become correlated as a indivisible entity, regardless of how distant they are physically separated. Entangled qubits are well suited for tasks that require coordination, since any alteration made to one particle's state will be instantly reflected in the state of other particles. Also, the laws of quantum mechanics ensure that entanglements are inherently private and hence resilient to eavesdropping from a third party. Generally, the word "entanglements" in this article refers to the Bell states $|\Phi^{\pm}\rangle = \frac{\sqrt{2}}{2}(|00\rangle \pm |11\rangle)$ and $|\Psi^{\pm}\rangle = \frac{\sqrt{2}}{2}(|01\rangle \pm |10\rangle)$. So far, short-lived entanglement distribution between adjacent nodes has been achieved on the ground by sending photons over standard optical fibre [2], or from free-space based on a satellite [3]. Long-lived, remote entanglement between two quantum nodes are generated step-by-step: link-level heralded entanglements are established between adjacent nodes first, followed by entanglement swapping operations performed at intermediate nodes known as quantum repeater nodes. Additionally, entanglement purification allows for higher fidelity with consumption of entangled pairs of lower fidelity, as shown in Fig.1.



(a) quantum teleportation (b) purification

(c) entanglement swapping

Fig. 1: Common applications of entanglements

Quantum memory plays a significant role in the entanglement distribution process, with the challenge of balancing between capability, lifetime, and store-and-retrieve fidelity. Limited resources may lead to severe contention under the multi-request scenario, calling for an effective routing design. At specific moment, it's possible that most nodes lack enough resources for a newly arrived request. Furthermore, the stochastic physical mechanisms underlying the entanglement distribution will undoubtedly lead to a decrease in the entanglement generation success rate.

Recently, huge efforts have been devoted to the development of efficient entanglement routing protocols, aiming to obtain higher throughput with finite qubit resources [4]. Reference [5] focused on capacity allocation on the edges of a quantum network for multiple requests of entanglement generation. Shouqian Shi *et al.* discussed a four-phase routing realization based on the "advance generation" model [6]. Jian Li *et al.* proposed a method to find the optimal purification decisions along the routing path with minimum entangled pair cost [7]. Yiming Zeng *et al.* studied the problem to simultaneously maximize the number of quantum-user pairs and their expected throughput [8]. Laszlo Gyongyosi *et al.* introduced the service rate of quantum repeater nodes based on G/G/1 priority queuing model and existing entangled pairs of different levels

[9].

Almost all papers mentioned above introduced phases in the entanglement distribution process, and primarily focused on link status. Therefore, these studies will fail to avoid the most congested nodes under continuous network environment, leading to unacceptable response time for individual requests and unbalanced resource utilization of the entire network. Compared to them, we explore a solution for reliable entanglement distribution under limited node resources. In this paper, we model the service process of quantum repeater nodes which support link-level entanglement establishment, multi-round purification and entanglement swapping, and derive their service availability based on the concept of queuing theory. Taking the specific queuing process of each repeater node into account, the routing algorithm can choose nodes with less waiting time to mitigate the impact of limited resources, which will help to provide a bandwidth guarantee for every request. A link-disjoint multi-path routing algorithm is proposed accordingly in the prevention of potential failure or memory shortage, with repeaters' availability, links' fidelity and classical delay taken into consideration. We use NetSquid simulation platform [10] to analyze the performance of our scheme, including end-to-end entanglement fidelity and throughput of requests.

II. Network Model

In this section, we will first clarify the key components of a typical quantum network, provide individual descriptions of their essential capabilities separately, and then supplement additional assumptions.

A. Network Components

End nodes. Analogous to end users in classical networks, end nodes in quantum networks communicate and run quantum applications. Each end node is equipped with a quantum processor which has a certain number of memory qubits and necessary hardware to perform basic quantum gates on the qubits, enabling computation, storage, measurement and other functions. We assume that all memory qubits within an end node can be manipulated in parallel, allowing the end node to initiate multiple entanglement attempts with its neighbour simultaneously with the help of multiplexing.

Repeaters. Since the quantum channel built by optical fibre is inherently imperfect, the success rate of entanglement establishments decays exponentially with the channel length. Through entanglement swapping, quantum repeaters serve as intermediate nodes to support entanglement distribution between remote nodes. Here we assume that repeaters are also capable of link-level entanglement purification to improve entanglement fidelity, which means all necessary purification is performed before the swapping process. Additionally, we assume each quantum node has the full functionality of both an end node and a repeater. Within a specific entanglement distribution scenario, these two concepts are employed to differentiate the roles of quantum nodes along the path.

Quantum channels. As discussed above, quantum channels connect adjacent nodes through optical fibres. Under the condition that both nodes have sufficient qubit storage, multiple entanglement attempts can be made at the same time under our assumptions.

B. Network topology

For a quantum network $G = (V, E, C)$ of V quantum nodes and E edges, C_i is defined as the number of memory qubits able to be manipulated in a node v_i. We assume the fidelity of entanglements generated on each edge to follow a normal distribution $N(\mu, \sigma^2)$.

III. Availability of Quantum Repeater

With the assumption that each qubit is able to build entangled pairs with other nodes independently, an arbitrary quantum repeater node $v_i \in V$ with qubit capacity C_i is regarded as a $G/G/C_i$ queuing system and works in an FCFS manner.

Our entanglement distribution scheme utilizes the "on-demand generation" model, in which the entangled pairs are distributed on demand, along the selected path(s) to save qubit resources and avoid decoherence in quantum memories. A remote entanglement generation request will experience path selection and resource allocation phase in order, finally be converted to requests on the source node, the destination node and several repeater nodes. Different from customers in a general queuing system, the service process of a request r on node v_i is described from two dimensions: resource demand, and service time. The qubit resource demand n_r includes basic bandwidth requirements for link-level entanglement establishment with both the previous-hop node v_f and the next-hop node v_l, and potential extra qubits consumed in entanglement purification to meet fidelity requirements. The service time can be divided into three stages in order according to the service process, as shown in Fig.2:

Entanglement establishment: Due to the existence of quantum link layer protocol [11] providing robust link layer entanglement service, the average service time of this stage is calculated by $\max\{t_{fe}, t_{le}\}$, where t_fe, t_le denote the time consumed with v_f and v_l separately.

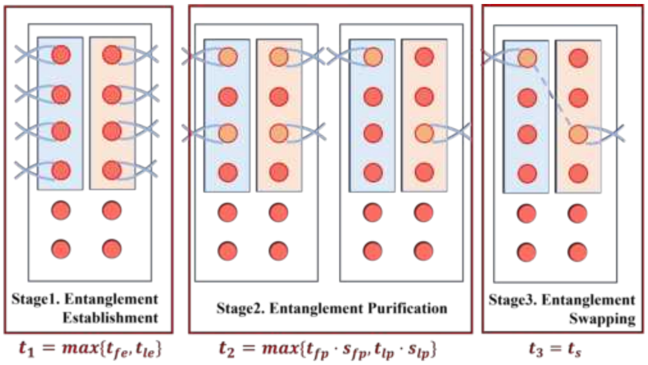

Stage1. Entanglement Establishment
Stage2. Entanglement Purification
Stage3. Entanglement Swapping
$t_1 = max\{t_{fe}, t_{le}\}$
$t_2 = max\{t_{fp} \cdot s_{fp}, t_{lp} \cdot s_{lp}\}$
$t_3 = t_s$

Fig. 2: Service process of a request in repeater v_i. Here $n_r = 8, s_{fp} = s_{lp} = 2$.

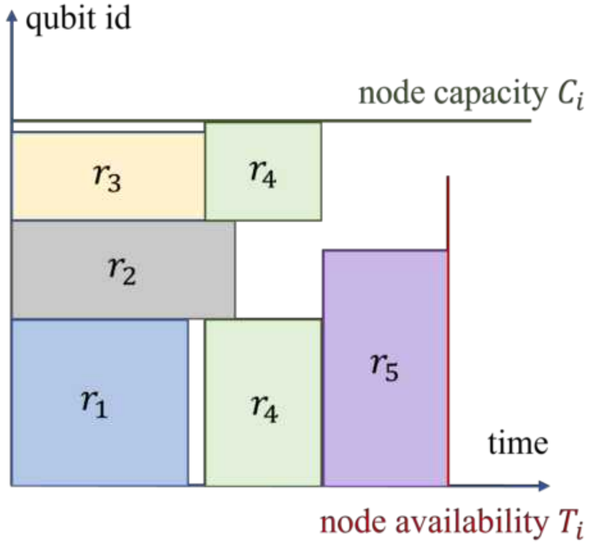


Fig. 3: Example of availability estimation in repeater v_i. Here, $R = \{r_1, r_2, r_3, r_4, r_5\}$, each request is represented as a block, whose length and width denote the resource demand and service time separately.

Entanglement purification: The time purification stage required depends on single-round purification time t_{fp}, t_{lp} in the node, and the number of rounds s_{fp}, s_{lp} with v_f and v_l.

Entanglement swapping: At last, all entanglement swapping processes take constant time t_s in node v_i.

Combining the time spent on different phases above together, request r will occupy n_r qubits for a period of

$$t_r = \max\{t_{fe}, t_{le}\} + \max\{t_{fp}s_{fp}, t_{lp}s_{lp} + t_s\}$$

.

Since both the resource demand and the service time vary from specific request r, the service rate is inaccurate in representing a repeater node's service capability. Instead, we use the whole waiting time of current requests in a node to characterize its availability. To estimate the queuing time of a newly arrived request u when the sorted list of tasks in the node v_i is $R = \{1, \ldots, r, \ldots, u-1\}$, a recursive algorithm is utilized. The departure time of requests being serviced currently can be sequentially predicted using their estimated service time, revealing the next task to be completed and the number of qubits it will release. Afterwards, the moment when the next task to be serviced is able to enter the system is determined according to its resource demand, and then its departure time is confirmed by service time. This process is repeated until the completion time of all tasks is obtained, the latest time among which is the availability attribute T_i of node v_i. Fig.3 illustrates how this algorithm works.

IV. Entanglement Routing Algorithm

Introducing multi-path methods in the entanglement distribution routing process can reduce the storage capacity requirements for individual quantum repeater nodes. Also, it allows for a more efficient utilization of resources from different nodes to handle probabilistic failures that may occur on some certain paths, improving the overall robustness. In this paper, we adopt a link-disjoint multi-path algorithm Q-PSBA(Path Selection Based on Availability), to avoid resource contention on a single link caused by different paths of the same request. It is achieved by an extended multi-round Dijkstra algorithm, where the edges associated with the optimal path found in the previous round are deleted between rounds. An example of link-disjoint path exploration is shown in Fig.4.

Compared to the original Dijkstra algorithm, we introduced innovative routing metrics with nodes' resource availability and links' entanglement quality along the path included. To maximally loosen the requirement for qubit coherence time, an end-to-end entanglement generation will be started when all nodes on the path are prepared. Therefore, resource availability is depicted by the minimum node qubit capacity C_i and the maximum node availability T_i among the repeater nodes of a path. The entanglement quality is represented by both the product $\prod_{i=1}^{n} F_i$ of fidelity, and the sum $\sum_{i=1}^{n} D_i$ of classical communication delay on each edge of the path. The composite metric K is defined as:

$$K = k_1 T_i / C_i - k_2 \log \prod_{i=1}^{n} F_i + k_3 \sum_{i=1}^{n} D_i$$

, where $k1, k2, k3 \geq 0$ are parameters to adjust the routing strategy.

V. Performance Evaluation

In this section, we evaluated the performance of the simulation outcomes from multiple perspectives compared to another link-status-based Dijkstra routing algorithm Q-CAST [6] using metrics of request throughput and average fidelity. The simulation platform is based on NetSquid, a discrete event simulator for quantum networks. In addition to routing algorithms, several protocols were deployed on each node to simulate qubit resource management and quantum operations.

Network parameters: We constructed a 10×10 grid network that contains 100 quantum nodes, and set the number of quantum memory positions in each node to 100(i.e. ignore the influence of node qubit capacity C_i). The entanglement establishment between two neighbor nodes is regarded as a robust service with a 100% success rate, and an average completion time of 10ms. The success probability of imperfect entanglement swapping operations is set to 0.9 in all nodes.

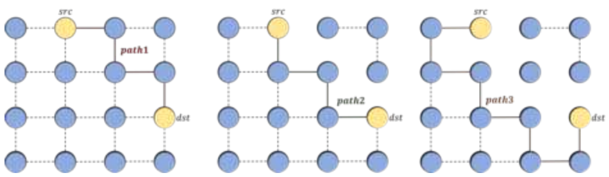


Fig. 4: A three-round routing example on a 4×4 grid network, with network topology updated between rounds.

The average operation time of entanglement purification and swapping is set to 2ms and 10ms, separately. Without loss of generality, we also set some random parameters in the network. First, we randomly select an initial fidelity from a Gaussian distribution $N = (0.8, 0.001)$ for each edge and assume the fidelity of multiple entangled pairs generated along the same edge to be identical. Then we defined a random request generator, which simulates a request flow following a Poisson distribution with a mean inter-arrival time of 50ms. Each request in the flow has random, non-adjacent source and destination nodes. Except for entanglement establishment and entanglement swapping, necessary entanglement purification on a single link is engaged based on the end-to-end fidelity calculation model in [12] to ensure that the final entanglement's fidelity is above 0.5.

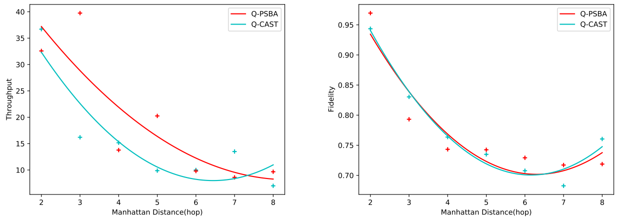
Performance with end-to-end distance: For the multi-path routing algorithm Q-PSBA, the S-D pair distance is defined as the Manhattan distance between the source node and the destination node. Under the experimental setup described above, 100 requests were generated randomly and finished with paths selected by Q-PSBA and Q-CAST. All available qubits are utilized in nodes on each path, and the fidelity of every successfully constructed target entanglement is measured. Then we group the requests by their S-D pair distance to calculate the average performance, as shown in Fig.5. The results imply that Q-PSBA has an advantage of about 5 entangled pairs in terms of throughput compared to Q-CAST, while the fidelity performance is similar.

Stability with time: To study the sustainability of our method, we randomly generate 25 rounds of requests with S-D pair distance of 5 hops, with 4 requests in each round. The performance of requests with different sequences (i.e. arrival time) is shown in Fig.6. We found that the difference in average fidelity and throughput between requests with different sequences is within 0.04 and 3, thus the stability with time indicated.

VI. Conclusion

In this paper, we have proposed a reliable routing protocol for entanglement distribution in quantum networks to deal with limited resources and existing multiple requests. We have modelled the service process of quantum repeaters in both qubit usage and service time dimensions, which helps to characterize the node's service availability. A flexible multi-path routing scheme has been developed, with qubit capacity, service availability, link fidelity and classical channel delay considered. Our algorithm has been evaluated on a grid network of small scale and proved to provide higher throughput and better fidelity for each request with time stability. To strengthen our theory, more experiments based on other topology network structures remain to be conducted in both simulation and physical environments.

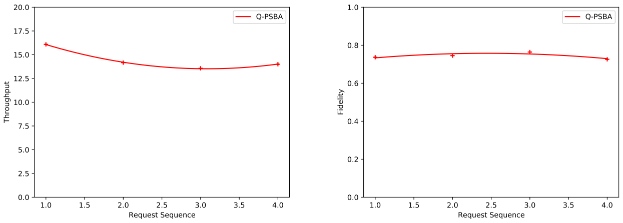
To our best knowledge, we have proposed the first multi-request routing scheme without the setting of processing time windows. In future work, it would be interesting to include historical request information in the repeater node's service model, and an optimal resource allocation algorithm to meet the bandwidth requirements of requests is desired.



(a) Throughput

(b) Fidelity

Fig. 5: Performance with different S-D pair distance



(a) Throughput

(b) Fidelity

Fig. 6: Performance with different request sequence

Acknowledgment

The corresponding author Xiaofeng Jiang would like to acknowledge the financial support of National Natural Science Foundations of China (Grant No. 62173315, 62101525, 62021001), Youth Innovation Promotion Association CAS (Grant No. 2020450), Strategic Priority Research Program of CAS (Grant No. XDC07020200), Fundamental Research Funds for the Central Universities, China Environment for Network Innovations (CENI).

References

[1] S. Wehner, D. Elkouss, and R. Hanson, "Quantum internet: A vision for the road ahead," *Science*, vol. 362, no. 6412, 2018. doi:10.1126/science.aam9288.

[2] J. F. Dynes *et al.*, "Efficient entanglement distribution over 200 kilometers," *Optics Express*, vol. 17, no. 14, p. 11440, 2009. doi:10.1364/oe.17.011440

[3] J. Yin *et al.*, "Satellite-based entanglement distribution over 1200 kilometers," *Science*, vol. 356, no. 6343, pp. 1140–1144, 2017. doi:10.1126/science.aan3211

[4] F. Dupuy, C. Goursaud, and F. Guillemin, "A survey of quantum entanglement routing protocols—challenges for wide-area networks," *Advanced Quantum Technologies*, vol. 6, no. 5, 2023. doi:10.1002/qute.202200180

[5] C. Li, T. Li, Y.-X. Liu, and P. Cappellaro, "Effective routing design for remote entanglement generation on Quantum Networks," *npj Quantum Information*, vol. 7, no. 1, 2021. doi:10.1038/s41534-020-00344-4.

[6] S. Shi and C. Qian, "Concurrent entanglement routing for Quantum Networks," *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, 2020. doi:10.1145/3387514.3405853.

[7] J. Li *et al.*, “Fidelity-guaranteed entanglement routing in Quantum Networks,” *IEEE Transactions on Communications*, vol. 70, no. 10, pp. 6748–6763, 2022. doi:10.1109/tcomm.2022.3200115.

[8] Y. Zeng, J. Zhang, J. Liu, Z. Liu, and Y. Yang, “Multi-entanglement routing design over Quantum Networks,” *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 2022. doi:10.1109/infocom48880.2022.9796810

[9] L. Gyongyosi and S. Imre, “Routing space exploration for scalable routing in the Quantum internet,” *Scientific Reports*, vol. 10, no. 1, 2020. doi:10.1038/s41598-020-68354-y

[10] T. Coopmans *et al.*, “NetSquid, a network simulator for quantum information using discrete events,” *Communications Physics*, vol. 4, no. 1, 2021. doi:10.1038/s42005-021-00647-8

[11] A. Dahlberg *et al.*, “A link layer protocol for Quantum Networks,” *Proceedings of the ACM Special Interest Group on Data Communication*, 2019. doi:10.1145/3341302.3342070.

[12] H.-J. Briegel, W. Dür, J. I. Cirac, and P. Zoller, “Quantum repeaters: The role of imperfect local operations in quantum communication,” *Physical Review Letters*, vol. 81, no. 26, pp. 5932–5935, 1998. doi:10.1103/physrevlett.81.5932

Energy Efficiency Analysis of novel Index Modulation-based Non-Orthogonal Multiple Access (IMNOMA) system for 5G Networks

Shwetha H M
Department of Electronics and Communication Engineering
National Institute of Technology,Warangal
Telangana, India
shwethahmutt@gmail.com

Anuradha S
Department of Electronics and Communication Engineering
National Institute of Technology,Warangal
Telangana, India

***Abstract*—: The requirements for high data rates, spectrum efficiency, energy efficiency, and worldwide connectivity should be addressed by future generations of wireless communication networks. The orthogonal multiple access systems cannot meet these demands. Resources are allocated to each user orthogonally in OMA techniques, including TDMA, FDMA, CDMA, and OFDMA. In a NOMA system, all users can concurrently share resources that are available to them. NOMA is the novel multiple access technology as it satisfies the requirements for high data speeds, spectrum efficiency, widespread connectivity. Users are informed in Index modulation utilising both constellation symbols as well as index symbols. This work proposes IM-NOMA, a unique Index Modulation-based NOMA system. A detailed analysis of the proposed scheme's energy efficiency is performed. Energy efficiency of the proposed scheme is analyzed in detail for different parameters. When compared to the existing modulation schemes, the suggested IM-NOMA performs significantly better.**

***Index Terms*—: Index modulation, Non-orthogonal Multiple Access, Energy efficiency, Spectral efficiency.**

I. Introduction

Due to the advancement of wireless technology, massive connectivity and data rates are in high demand. OMA approach includes the multiple access methods used in 1G, 2G, and 3G that use frequency, time, and code, as well as OFDMA in 4G. Resources such as frequency, time, code are orthogonally allotted to each user in OMA [1].

The users supported by OMA technology depend on the resources availability. High spectral efficiency, massive connection, and demand for data rates have all increased as a result of the development of 5G. OMA-based networks are unable to satisfy these needs since the number of concurrent users a network can handle greatly depends on the availability of orthogonal resource blocks. NOMA supports high spectral efficiency, massive connectivity because users can use the same frequency and temporal resources as traditional OMA techniques. Consider NOMA technology instead of OMA technology to dramatically increase 5G network capacity. Fig.1 provides a graphic comparison between NOMA and OMA. [2] Code-domain and Power-domain are the two main NOMA subcategories. NOMA multiplexing uses power domains and code domains to allow more users to share the limited resources available, boosting capacity of 5G networks. Users in PD-NOMA are given varying levels of power so that they can all access the same resource blocks for time, frequency, and code. [3] A unique code is issued to each user in CD-NOMA so they can share the same resource blocks. Superposition coding (SC) at the transmitting side, successive interference cancellation (SIC) at the receiver are the core techniques of NOMA [4].

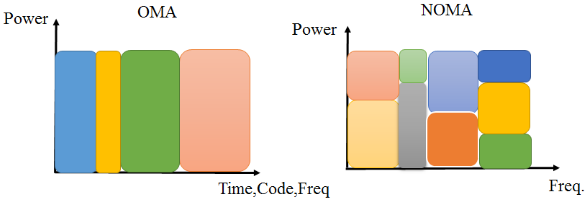


Fig. 1. Graphic comparison of NOMA and OMA

Index modulation (IM) is a simple but effective digital modulation method that sends extra information bits via the resource indices of a matching communication system. Compared to conventional communication systems, IM offers significant spectrum and energy efficiency, making it a potential approach for the future. By using subcarriers in addition to the conventional M-ary signal topologies, index modulation transmits data in a novel way. The benefits of IM-assisted NOMA over orthogonal approaches are high SE as well as EE [5] [6]. The IM-NOMA for downlink has been addressed in this study. A network comprising two users is intended to implement the proposed system. In this paper, a novel IMNOMA system is proposed. The results demonstrate the fact that the IM-

Identify applicable funding agency here. If none, delete this.

NOMA SE performs well than the OMA system. The topics of spectrum and energy efficiency are briefly discussed.
The proposed IM-NOMA system is discussed in further detail in Section 2 of the study. The simulation results for SE and EE are shown and discussed in Section 3. In Section 4, the paper is concluded.

II. SYSTEM MODEL

NOMA system shown in Figure 2 consists of a single Base station (BS) connected with two users.

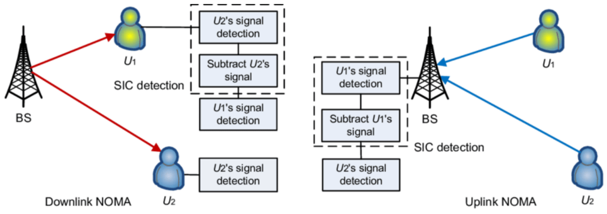


Fig. 2. Two User NOMA Network

A. IMNOMA system

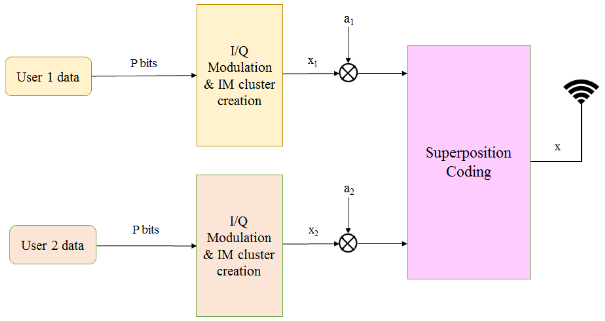

User 1 data
P bits
I/Q Modulation & IM cluster creation
x1
a1
User 2 data
P bits
I/Q Modulation & IM cluster creation
x2
a2
Superposition Coding
x

Fig. 3. Downlink IM-NOMA system model

Consider an IMNOMA network that has only one BS serving N_T users. The whole bandwidth is broken up using L_T orthogonal subcarriers. There is a B_c bandwidth for each subcarrier. L close-by subcarriers provide service to the $N \leqslant N_T$ users. Each user gets P bits. As in the index modulation idea, every user sends P bits that are split into the two blocks: - The indices of active subcarriers combination and M-ary constellation symbols. In the first part, which is made up of $Klog_2M$ bits, and the second part, which is made up of $log_2\binom{L}{K}$ bits, each user will receive a total of P bits, where K ranges from 1 to L combined make up the P bits that will be sent to user. The first part is modulated by using conventional modulation technique. The next part is taken into consideration while choosing an active subcarrier to carry the associated modulated complex symbol C_n. As a result, $log_2\binom{L}{K} + Klog_2M$ bits reach the system's transmitter during each transmission interval.
The signal transmitted over subcarrier 1 is specifically represented as, for the nth user. [7]

$$x(\varepsilon) = \begin{cases} s(\varepsilon) & \text{for } \varepsilon \in I \\ 0 & \text{for } \varepsilon \notin I \end{cases}, \varepsilon \in 1, 2...N \qquad (1)$$

Similar procedure is followed by each user, the modulated M-ary symbol and active sub-carriers are transmitted. The processes are then followed as in power domain NOMA. FU having weak channel coefficient is given more power than near user. The superimposed signal after SC process that is transmitted for user n is given by,

$$x_{SC} = \sum_{n=1}^{N} \sqrt{a_n P_{BS}} \left[x_n \left(\alpha_k \right) \right] \qquad (2)$$

Where, P_{BS} - transmit power per sub-carrier at the base station. Hence, each user follows the same process for selecting the complex symbol and active subcarriers which carries the symbols. As in conventional Power domain NOMA systems, the signals to be transmitted are multiplexed over all active subcarriers and simultaneously transmitted to the different users. According to the superposition coding principle, additional power is apportioned to the far user which has weak channel coefficient compared to the near user. Information signal received by the user is given by,

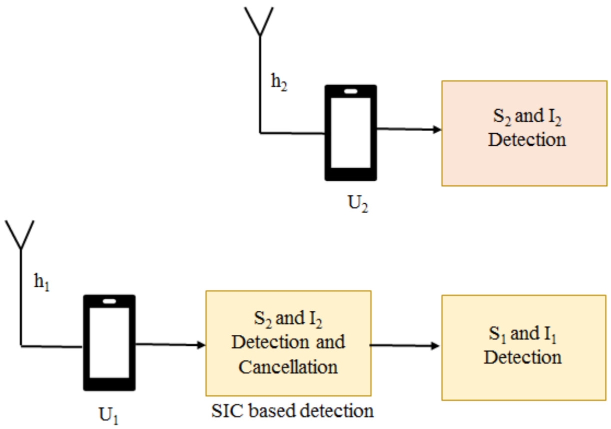


Fig. 4. SIC detection of IM-NOMA

$$Y = Hx_{SC} + W \qquad (3)$$

The channel matrix is as follows,

$$H = diag \left\{ h\left(1\right) ... h\left(N\right) \right\} \qquad (4)$$

w symbolizes the AWGN vector,

$$w = \left[w\left(1\right) ... w\left(N\right) \right]^T \qquad (5)$$

$W \sim CN(0, \sigma^2)$. H follows the rayleigh fading channel distribution. According to the Superposition Coding, near user with good channel conditions is allotted with less power coefficient than that of far user. In SIC, each user nearer to the Base Station (BS) detects and removes the signals that are far from the BS compared to that user and treats all other user's signal as noise. The decoded signal will be removed from the received signal, the same process continues till it decodes its own signal. All user's signals are allotted with power coefficients by the BS according to the SC technique

III. Energy efficiency

Spectral efficiency is the fraction of the sum rate to the bandwidth.

$$SE = \frac{R_T}{W} \tag{6}$$

In orthogonal multiple access technology, each user is assigned with the single channel. The spectral efficiency is,

$$SE_{OMA} = log_2(M) \tag{7}$$

In OMA, SE is independent of L.
In NOMA technique, all users use the available channels available in the network. The SE for NOMA is given below,

$$SE_{NOMA} = Llog_2(M) \tag{8}$$

The spectral efficiency of IM-NOMA is given below [8],

$$SE_{IM-NOMA} = log_2\binom{L}{K} + Klog_2M \tag{9}$$

Where, $K \leqslant L$ - total active channels.
Trade-off between BER and SE is achieved using flexible setups since only active channels are used to carry symbols. Only active channels will carry symbols so that trade-off between BER and SE can be made through flexible setups. [bpcu] stands for bits per channel use.
Energy efficiency (EE) is the ratio between the sum rates to the base station's power. The message signal power and the circuit power represent the total power consumption at the transmitting side.

$$EE_{IM-NOMA} = R_T/P_T \tag{10}$$

According to Shannon's theory, the relationship between EE and SE will not consider the power consumption of the circuit, so it is monotonic. SE and EE are inversely proportional to each other i.e., EE increases for lesser SE region and EE decreases for higher SE region. Peak of the curve denotes the EE at its maximum value. EE-SE relationship is linear for the positive slope of the curve. SE increases with the rise in EE.

IV. 4. Results And Discussion

The energy efficiency of the proposed IMNOMA system is analyzed for different values of subcarriers. Following are the different system parameters chosen for Monte-Carlo simulation. Modulation order range is $M = 2, 4$ and the values of active subcarriers considered are $k = 1, 2, 3$. Circuit Power is assumed to be 60W and the bandwidth considered is $6*10^6$. From the Figure 5, we can observe that for a given M and L, the energy efficiency increases with the increase in the number of active subcarriers. Energy efficiency also increases with the increase in the modulation order. The Figure 6 shows the EE versus the number of active subcarriers K for different values of $L = 2, 4, 8$ with the modulation order M=4. For a given value of L, the energy efficiency increases with the increase in active subcarriers K. The graph indicates that when $K = L/2$, the energy efficiency attains its maximum point. Once the EE attains the maximum value, for the next values of K, the EE decreases with the increase in K.

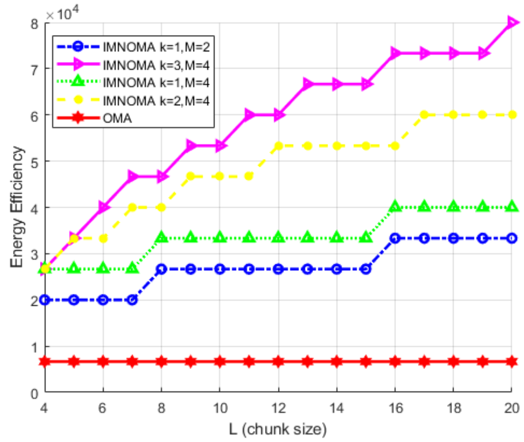


Fig. 5. Energy efficiency versus L for different values of K

The Figure 7 indicates the Energy Efficiency of IMNOMA scheme versus the modulation order M for different values of K and L. OMA and IMNOMA spectral efficiency results are compared. In OMA system, the energy efficiency is independent of L and hence there is no enhancement in EE. However, IMNOMA has additional $log_2\binom{L}{K}$ bps/Hz compared to OMA system. In Figure 7, IMNOMA achieves higher EE compared to OMA irrespective of the modulation order.

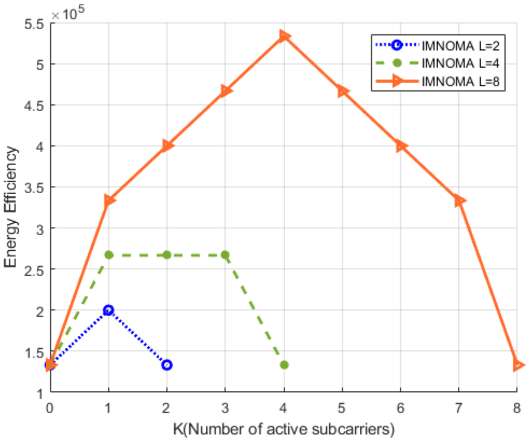


Fig. 6. Energy efficiency versus K for different values of L

The Figure 8 shows the energy efficiency versus the transmit power in dBm. Two subcarriers are made active out of four subcarriers. i.e., $L = 4, K = 2$. Circuit power range is assumed to be $P_c = 40W, 50W, 60W$. Modulation order M=2 is considered. From the figure 8, we can observe that when P is smaller than or equal to a certain threshold, the graph indicates the same energy efficiency value. Once P reaches the threshold value, the EE begins to decline. Hence, the optimal energy efficient systems work at the exact value of P when the power crosses the threshold value. The tradeoff between

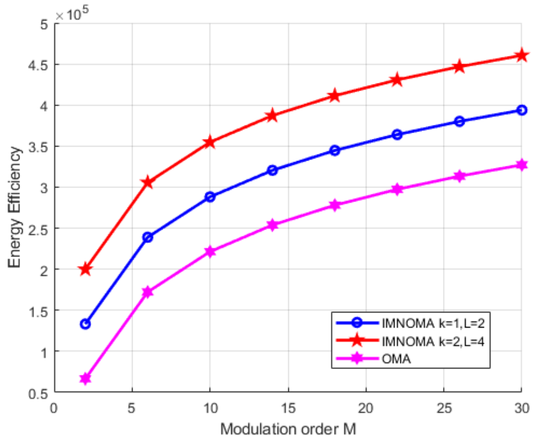


Fig. 7. Energy efficiency versus M for different values of L and K

EE and SE can be accomplished by adjusting the power. Since OMA system is independent of L, IMNOMA outperforms the OMA as shown in the Figure 8. We can also observe that the energy efficiency increases with the decrease in circuit power.

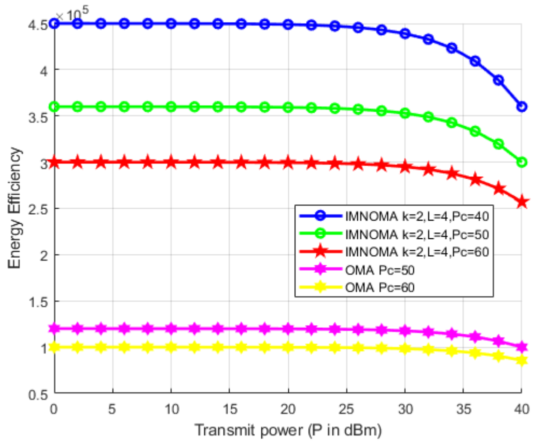


Fig. 8. Energy efficiency versus transmit power P for different values of $L = 4, K = 2, Pc = 40W, 50W, 60W$.

V. Conclusion

A novel IMNOMA system is proposed and energy efficiency of the proposed system is analyzed. According to the results of the simulation, IM-NOMA offers a greater energy efficiency than the existing OMA systems. The number of active subcarriers rises along with the energy efficiency for a given M and L. The point at which energy efficiency reaches its greatest is when $K = L/2$. P must be less than or equal to a predetermined threshold in order for the graph to display the same energy efficiency value. In comparison to OMA, IMNOMA performs better in terms of energy efficiency for different system parameters.

References

[1] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 28–41, 2019.

[2] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Communications Surveys and Tutorials*, vol. 20, no. 3, pp. 2294–2323, 2018.

[3] H. M. Shwetha and S. Anuradha, "Analysis of Downlink and Uplink Non-orthogonal Multiple Access (NOMA) for 5G," in *Proceedings of Third International Conference on Sustainable Computing SUSCOM 2021*, 2022, pp. 385–395. [Online]. Available: https://link.springer.com/10.1007/978-981-16-4538-9_38

[4] M. Liaqat, K. A. Noordin, T. Abdul Latef, and K. Dimyati, "Power-domain non orthogonal multiple access (PD-NOMA) in cooperative networks: an overview," *Wireless Networks*, vol. 26, no. 1, pp. 181–203, 2020. [Online]. Available: https://doi.org/10.1007/s11276-018-1807-z

[5] H. M. Shwetha and S. Anuradha, "Index Modulation-Based Non-orthogonal Multiple Access (IM-NOMA): Spectral Efficiency Analysis," in *Proceedings of Second International Conference on Computational Electronics for Wireless Communications*, 2023, pp. 517–526. [Online]. Available: http://link.springer.com/10.1007/978-981-13-8222-2

[6] M. Irfan, B. S. Kim, and S. Y. Shin, "A spectral efficient spatially modulated non-orthogonal multiple access for 5G," *2015 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2015*, pp. 625–628, 2016.

[7] H. M. Shwetha and A. Sundru, "A novel multicarrier index keying-aided non-orthogonal multiple access systems in presence of uncertain channel state information," *International Journal of Communication Systems*, no. May, pp. 1–22, 2023.

[8] H. M. Shwetha and S. Anuradha, "Performance Analysis of Novel Index Modulation-Based Non-Orthogonal Multiple Access Systems over Nakagami-m Fading Channels with Imperfect CSI," *Radioengineering*, vol. 32, no. 3, pp. 425–437, 2023.

Session 5B: Smart IoT & Software Platform

Chair: Prof. Kwanghoon Kim, Kyonggi University, Korea, ,

1 Paper ID: 20240333, 369~372

The Data Alignment Method between GPS and IMU based on ICP for Indoor Positioning

Mr. Yong Hee Park, Mr. Min Gu Kang, Mr. Jang Hyeon Jeong, Prof. Seong Gon Choi,

Pabat. Korea(South)

2 Paper ID: 20240299, 373~375

Incorporation of waypoint following logic into ROS publish and subscribe mechanism

Mr. Minhwa Hong, Prof. Seonggon Choi, Dr. Heonjong Yoo,

Chungbuk National University. Korea(South)

3 Paper ID: 20240438, 376~383

Development and Implementation of BOSESKO: A Synoptic Multi-platform Digital Citizen Participatory System

Dr. Jennifer Llovido, Dr. Michael Angelo Brogada, Dr. Floradel Relucio, Dr. Lea Austero, Dr. Lany Maceda, Dr. Mideth Abisado,

Bicol University. Philippines

4 Paper ID: 20240348, 384~388

An IoT-Based Early Warning System for Settlement Monitoring Using Differential Pressure Static Level

Mr. Tieyan CHAO, Ms. Hui LIANG, Mr. Yuwei GE, Mr. Kai HOU, Ms. Xiang DONG, Mr. Ting PENG,

Shaanxi Huashan Road and Bridge Group Co., LTD. China

5 Paper ID: 20240073, 389~394

Tunnel Construction Site Monitoring and Digital Twin System

Mr. Wei CHENG, Mr. Yuxing PAN, Mr. Zhi MA, Mr. Yincai CAI, Dr. Yuan LI, Prof. Ting PENG,

Sichuan Chuanjiao Road and Bridge Co., LTD. China

6 Paper ID: 20240339, 395~401

Research on LSTM-based Model for Predicting Deformation of Tunnel Section During Construction Period

Mr. Jiwen ZHANG, Mr. Kai YUAN, Mr. Jianjun MAO, Mr. Yincai CAI, Mr. Dongfeng LEI, Mr. Jinyang DENG, Mr. Ting PENG,

Sichuan Chuanjiao Road and Bridge Co., LTD. China

The Data Alignment Method between GPS and IMU based on ICP for Indoor Positioning

Yong Hee Park*, Min Gu Kang**, Jang Hyeon Jeong***, Seong Gon Choi**

**Pabat Corp., Seocho-gu, Seoul, South Korea*

***Information & Communication Engineering, Chungbuk National University, Cheongju-si, Chungcheongbuk-do, South Korea*

****JJsolution Corp., Chungbuk National University, Cheongju-si, Chungcheongbuk-do, South Korea*

dydgml1994@gmail.com, kkmg0157@chungbuk.ac.kr, wkdgus4788@chungbuk.ac.kr, choisg@cbnu.ac.kr

Abstract— **We propose a method to align IMU data with GPS data to measure precise data indoors. While GPS can provide absolute coordinates, it has problems with signal strength reduction and measurement failure in indoor environments. Compared to GPS, IMU data offers high precision and less location constraint. In contrast, it does not have absolute coordinates, and cumulative errors can occur. Measurement of absolute coordinate in indoor can demand in various fields such as factories, and related research has been studied. Previously, combining GPS and IMU data was commonly used to complement real-time errors. In this paper, we propose a method that aligns IMU data with outdoor GPS data to measure indoor absolute coordinates. Simulation results confirm the successful alignment of these two datasets.**

Keywords— **IoT, Positioning, Communication, Location Estimation, Database**

I. Introduction

We propose a method for measuring indoor absolute coordinates by utilizing the Iterative Closest Point (ICP) algorithm to align Global Positioning System (GPS) and Inertial Measurement Unit (IMU) data. GPS is a commonly used sensor for coordinate measurement, providing latitude and longitude information. However, GPS is troubled by accuracy issues, and research efforts such as GPS-RTK have been studied to address these shortcomings. Nevertheless, indoor GPS measurements remain challenging due to signal strength issues.

IMU is a sensor that combines accelerometers, gyroscopes, and magnetometers, allowing for the measurement of relative coordinates. While IMU data ensure high sampling rates and resolution, it does not provide absolute coordinates like GPS. To overcome their respective limitations, GPS and IMU are often fused using techniques such as the Kalman filter [1].

Indoor positioning requirements vary depending on the application, such as in factories, and various technologies like Ultra-wideband (UWB) and Bluetooth beacons are being researched [2]. However, these technologies also suffer from large error ranges and the inability to determine absolute coordinates. Combining GPS and IMU technology faces challenges when used indoors.

We propose a measuring method that indoor absolute coordinates by leveraging the well-established GPS and IMU technologies. To achieve this, we employ the ICP algorithm to align GPS and IMU data and infer absolute coordinates within indoor environments.

II. Related work

GPS is commonly used for coordinate measurement, but it faces challenges when used indoors due to signal degradation caused by roofs and walls. Consequently, numerous research efforts have been dedicated to indoor positioning. Ultra-wideband (UWB) technology has been developed for this purpose, utilizing wireless communication with a bandwidth of over 500MHz. While UWB is highly effective with an error range of around 10 cm, it falls short in precision compared to IMU [3].

Research is also being conducted on indoor positioning using Bluetooth Low Energy (BLE). Through the commonly used Bluetooth module, the indoor location can be measured using techniques such as triangular positioning. However, signal information with RSSI as the criterion for positioning is sensitive to various interferences, and there is a problem that the measurement value is very unstable [4].

The Iterative Closest Point (ICP) algorithm is used to align one point with another point scanned from a different viewpoint. ICP algorithm is a widely employed technique in computer vision and robotics. It aims to align or register two sets of 3D data points. This process includes an initial estimate of the transformation (translation and rotation), finding corresponding points between the datasets, iteratively optimizing the transformation to minimize the squared differences between corresponding points, and refining it until convergence criteria are met. ICP is essential for applications like 3D reconstruction, simultaneous localization, and mapping (SLAM), and object recognition.

We propose method to measure coordinate data in constraint environment. To achieve this, it aligns the measurable GPS data with IMU data.

III. Proposal Method

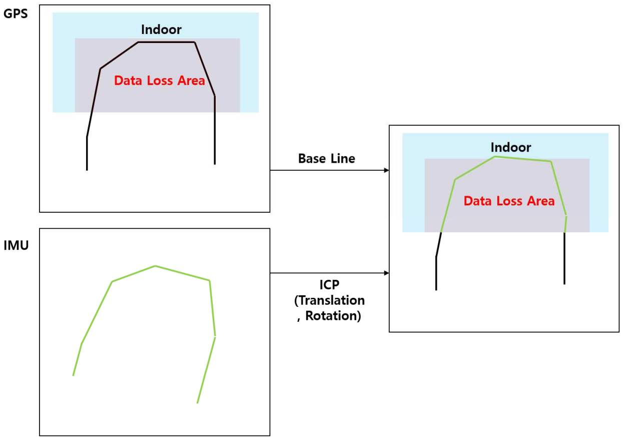


Figure 1. Indoor data measurement plan

The method we propose is specifically illustrated in Figure 1. GPS experiences a significant decrease in accuracy upon entering indoor environments. IMU remains unaffected by indoor conditions but lacks knowledge of its absolute coordinates. To compensate for this, we move and rotate IMU data using data from GPS measured outdoors as a reference point to align it with absolute coordinates. During this process, IMU data is recorded indoors and can be utilized as indoor positioning data.

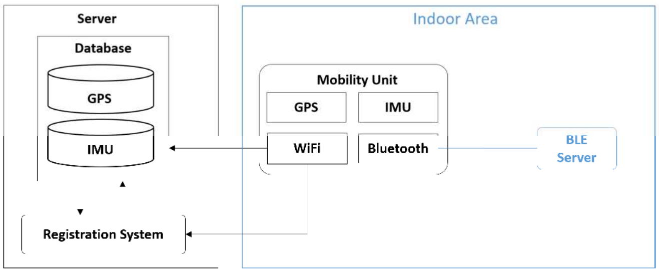


Figure 2. Functional diagram

The functional overview diagram is illustrated in Figure 2. The Mobility Unit must be capable of collecting GPS and IMU data, while also has Wi-Fi and Bluetooth functionalities. The BLE Server, upon the Mobility Unit entering an indoor area, determines indoor access through Bluetooth MAC scanning. GPS and IMU data from the mobile unit are stored in the server's database using Wi-Fi. The server interacts with the Registration System to execute the ICP alignment of GPS and IMU data.

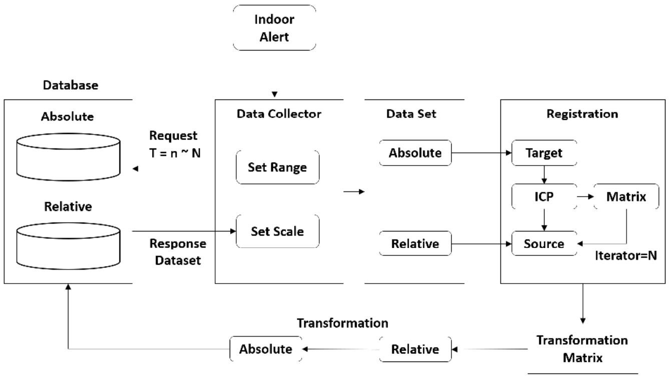
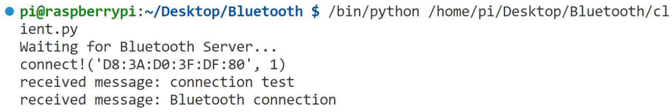


Figure 3. System configuration diargram

The system configuration diagram is presented in Figure 3. When the Mobility Unit is detected to have entered indoors, the Data Collector initiates the data shifting process. The measured latitude and longitude data from GPS and the data from IMU, which are in X and Y coordinate format, are first converted. Since ICP involves a considerable computational load, we scale both datasets relative to their respective means. Additionally, since IMU data is sampled at a much higher rate compared to GPS, it is adjusted to match the GPS sampling frequency. The reason for the higher sampling rate of IMU data compared to GPS data is due to the significantly faster measurement rate of IMU. Typically, IMUs collect data such as acceleration, angular velocity, and magnetic field at a very rapid pace. This higher data acquisition rate allows IMUs to provide more detailed and fine-grained information about movements and actions. GPS operates at a relatively slower measurement rate, with slower updates for altitude and position information. As a result, IMU data is collected at a much faster sampling rate than GPS data. It also aligns with the IMU's data range appropriately, as it calibrates using outdoor data rather than corrupted GPS data. Thereafter, ICP iterations are performed as configured.

IV. Simulation

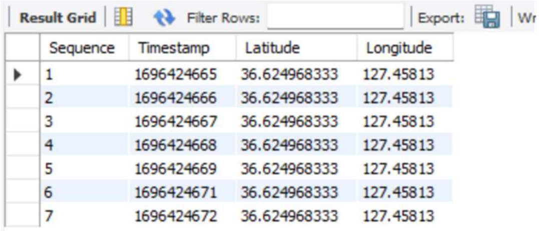
The GPS data used in the simulation was taken from ETH Zurich, which provides a localization dataset for SLAM [5]. IMU data was randomly generated based on GPS data. The number of data is 1294 each for GPS and IMU.



```
● pi@raspberrypi:~/Desktop/Bluetooth $ /bin/python /home/pi/Desktop/Bluetooth/cl
  ient.py
  Waiting for Bluetooth Server...
  connect!('D8:3A:D0:3F:DF:80', 1)
  received message: connection test
  received message: Bluetooth connection
```

Figure 4. Bluetooth MAC scanning

The server determines indoor access by establishing Bluetooth pairing with the mobility unit. Therefore, the server must periodically scan Bluetooth MAC addresses. The results of the Server-to-Mobility unit Bluetooth MAC sensing is in Figure 4. When MAC sensing is completed, the MAC address of the mobility unit and the connected port number are shown.



Sequence	Timestamp	Latitude	Longitude
1	1696424665	36.624968333	127.45813
2	1696424666	36.624968333	127.45813
3	1696424667	36.624968333	127.45813
4	1696424668	36.624968333	127.45813
5	1696424669	36.624968333	127.45813
6	1696424671	36.624968333	127.45813
7	1696424672	36.624968333	127.45813

Figure 5. Database stored on the server

The result of storing GPS data of the mobility unit in the server's database is in Figure 5. The database stores the timestamp, latitude, and longitude information of the mobility unit. The timestamp information is stored in UNIX time format. GPS data is collected in string form and uses the NMEA 0183 protocol. Latitude and longitude data are collected in the GPGGA format among the collected data.

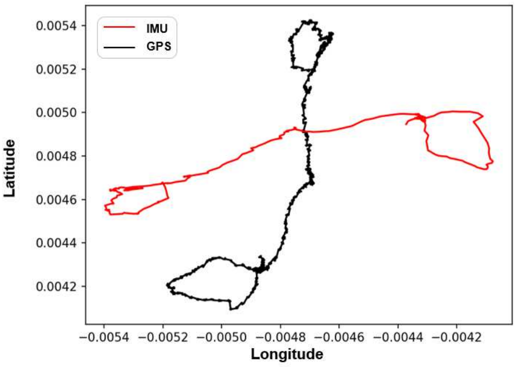
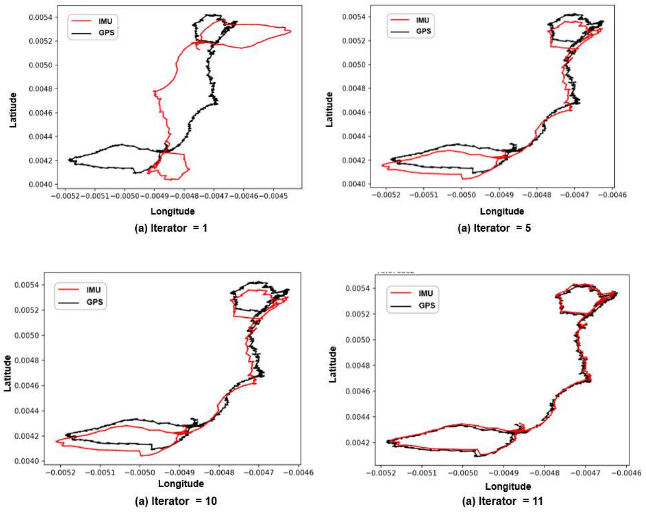


Figure 6. GPS/IMU data samples

The value of the shifted data in the IMU as much as the GPS average data is in Figure 6. The scaling of GPS data and IMU data was standardized through a process known as data shifting.



(a) Iterator = 1

(a) Iterator = 5

(a) Iterator = 10

(a) Iterator = 11

Figure 7. Alignment results by iteration count of ICP

The result of data matching by the number of repetitions of ICP is in Figure 7. From the results of this simulation, it can be confirmed that the change for matching is large when the initial ICP is performed, but the change gradually decreases as the number of times increases. It can be confirmed that the data of the two are matched when the performance of ICP is repeated 11 times.

V. Conclusion

We proposed a method for aligning GPS and IMU data to measure indoor positioning. To achieve this, several data preprocessing steps were carried out, followed by the iterative execution of ICP until the data alignment was reached. Simulation results confirmed the successful alignment of data from both GPS and IMU data.

IMU has the issue of accumulating errors. The method proposed in this paper can accurately determine the position upon entering indoors from outdoors, but it struggles to address accumulated errors with prolonged indoor stay. There is a need for improvement in addressing this issue, and we plan to conduct future research to explore supplementary methods, such as collaboration with other devices.

Acknowledgment

This work was partly supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2024-2020-0-01462, 33%) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). This work was partly supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2020R1A6A1A12047945, 34%). This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C1006167, 33%).

*corresponding author is S.G. Choi(choisg@cbnu.ac.kr)

References

[1] Francois Caron, "GPS/IMU data fusion using multisensor Kalman filtering: introduction of contextual aspects", Information Fusion, pp. 221-230, 2006.

[2] Tae Yun Jung, Eui Rim Jeong, Recurrent Neural Network Based Distance Estimation for Indoor Localization in UWB System, Journal of the Korea Institute of Information and Communication Engineering, Vol. 24, No. 4: pp. 494-500, 2020.

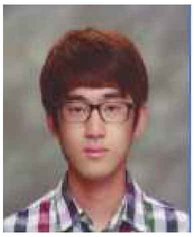
[3] Chang Eum Lee, Tae Gyeong Seong, "UWB positioning technology introduction and technology trends," The Journal of Korean Institute of Communications and Information Sciences, vol. 34 NO.4: pp. 0003–0009, Apr. 2017.

[4] Chang Pyo Yoon, Chi Gon Hwang, "Efficient indoor positioning systems for indoor location-based service provider", journal of the Korea Institute of Information and Communication Engineering, pp. 1368-1371, 2015.

[5] A.L. Majdik, "The Zurich Urban Micro Aerial Vehicle Dataset for Appearance-based Localization, Visual Odometry, and SLAM", ETH zurich, 2017.

Yong Hee Park received B.S. and M.S degree in the College of Information & Communication Engineering, Chungbuk National University, Korea in 2019 and 2022. His research interests include Autonomous Vehicle, AI, Smart Grid. He is currently researcher in Rejuvenor Inc. and representative in Pabot Co.

Min Gu Kang is currently undergraduate in the College of Information & Communication Engineering, Chungbuk National University. His research interests include Autonomous Vehicle, Smart Grid, Cloud Computing.



Jang Hyeon Jeong received B.S. and M.S. degree in the College of Electrical & Computer Engineering, Chungbuk National University, Korea in 2019 and 2021. His research interests include Network Security, Smart Grid. He is currently researcher in Xabyss Inc and CEO in JJsolution Inc. His research interest is network security.

Seong Gon Choi received B.S. degree in Electronics Engineering from Kyungpook National University in 1990, and M.S. and Ph.D. degree from KAIST in Korea in 1999 and 2004, respectively. He is currently a professor in College of Electrical & Computer Engineering, Chungbuk National University. His research interests include V2X, AI, smart grid, IoT, mobile communication, high-speed network architecture and protocol.

Incorporation of waypoint following logic into ROS publish and subscribe mechanism

*Corresponding author

1st Minhwa Hong
The Department of Electrical and Computer Engineering
Chungbuk National University
Chungdae-ro, Seowon-gu, Cheongju-city,
Chungcheongbuk-do, 28644 Republic of Korea
email address: alsghk0429@naver.com

2nd Seonggon Choi*
The Department of Electrical and Computer Engineering
Chungbuk National University
Chungdae-ro, Seowon-gu, Cheongju-city,
Chungcheongbuk-do, 28644 Republic of Korea
email address: sgchoi@chungbuk.ac.kr

2st Heonjong Yoo
The Department of Electrical and Computer Engineering
Chungbuk National University
Chungdae-ro, Seowon-gu, Cheongju-city,
Chungcheongbuk-do, 28644 Republic of Korea
email address: 622061@chungbuk.ac.kr

***Abstract*—The waypoint following logic is developed and it is incorporated into ROS publish and subscribe block for 2 wheel mobile platform. Firstly, the waypoint follower in mobile robotics simulation toolbox in simulink is introduced. On the other side of research, platform can be moved through ROS connection. In this presentation, the development of waypoint following logic is experimented with the real mobile robot with ROS connection.**

***Index Terms*—Waypoint following logic, ROS connection**

I. Introduction

The waypoint following logic is developed for decades, see in . For example, pure pursuit is utilized for 2 wheel mobile platform, in 2. In this presentation, we use different way point follow logic using state flow box in Simulink.

New models are often encountered in the agricultural, protection, and precision sectors. Computers, artificial intelligence, and big data technologies have been used to develop intelligent farming systems, particularly for farming robots [1]. With the rapid development of advanced technologies, many new techniques, such as big data, artificial intelligence, the Internet of Things, machine vision, and agricultural robotics, have been applied to agricultural production [2]. Recently, the pure pursuit method has been widely used for the path tracking of outdoor and indoor mobile robots. However, these algorithms and methods are controlled using the vector interval, which causes the mobile platform to vibrate during implementation. Automated navigation technology plays a crucial role in the autonomous navigation of mobile robots in engineering field [3]. The advantage of the pure pursuit algorithm is that the path of the mobile robot follows waypoints. Several methods have been applied to agricultural field operations for robotic localization. Localization systems such as the global positioning system (GPS) [4], real-time kinematic GPS (RTK-GPS) [5], geographic information systems [14], and LiDAR-based systems have been applied to agricultural mobile robotic systems [6]. Recently, linear active disturbance rejection control (LADRC) was designed for the trajectory tracking control problem of a differential-type model [10]. A recursive technique was applied to the path-tracking problem of the differential-type model by composing a chained form of the system [6] [8]. This paper introduces a novel Stateflow algorithm recently introduced by MathWorks. Several researchers [7], [8] have investigated the Stateflow method, namely, its applicability to flight control, hybrid energy control system design, and simulation [9]. The benefit of the waypoint following logic is that the platform is controlled using time interval control rather than vector interval control, in which the vibration of the platform is reduced, on the other hand, the pure pursuit and, Stanley methods have that of the platform. Therefore, we determined that the waypoint following logic is appropriate for the path-tracking problem in an UGV(Unmanned guided vehicle) environment.

Thus, the contribution of this paper can be expressed as

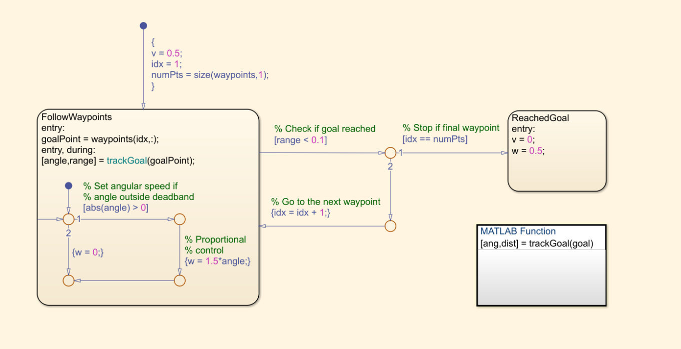
- Despite the various aforementioned control and estimation methods, the experiment was conducted focusing on the waypoint following logic incorporated into ROS publish and subscribe block introduced in MATLAB/SIMULINK recently.
- The actual path is implemented based upon the waypoints from ROS subscribe block.

II. Methods

The waypoint following logic is coded by

First of all, angle and range are extracted using ”trackGoal” code. If absolute value of angle is positive, the angular velocity

Identify applicable funding agency here. If none, delete this.



```
{
v = 0.5;
idx = 1;
numPts = size(waypoints,1);
}
```

```
FollowWaypoints
entry:
goalPoint = waypoints(idx,:);
entry, during:
[angle,range] = trackGoal(goalPoint);
```

```
% Set angular speed if
% angle outside deadband
[abs(angle) > 0]
```

```
{w = 0;}
```

```
% Proportional
% control
{w = 1.5*angle;}
```

```
% Check if goal reached
[range < 0.1]
```

```
% Stop if final waypoint
[idx == numPts]
```

```
% Go to the next waypoint
{idx = idx + 1;}
```

```
ReachedGoal
entry:
v = 0;
w = 0.5;
```

```
MATLAB Function
[ang,dist] = trackGoal(goal)
```

Fig. 1. Waypoint Code

signal is present, otherwise, angular velocity signal is zero. If range approaches zero, the linear velocity is forced to be zero, which means platform stops. The waypoint following logic is incorporated into ROS publish and subscribe blocks in Simulink. The input of the subsystem is longitude,latitude $x(t), y(t)$ and sideslip angle $h(t)$. Furthermore, the output of the subsystem is linear and angular velocity signal $v(t), w(t)$

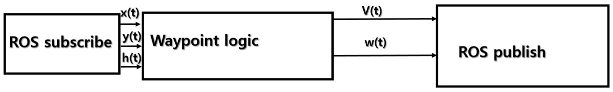


Fig. 2. The overall scheme of waypoint following into ROS subscribe and publish block

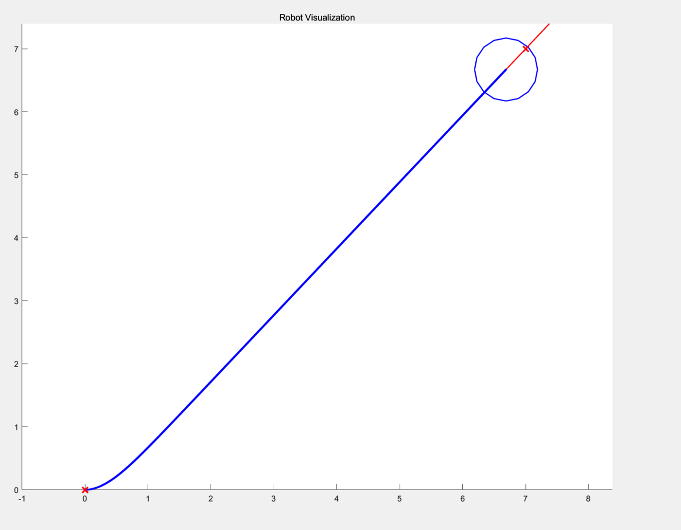


Fig. 3. The actual path using waypoint following logic

A. Maintaining the Integrity of the Specifications

The experimented waypoint condition is given as

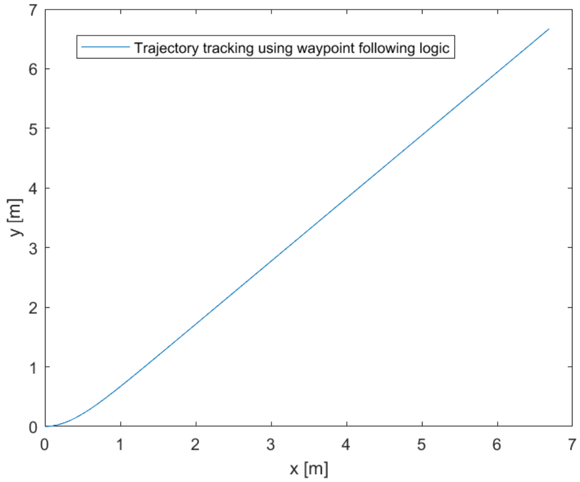


Fig. 4. Plot on MATLAB prompt

TABLE I
WAYPOINT FOLLOWING LOGIC

Table Head	**Waypoint following condition**		
	Waypoint	***Starting point***	***End point***
a	Predetermined set waypoints	(0,0)	(7,7)

III. EXPERIMENT

The linear and angular velocity signal is automatically calculated using way point code in Fig. 1.

IV. OUTPUT FEEDBACK CONTROL METHOD

The original system modeling have the following relation, given as The above system is MIMO(Multi-input and multi-output) system. Since linear velocity signal doesn't affect the output signal only by using open-loop experiment, we consider The above system is SISO(Single input and single output) system, so that it is more simplistic for estimating transferfunction. For the system modeling, transfer estimation code ”tfest” is used in MATLAB, then ”tf2ss” code results in state space model, given as

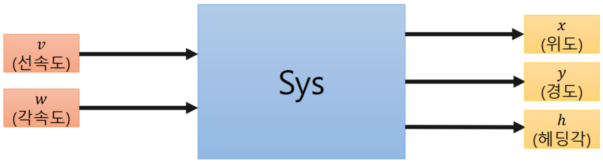


Fig. 5. Original system modeling

$$\begin{bmatrix} \dot{h}_1'(t) \\ \dot{h}_2'(t) \end{bmatrix} = \begin{bmatrix} -0.001 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} h_1'(t) \\ h_2'(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$
$$\begin{bmatrix} h(t) \end{bmatrix} = \begin{bmatrix} 0.01 & 0 \end{bmatrix} \begin{bmatrix} h_1'(t) \\ h_2'(t) \end{bmatrix} \quad (1)$$

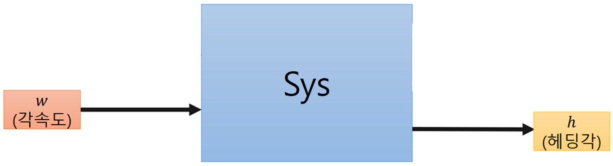


Fig. 6. Original system modeling

The output feedback control input is given as

$$u(t) = y(t) \tag{2}$$

After that, from workspace block is used with the following code, given as

$$\begin{aligned} &sampleTime = 60/1901;\\ &numSteps = 60;\\ &time = sampleTime * (0 : numSteps - 1);\\ &time = out.t;\\ &simin = [time, out.w2];\\ &simin = [time * 10^{(} - 2), out.w2]; \end{aligned} \tag{3}$$

60 is the size of the first experiment, on the other hand, 1901 is the size of the output feedback simulation.Using above code, we can modify the shape of controlled input signals for closed loop experiment by using the input and output data sets from open-loop experiment. The experiments we implemented is described as Originally, the path is deviated from $(7, 7)$ in open loop experiment. By using output feedback inputs $w(t)$, the closed-loop experiment is implemented to target $(7, 7)$, final coordinates.

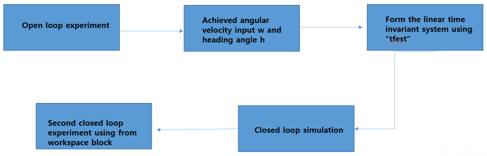

Open loop experiment
Achieved angular velocity input w and heading angle h
Form the linear time invariant system using "tfest"
Closed loop simulation
Second closed loop experiment using from workspace block

Fig. 7. The description of open-loop, closed-loop experiments

Future Work and Conclusion

In summary, this study demonstrated the successful implementation of a waypoint following with ROS publish and subscribe block using MATLAB/Simulink. Firstly, the waypoint following logic is introduced. Based on the system design, ROS publish and subscribe block is incorporated into waypoint logic. In the future, the interface will be implemented for the 4 wheel independent steering system. Furthermore, the output feedback control is applied to the real experiment using open loop experiment by achieving angular velocity, input and heading angle output.

Acknowledgment

This work was partly supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2024-2020-0-01462, 33%) supervised by the IITP (Institute for Information communications Technology Planning Evaluation). This work was partly supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2020R1A6A1A12047945, 34%). This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C1006167, 33%).

References

[1] Kang D. H. and Lee, S. Y., Kim, J. K., Park, M. J., Son, J. K. and Yum, S.W., "Development of an automatic grafting robot for fruit vegetables using image recognition," *The Korean Society for Bio-Environment Control,* Vol. 28, No. 4, pp. 322-327, 2019.

[2] Sabanci. K. and Aydin, C., "Image processing based precision spraying robot," *Tarim Bilimleri Dergisi,* Vol. 20, No. 4, pp. 406-414, 2014.

[3] Samuel, M., Hussein, M. and Mohamad, M.B., "A review of some pure-pursuit based path tracking techniques for control of autonomous vehicle," , *International Journal of Computer Applications,* Vol. 135, No. 1, pp. 35-38, 2016.

[4] Xue, J., Zhang, L. and Grift, T.E., "Variable field-of-view machine vision based row guidance of an agricultural robot," , *Computers and Electronics in Agriculture,* Vol. 84, pp. 85-91, 2012.

[5] Weiss, U. and Biber, P., "Plant detection and mapping for agricultural robots using a 3D LIDAR sensor," , *Robotics and Autonomous Systems,* Vol. 59, No. 5, pp. 265-273, 2011.

[6] Torii, T., "Research in autonomous agriculture vehicles in Japan," , *Computers and Electronics in Agriculture,* Vol. 25, No. 1-2, pp. 133-153, 2000.

[7] Bakht, M.P.; Salam, Z.; Bhatti, A.R.; Anjum, W.; Khalid, S.A.; Khan, N., "Stateflow-based energy management strategy for hybrid energy system to mitigate load sheding" , *Applied Science* , Vol. 11, No. 10, 2021.

[8] Heonjong, Y., Kyeonghwan, L., Tean, C., "Development of the path following platform for a unicycle-type mobile robot in the indoor environment based on state flow method" , *Korean Society of Agricultural Machine* , 2020.

[9] Jarrod M. S., "Automatic steering methods for autonomous automobile path tracking" , *Doctoral Dissertation* , USA, 2009.

[10] Wang, X., Wang, F. and Wei, W., "Linear active disturbance rejection control of dissolved oxygen concentration based on benchmark simulation model number 1," , *Mathematical Problems in Engineering,*Vol. 2015, No. 178953, 2015.

[11] Jiang, Z.P. and Nijmeijer, H., "A recursive technique for tracking control of nonholonomic systems in chained form," , *IEEE Transactions on Automatic Control,* Vol. 44, No. 2, pp. 265-279, 1999.

BOSESKO: Designing A Synoptic Multi-platform Digital System for Citizen Participation

Jennifer L. Llovido*, Michael Angelo D. Brogada*, Floradel S. Relucio**, Lea D. Austero*, Lany L. Maceda*, Mideth B. Abisado***

*Computer Science and Information Technology Department, College of Science, Bicol University, Legazpi City, Albay, Philippines

**Computer Studies Department, Bicol University Polangui, Polangui, Albay Philippines

***National University, Manila, Philippines

jllovido@bicol-u.edu.ph, madbrogada@bicol-u.edu.ph, fsrelucio@bicol-u.edu, ldaustero@bicol-u.edu, llmaceda@bicol-u.edu.ph, mbabisado@national-u.edu.ph

Abstract— **Digital citizen participatory toolkits are gaining interest among researchers and practitioners for their crucial role in empowering citizens, promoting accountability, and ensuring diverse voices are heard in policymaking. This study aims to develop and implement BOSESKO: Building on Opinions and Sentiments for Sustainability and Knowledge Opportunities (formerly known as Kalahok) - a multilingual, inclusive, deliberative, synoptic, digital participatory toolkit that digitized data collection and analysis to engage communities in governance using technology-based methodologies. BOSESKO is available in English, Filipino, Ilokano, and Bikol versions for web and mobile devices. It primarily encourages public feedback on disaster preparedness and Universal Access to Quality Tertiary Education (UAQTE) implementation in the Philippines. Its adaptable design extends its utility beyond its initial scope. BOSESKO explored machine learning, natural language processing, and software integration for data gathering, processing, visualization, and system development while employing a hybrid approach with Extreme Programming (XP) and Scrum. Significant findings demonstrated that BOSESKO enabled the orderly solicitation and submission of inputs from local communities through the creation, management, consolidation, analysis, and visualization of responses. The result of the analysis based on the performance of BOSESKO's web application and mobile application 4.78 and 4.40, respectively, and this can guide agencies in formulating data-driven policies for UAQTE, Disaster Risk Reduction Management, Climate Adaptation (DRRM/CA), among others.**

Keywords— **Digital citizen participation, E-participation, System development, Integrated systems, Multi-platform systems, Hybrid software methodology**

I. Introduction

Technology influences every aspect of life; undoubtedly, it has become a necessary part of everyday activity. Its widespread use has changed how individuals interact with one another, conduct business, learn, and even participate in public activities. Due to the digitization of governments, e-participation has developed into a significant component for the collaborative democratic development of politics and domestic countries with the involvement of citizens [1]. The importance of citizen participation in shaping public policy has increased over the years [2]. Every democratic government strives to be inclusive of all decision-making processes, and this endeavor will succeed in realizing the entire principles of democracy, accountability, and good governance [3].

The core of an e-participation system is typically through websites, an information and communication technology (ICT) infrastructure. This website is easily accessible on PCs, laptops, netbooks, tablets, and smartphones and offers constant internet connectivity. It is beneficial to create applications that can be downloaded on smartphones and tablets in order to improve information accessibility [4]. E-participation has developed into a useful collection of tools that closes the digital gap between citizens and their governments by utilizing developments in ICT [3], and it significantly contributes to the evolution of the political landscape [5]. Citizen involvement is crucial in forming comprehensive national policies, and is a pivotal sign of a successful democratic government [6], [3].

Building on Opinions and Sentiments for Sustainability and Knowledge Opportunities, or BOSESKO, is a system developed to involve local communities in governance. It is implemented across various platforms, aiming to create a robust and efficient digital citizen participation platform. The main objective of this study is to effectively develop and implement BOSESKO using carefully chosen and appropriate technology. Through this, it aims to promote public participation, facilitate the expression of thoughts and sentiments on the implementation of Universal Access to Quality Tertiary Education (UAQTE) or other domains, and offer possibilities for knowledge-sharing.

II. Related works

The core of an e-participation system is typically through websites, an information and communication technology (ICT) infrastructure. This website is easily accessible on PCs, laptops, netbooks, tablets, and smartphones and offers constant internet connectivity. It is beneficial to create applications that can be downloaded on smartphones and tablets in order to improve information accessibility [4]. E-participation has developed into a useful collection of tools that closes the digital gap between citizens and their governments by utilizing developments in ICT [3], and it significantly contributes to the evolution of the political landscape [5]. Citizen involvement is crucial in forming comprehensive national policies, and is a pivotal sign of a successful democratic government [6], [3].

System integration is the process of assessing and deconstructing a system's needs, assigning each one to a separate component, including software, hardware, or other systems [7], and seamlessly integrating these components to meet the system's objectives [8]. In a study by [9], integrated information systems in HEIs led to improved staff management and student workload efficiency. This integration streamlines daily tasks, enhances data processing accuracy and speed, and facilitates informed decision-making, especially concerning student support and reducing delays.

Multi-platform programs, commonly known as cross-platform applications, have completely changed the information technology industry. Applications that can run on all operating systems [10] and communication devices, including Android, Blackberry, and iOS mobile phones, are known as multi-platform applications [11]. In this field, frameworks such as Flutter and React Native have become effective tools. Using a single codebase, developers can produce visually appealing applications for mobile, web, and desktops using Google's Flutter [12], [13]. React Native combines web technologies with native smartphone programming to produce quick and responsive applications [14]. The advantages of multi-platform technology as a learning aid are highlighted in the study of [10], which shows how it lowers costs and improves accessibility for students. The importance of responsive web design (RWD) has become increasingly important as more people view websites using different devices. A responsive website dynamically adjusts depending on the technology and device used [15], [16]. The importance of device responsiveness in increasing user experience has been emphasized [17], particularly given the pervasive use of mobile devices for website access. In an effort to reduce performance trade-offs when used independently, hybrid software development, a combination of coordinated techniques and processes, has become popular [18], [19]. The effective fusion of Scrum and Extreme Programming (XP) into a single framework is known as ScrumXP [20]. The advantages of Scrum and XP are combined in this agile methodology, enabling organizations to continually improve through efficient technical techniques and customer-centric strategies. Pair programming and code refactoring, two essential engineering XP methods, improve both the speed and quality of software development [21].

Several studies have provided insights into the integration and deployment of digital systems, notably platforms for public participation. Cristobal et al. [22] developed an e-participation mobile application with system features, aiding district administrators and barangay officials with specific modules for effective communications and incident management. Llovido and Palaoag [23] developed an e-participatory platform for disaster risk reduction and management, making it easier to report incidents and provide feedback. In order to efficiently handle data processing and display visual representations of equipment management, a previous study [24] developed a platform with a focus on data visualization features. Microservice architecture supports modularity, scalability, and maintainability, providing real-time access to crucial data for administrators and users to make educated decisions about equipment management.

III. Method

In congruence with the approach employed in the study by [25], the researchers utilized a hybrid methodology as shown in Figure 1, combining the best practices and strengths of Extreme Programming (XP) and Scrum to address user requirements. Scrum and XP are effective methodologies for assisting software teams in their projects. In certain instances, these two strategies complement each other. While Scrum focuses on project management, XP provides useful software development practices [26]. Both approaches were determined to be appropriate for the design and development of integrated systems.

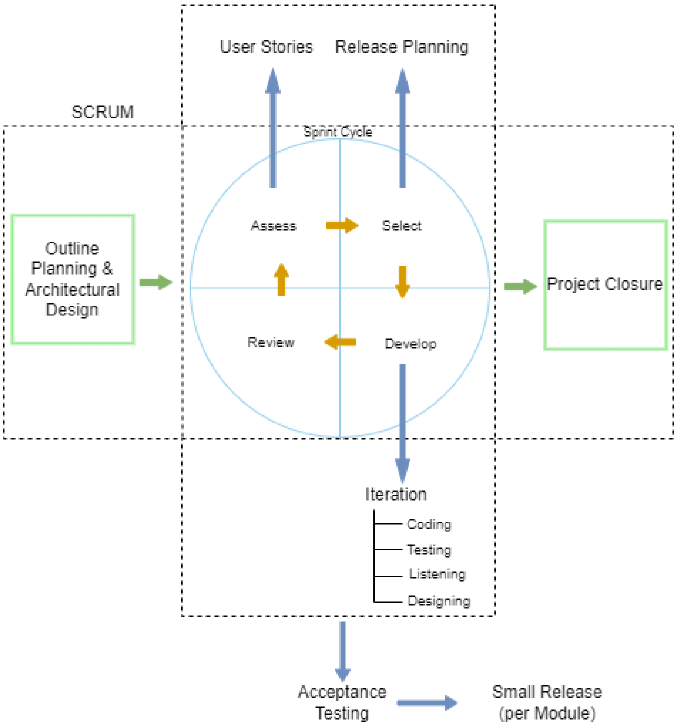


Figure 1. Scrum and XP Methodology

A. Outline Planning and Architectural Design

The main components of the BOSESKO toolkit were subjected to unit testing. Before proceeding to the next feature, each piece of code written for each class was tested. The system testing runs were documented.

The researchers outlined the precise goals of the proposed BOSESKO toolkit. The involved sectors were the subject of extensive efforts to establish collaborations and hold meetings and field visits. The preliminary interview results, along with other crucial inputs, provided insightful information and guidance, highlighting the value of digital public participation. These contributions emphasize the factors that should be taken into consideration, particularly in the design and development of the system.

B. Sprint Cycle

In sprint cycles, the system progresses incrementally during each cycle. The four phases of a sprint are assess, select, develop, and review. This displays logs from the time when each sprint cycle is finished, implying the availability of an updated system version for evaluation. The project's backlog, a list of tasks required to be completed, serves as the basis for planning. It is a collection of prioritized features that are presented as user stories. The stakeholders' suggestions made during the exploratory meetings served as a foundation for the researchers as they established the goals and scope of their project. This aids the researchers in grasping the various system components that require their focus. The selection phase involved a Focus Group Discussion (FGD) in which participants were divided into beneficiary and implementer groups to select features and functionality that would be developed during the sprint. Researchers have integrated user requirements and specifications into technological designs, to establish functionality. A use-case design was created. Partially outlined system features served as a guide for the design of the system's processes and database. Subsequently, a release schedule was established. Jira was used to track all work-related activities to ensure the timely deployment of the system.

1) ***Develop*****:** Coding, testing, listening, and designing undergo iterations at this stage. It is necessary to have a well-designed system architecture that enables seamless interaction between various platforms and identification of connectivity for particular component features. To create a cohesive system structure, a use case diagram—a visual representation of the connections and interactions between various actors and modules in the system was employed. The use case diagram can be used to guide the application of its features and functionalities.

2) ***Coding:*** In this study, cutting-edge front-end technologies such as TypeScript, React, Next.js were used to produce an interactive user experience. Nest.js was used on the backend because it provides functional and scalable features. Chart.js and TailwindCSS were utilized to improve the design aesthetics and data visualization. VSCode and GitHub are essential tools for code editing and management. Postman was used to thoroughly test and validate the APIs to guarantee top performance. MySQL, HeidiSQL, and TypeORM were used to manage and interact with data. In parallel, Android Studio and Flutter were used for the mobile app to produce versatile and high-performance applications.

3) ***Testing:*** The main components of the BOSESKO toolkit were subjected to unit testing. Before proceeding to the next feature, each piece of code written for each class was tested. The system testing runs were documented.

4) ***Listening:*** Series of meetings with stakeholders were conducted to create and evaluate the questionnaire to be implemented in the toolkit. Based on the comments and suggestions from stakeholders, changes to the system were made. To ensure that these recommendations were carried out systematically until all specified system functionalities were fully realized, an incremental method was adopted.

5) ***Designing:*** The design of the toolkit in the various platforms prioritizes user-centered principles, incorporating feedback from technical experts and adhering to accessibility standards. It employs a visually appealing and responsive interface, ensuring a seamless experience across devices. Scalability and inclusivity were central considerations, with plans for ongoing improvements and user-driven refinements.

6) ***Review:*** The completed work was examined and delivered to the stakeholders at the end of the sprint. Subsequently, the following sprint cycle was started.

7) ***Acceptance Testing:*** The BOSESKO toolkit underwent an initial evaluation by researchers and users, as well as a thorough technical validation process. The system is functional, but evaluators noted that further work is needed to improve its usability and performance. The evaluation of the toolkit is still in progress to ensure that all modules are properly examined to meet the highest standards of functionality and user experience.

8) ***Small Release:*** For every essential feature developed, small releases were presented to stakeholders to collect feedback for further refinements in future releases.

9) ***Project Closure:*** A user manual, brochure, and video tutorial were all created to fully describe the features and functionalities of the system. These would guide the respective users of the toolkit.

IV. Results and Discussion

A. BOSESKO System Functions

BOSESKO is composed of software components available in web and mobile applications. It is a comprehensive data-gathering, analysis, and visualization system, making it a synoptic software. The proponents used multiple programming tools and techniques to develop the software system. Software integration played a vital role in development.

The mobile app is developed using Flutter with BLOC framework. SQLite serves as the local database of the mobile app. NodeJS was utilized for the backend development of the web app, while React.JS was for the front-end development and visualization. MYSQL is the central database of the web app. Additionally, the synchronization of the mobile and web applications is made possible using the REST application programming interface (API). Python is the primary tool for the data miner / NLP component. Microservices technique prepared the survey raw data for the data miner program and performed data cleaning. The web application and data miner integration is done using the FAST application programming interface (API). The following are specific system functions.

B. Web App

The web application is hosted in the cloud server with access to the database server, which serves as the primary data repository of the system. This application provides a responsive layout for multiple platforms and browsers. The following are its specific modules

1) Administrator Dashboard. This module allows administrators to manage different software system functions. This page, as demonstrated in figure 2, serves as the backend manager of the system accessible in the web application.
2) Access Manager. This module manages the system accounts, including their types, roles, and privileges. Figure 3 demonstrates that with the access manager, permissions can be read, created, updated, or deleted for a specific system module.

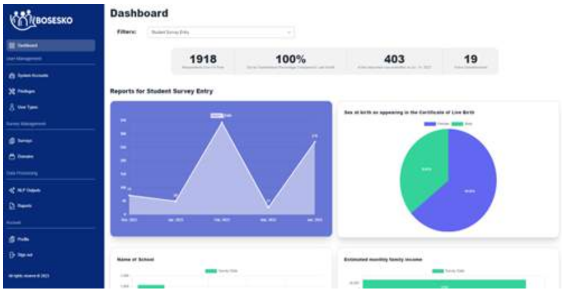
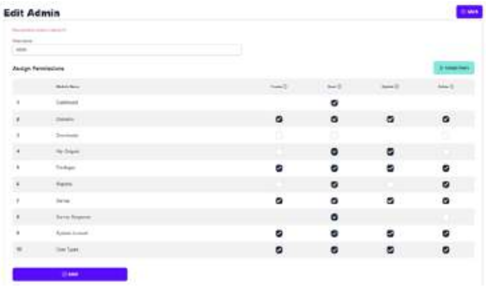

BOSESKO
Dashboard
Filters:
1918
100%
403
19
Reports for Student Survey Entry
Sex at birth as appearing in the Certificate of Live Birth
Name of School
Estimated monthly family income

Figure 2. Administrator Dashboard

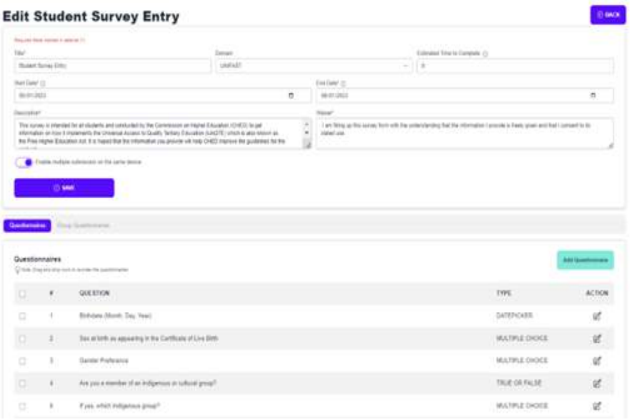


Edit Admin

Assign Permissions

Figure 3. Access Manager

3) Domain Manager. BOSESKO is designed to allow customization, creation, and implementation of the toolkit for other e-participation domains, not exclusively for DRRM and UAQTE only. Figure 4 depicts the user interface for the mentioned purpose.
4) Survey Manager. This module allows creating, updating, deleting, activating, and deactivating surveys within an active domain. Users can specify the survey's description and start and end dates, as shown in Figure 5.
5) Question Manager. Figure 6 is the question manager used to administer questions of an active survey. Question types include multiple choice, open-ended, true or false, drop-down selection, date picker, and rating. Each type has its settings and configuration, like the required answer, audio recording, graph visibility, multiple answers, other answers, maximum and minimum rating values, and follow-up questions.

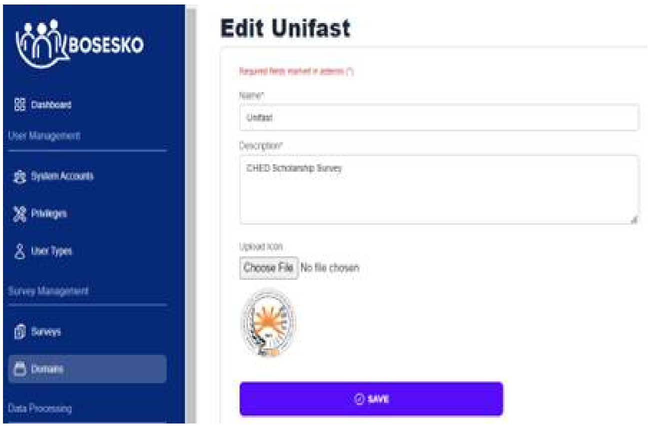


Edit Student Survey Entry

Questionnaires

#	QUESTION	TYPE	ACTION
1	Birthdate (Month, Day, Year)	DATEPICKER	
2	Sex at birth as appearing in the Certificate of Live Birth	MULTIPLE CHOICE	
3	Gender Preference	MULTIPLE CHOICE	
4	Are you a member of an indigenous or cultural group?	TRUE OR FALSE	
5	If yes, which indigenous group?	MULTIPLE CHOICE	

Figure 4. Domain Manager



Edit Unifast

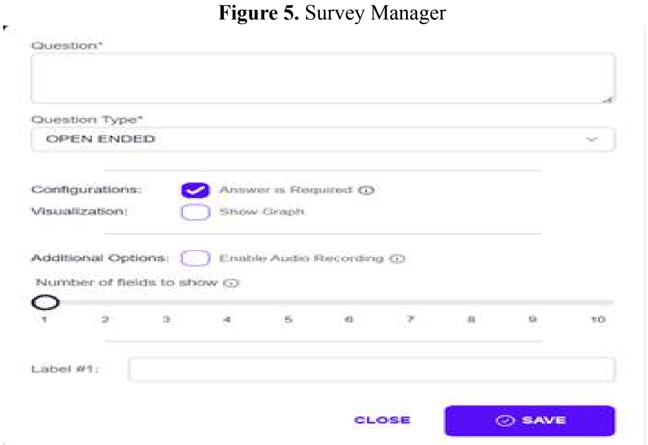
Required fields marked in asterisk (*)

Name*: Unifast

Description*: CHED Scholarship Survey

Upload Icon: No file chosen

Figure 5. Survey Manager



Question*

Question Type*: OPEN ENDED

Configurations: ☑ Answer is Required

Visualization: □ Show Graph

Additional Options: □ Enable Audio Recording

Number of fields to show

Label #1:

Figure 6. Question Manager

6) Data Miner: NLP Toolkit. As shown in Figure 7, this module performs data searching, cleaning implemented using microservices, and analysis of survey responses via text or voice recording to identify patterns and extract useful information. The toolkit implements pre-processing, feature extraction, model training, and performance evaluation.

7) Data Visualizer. The Data Visualizer (Figure 8) is a report-generation tool that creates data visualization of survey responses. It displays a storyboard, charts, graphs, and other visual representations.

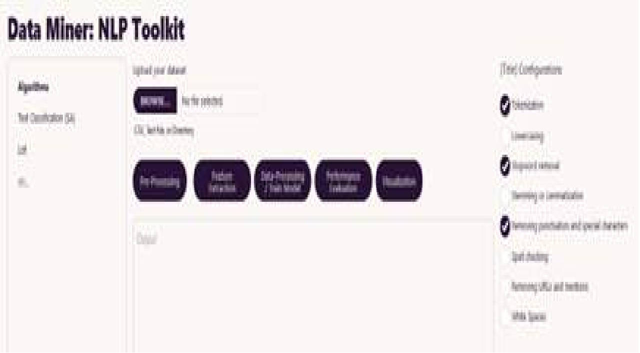


Figure 7. Data Miner

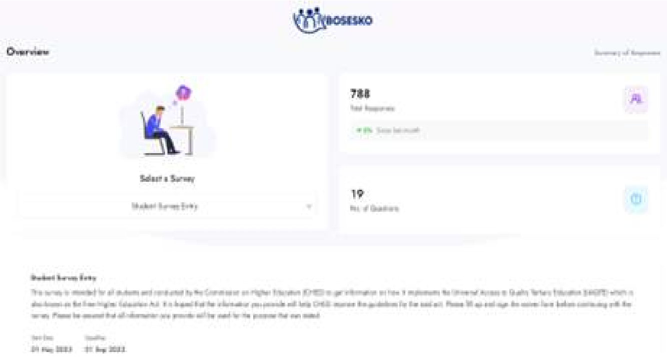
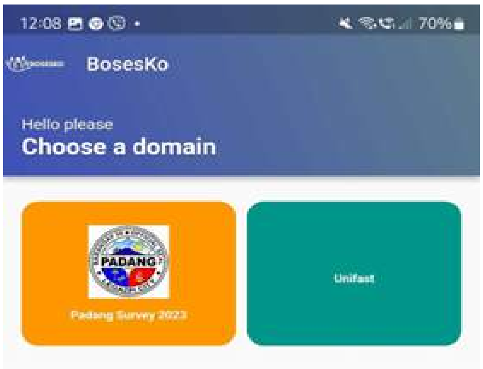


Figure 8. Data Visualizer

C. Mobile App

This mobile app is a tool for data gathering. It is compatible with Android and IOS devices. It has a local database that temporarily holds the user's data and is synchronized to the web application's database server when a stable WAN network is available. The following are the modules of the mobile application.

1) Domain/Survey Selector. This module (Figure 9) allows the respondents to choose their preferred domain and survey. The system can activate multiple domains with multiple surveys in it. This module is also accessible using the web application with a responsive layout.

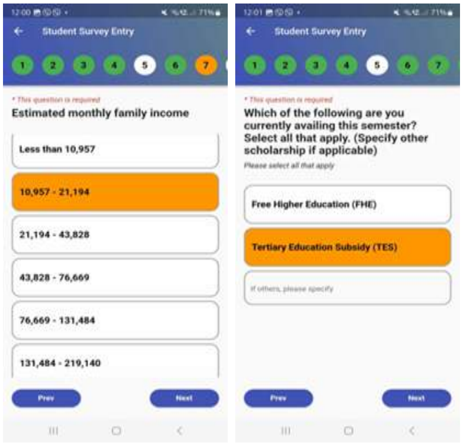


Hello please
Choose a domain

Padang Survey 2023

Figure 9. Domain / Survey Selector - Mobile App

2) Response Uploader. This module allows the respondents to actively respond using text input or voice recording. Respondents can also engage in the active survey using the web application. Figure 10, Figure 11, and Figure 12 show these features.



Student Survey Entry

* This question is required

Estimated monthly family income

Less than 10,957

10,957 - 21,194

21,194 - 43,828

43,828 - 76,669

76,669 - 131,484

131,484 - 219,140

Student Survey Entry

* This question is required

Which of the following are you currently availing this semester? Select all that apply. (Specify other scholarship if applicable)

Please select all that apply

Free Higher Education (FHE)

Tertiary Education Subsidy (TES)

If others, please specify

Figure 10. Response Uploader for Mobile App

Figure 11. Response Uploader for Web App

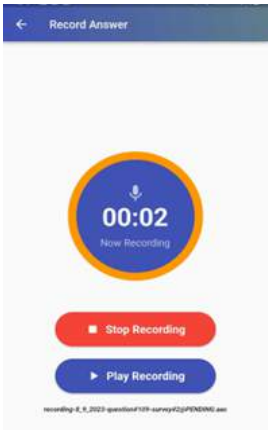


Figure 12. Response Audio Recorder for Mobile App

C. Evaluation

Table 2 presents the performance of BOSESKO's web application and mobile application across various functional aspects. The evaluation uses the ISO 25010 software product functional suitability test. The average scores indicate that BOSESKO excels in functional completeness, with a perfect score of 5.00 for the web application and a strong 4.33 for the mobile application. This suggests that the web application is more feature-rich and comprehensive than its mobile counterpart. Furthermore, both versions of BOSESKO exhibit high levels of functional correctness, scoring 4.67 for the web application and 4.33 for the mobile application. This indicates that the functionality of the applications is generally reliable and free from significant errors, with the web application being more consistent in this regard. Regarding functional appropriateness, both the web and mobile applications of BOSESKO received strong scores of 4.67. This suggests that BOSESKO's functionality, whether accessed through a web browser or a mobile device, is well-suited to its intended purpose.

Table 2. Functional Suitability Evaluation result for BOSESKO application

Quality Characteristic	**BOSESKO Web application**	**BOSESKO mobile application**
Functional completeness	5.00	4.33
Functional correctness	4.67	4.33
Functional appropriateness	4.67	4.67
Weighted Mean	**4.78**	**4.40**

V. Conclusion

BOSESKO, a synoptic digital participatory toolkit, has been developed to engage communities in governance by utilizing technology-driven approaches. With its multilingual and inclusive features, it has the potential to play a vital role in gathering public feedback on critical issues such as the implementation of Universal Access to Quality Tertiary Education (UAQTE) in the Philippines and disaster preparedness, providing a convenient channel for public participation. This study employs state-of-the-art technologies to build web and mobile applications, successfully creating a reliable, universally applicable, and effective digital platform for public involvement through a hybrid methodology.

The web and mobile application functions of BOSESKO give users the freedom to choose domains, take surveys, and offer feedback, improving citizen involvement and functionality. Based on the functional suitability test of software product quality evaluation, BOSESKO's web application obtained 4.78; on the mobile application, it perceived a 4.40 total weighted mean. Survey management is highly emphasized on the platform, leading to better data handling and user control for improved usability. Furthermore, BOSESKO is a multiplatform application that considerably increases its reach by making it accessible to all citizens. Additionally, it guarantees an agile and consistent user experience across all platforms, improving functionality, usability, and engagement. This study highlights the superior feature set of BOSESKO in comparison to other survey applications. The development and implementation of BOSESKO show the potential of online platforms for public participation in promoting inclusive and democratic decision-making. BOSESKO has the potential to close the gap between citizens and governments by utilizing technology, data mining techniques, and user feedback, enabling engagement and fostering transparency and accountability. Future work includes determining the level of user acceptance of the toolkit and understanding the relationship of each factor that could affect its utilization.

Acknowledgment

The researchers would like to thank the Philippine Commission on Higher Education (CHED) Leading the Advancement of Knowledge in Agriculture and Science (LAKAS) Project No. 2021-007, eParticipation 2.1: Harnessing Natural Language Processing (NLP) for Community Participation, for providing the necessary funds to make this research possible. The researchers are truly grateful for their contribution to the research.

References

[1] S. Leible, S. Götz, K. Meyer-Lüters, M. Ludzay, T. Kaufmann, and M. N. Tran, "ICT Application Types and Equality of E-Participation - A Systematic Literature Review," *ResearchGate* , Jul. 2022. https://www.researchgate.net/profile/Stephan-Leible/publication/361907528_ICT_Application_Types_and_Equality_of_E-Participation_-_A_Systematic_Literature_Review/links/62cc1b20cab7ba7426e4bd6b/ICT-Application-Types-and-Equality-of-E-Participation-A-Systematic-Literature-Review.pdf

[2] F. L. Benítez-Martínez, M. V. Hurtado-Torres, and E. Romero-Frías, "A neural blockchain for a tokenizable e-Participation model," *Neurocomputing*, vol. 423, pp. 703–712, Jan. 2021, doi: 10.1016/j.neucom.2020.03.116.

[3] H. A. Manaf and M. N. S. Man, "Mobile Application and Web 2.0 as an E-Participation Mechanism: A Literature Analysis," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 15, no. 06, p. 185, Mar. 2021, doi: 10.3991/ijim.v15i06.20673.

[4] A. Setyono, L. B. Handoko, Purwanto, A. Salam, E. Noersasangko, and D. E. Waluyo, "Development of Mobile e-Participation System to Enhance e-Government Performance," in *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Sep. 2019. Accessed: Jun. 30, 2023. [Online]. Available: http://dx.doi.org/10.1109/isemantic.2019.8884221

[5] K. B. V. Salvio, “Extending the Evaluation on Philippine E-Government Services on its Accessibility for Disabled Person,” in *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, Jul. 2020. Accessed: Dec. 15, 2023. [Online]. Available: http://dx.doi.org/10.1109/worlds450073.2020.9210374

[6] N. H. Basri, W. A. W. Adnan, and H. Baharin, “An Exploratory Study of Users’ Experiences with e-participation: a Case Study of Malaysia,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 3, p. 1138, Sep. 2019, doi: https://doi.org/10.11591/ijeecs.v15.i3.pp1138-1143.

[7] M. Sanchez, E. Exposito, and J. Aguilar, “Industry 4.0: survey from a system integration perspective,” *International Journal of Computer Integrated Manufacturing*, vol. 33, no. 10–11, pp. 1017–1041, Jun. 2020, doi: 10.1080/0951192x.2020.1775295.

[8] N. Kishi *et al.*, “Methodology and Organizational Design to Realize the System Integration Necessary for the Development of Commercial Aircraft,” *TRANSACTIONS OF THE JAPAN SOCIETY FOR AERONAUTICAL AND SPACE SCIENCES, AEROSPACE TECHNOLOGY JAPAN*, vol. 19, no. 1, pp. 1–8, 2021, doi: 10.2322/tastj.19.1.

[9] A. Kayanda, “Information systems integration for better decision making in Tanzanian Higher Education Institutions,” *International Journal of Education and Development using Information and Communication Technology (IJEDICT)*, vol. 18, no. 2, pp. 163–176, 2022, Available: https://files.eric.ed.gov/fulltext/EJ1359977.pdf

[10] F. Rini, Y. Yelfiza, and A. Y. Pernanda, “New Lms Mobile Framework Based on Multiplatform: A Literature Review of Mobile Lms Theory, Design and Implementation,” *INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND ANALYSIS*, vol. 06, no. 06, pp. 2557–2563, Jun. 2023, doi: 10.47191/ijmra/v6-i6-54.

[11] G. G. Mustofa, B. Mulyanti, and I. Widiaty, “The implementation of multi-platform technology,” *IOP Conference Series: Materials Science and Engineering*, vol. 1098, no. 2, p. 022112, Mar. 2021, doi: 10.1088/1757-899x/1098/2/022112.

[12] S. Y. Ameen and D. Y. Mohammed, “Developing Cross-Platform Library Using Flutter,” *European Journal of Engineering and Technology Research*, vol. 7, no. 2, pp. 18–21, Mar. 2022, doi: https://doi.org/10.24018/ejeng.2022.7.2.2740.

[13] Oscar Danilo Gavilánez Alvarez *et al.*, “Comparative Analysis of Cross-Platform Frameworks,” *Journal of Namibian Studies : History Politics Culture*, vol. 33, May 2023, doi: 10.59670/jns.v33i.874.

[14] A. Fentaw, “Cross platform mobile application development: a comparison study of React Native Vs Flutter,” 2020. Available: https://jyx.jyu.fi/bitstream/handle/123456789/70969/1/URN%3ANBN%3Afi%3Ajyu-202006295155.pdf

[15] N. O. Adelakun, B. A. Olanipekun, and S. A. Bakinde, “Easy Approach to a Responsive Website Design Using Artisteer Application Software,” *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3579919.

[16] I. Althomali, G. M. Kapfhammer, and P. McMinn, “Automated visual classification of DOM‐based presentation failure reports for responsive web pages,” *Software Testing, Verification and Reliability*, vol. 31, no. 4, Feb. 2021, doi: 10.1002/stvr.1756.

[17] K. A. Sathiyanathan, “SOURCEGEN : Mobile responsive source code generator,” *dlib.iit.ac.lk*, 2021. http://dlib.iit.ac.lk/xmlui/handle/123456789/875 (accessed Jul. 19, 2023).

[18] I. K. Kirpitsas and T. P. Pachidis, “Evolution towards Hybrid Software Development Methods and Information Systems Audit Challenges,” *Software*, vol. 1, no. 3, pp. 316–363, Aug. 2022, doi: 10.3390/software1030015.

[19] B. Bose, N. T. Khan, Sumaiya Ashreen, Faishal Ahmed Shuvo, Md. Mazid-Ul-Haque, and A. Bhowmik, “Hybrid Scrum-XP: A Proposed Model based on Effectiveness of Agile Model on Varieties of Software Companies in Bangladesh,” *AIUB Journal of Science and Engineering (AJSE)*, vol. 22, no. 1, pp. 35–44, May 2023, doi: 10.53799/ajse.v22i1.353.

[20] F. Fuior, “Key elements for the success of the most popular Agile methods,” *Revista Română de Informatică şi Automatică*, vol. 29, no. 4, pp. 7–16, Dec. 2019, doi: 10.33436/v29i4y201901.

[21] N. Koceska and S. Koceski, “Hybrid project management as a new form of project management,” Dec. 2022. Accessed: Jul. 19, 2023. [Online]. Available: https://eprints.ugd.edu.mk/31403/1/Hybrid%20project%20management.pdf

[22] I. A. A. Cristobal *et al.*, “Pasigueño Assistant: An E-Participation Mobile Application Framework for the City of Pasig, Philippines,” in *TENCON 2018 - 2018 IEEE Region 10 Conference*, Oct. 2018. Accessed: Jul. 19, 2023. [Online]. Available: http://dx.doi.org/10.1109/tencon.2018.8650177

[23] J. L. Llovido and T. D. Palaoag, “e-LAHOK: An e-Participatory Platform for Disaster Risk Reduction and Management,” *IOP Conference Series: Materials Science and Engineering*, vol. 803, no. 1, p. 012049, Apr. 2020, doi: 10.1088/1757-899x/803/1/012049.

[24] T. Lv, J. Zhang, and Y. Chen, “The Research on Data Acquisition and Analysis Platform for Lathe Machine based on Stream Computing,” *Journal of Physics: Conference Series*, vol. 1650, no. 3, p. 032060, Oct. 2020, doi: 10.1088/1742-6596/1650/3/032060.

[25] L. L. Maceda, J. L. Llovido, and J. E. Serrano, “System Design for Disaster Risk Damage Assessment,” in *2018 International Symposium on Computer, Consumer and Control (IS3C)*, Dec. 2018. Accessed: Jul. 20, 2023. [Online]. Available: http://dx.doi.org/10.1109/is3c.2018.00069

[26] M. Afshari and T. Javdani Gandomani, “A novel risk management model in the Scrum and extreme programming hybrid methodology,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 3, p. 2911, Jun. 2022, doi: 10.11591/ijece.v12i3.pp2911-2921.

Jennifer L. Llovido is a faculty member of the Computer Science and Information Technology Department at Bicol University College of Science, Legazpi City, Philippines, with an academic rank of Associate Professor V. She completed her Doctor in Information Technology (DIT) at the University of the Cordilleras, Baguio City, Philippines. Her published research works are centered on the fields of natural language processing, data mining, and system design and development. She can be reached at jllovido@bicol-u.edu.ph.

Michael Angelo D. Brogada is an Associate Professor at the College of Science of Bicol University-Main Campus, Legazpi City. He is managing a software development company, MAB Business Solutions, which has developed software applications and maintained computer networks and servers for businesses since 2011. He finished his doctorate in Information Technology at the Technological Institute of the Philippines. He passed certifications in IT, such as IBM DB2 Academic Associate and DICT – EDP Specialist in Computer Programming. His research interests include IT Protection and Security, Data Mining, Web Applications, and Cloud Computing. He can be reached at madbrogada@bicol-u.edu.ph.

Lany L. Maceda earned her Doctorate in Information Technology from University of the Cordilleras, Baguio City, Philippines, in 2020. She is a faculty member of the Department of Computer Science and Information Technology, holding an academic rank of Associate Professor V at Bicol University. Moreover, she also serves as the Director of the Research, Development and Management Division at the same institution. She has been actively promoting data-driven policy-making through her research papers published in reputable international journals and conferences with research interests on machine learning particularly on natural language processing and data mining. She can be reached at llmaceda@bicol-u.edu.ph.

Floradel S. Relucio is an Associate Professor I at Bicol University Polangui under the Computer Studies Department and is the current college research coordinator. She earned her Doctor in Information Technology (DIT) degree at the University of the Cordilleras, Baguio City, Philippines. The focal points of her research studies lie in the domains of natural language processing, system design and development, and the Internet of Things. She can be reached at fsrelucio@bicol-u.edu.ph.

Lea D. Austero is an Assistant Professor III at the Department of Computer Science and Information Technology at the College of Science of Bicol University in Legazpi City, Albay. She teaches computer programming and associated topics for the Bachelor of Science in Computer Science, Bachelor of Science in Information Technology, and Master of Science in Information Systems programs. Her published works include Determining resource capacity in disaster assistance using a model-driven decision support system; Discovering themes from internet news articles on the 2018 Mount Mayon Eruption; and Solving course timetabling problem using Whale Optimization Algorithm. These works may be found in IEEE and Scopus. She can be reached at ldaustero@bicol-u.edu.ph.

Mideth B. Abisado is an Associate Member of the National Research Council of the Philippines and a Board Member of the Computing Society of the Philippines Special Interest Group for Women in Computing. She is the Director of the CCIT Graduate Programs. She completed her Doctor in Information Technology (DIT) at the Technological Institute of the Philippines. Her research focuses on Empathic Computing, Social Computing, Human-Computer Interaction, and Human Language Technology. She can be reached at mbabisado@national-u.edu.ph.

An IoT-Based Early Warning System for Settlement Monitoring Using Differential Pressure Static Level

Tieyan CHAO*, Hui LIANG*, **, Yuwei GE*, **, Kai HOU***, Xiang DONG****, Ting PENG****

* Shaanxi Huashan Road and Bridge Group Co., LTD. Xi'an 710016

** Shaanxi Zhengcheng Road and Bridge Engineering Research Institute Co., LTD. Xi'an 712000

*** Sichuan Chuanjiao Road and Bridge Co.,LTD. Sichuan 618300

**** Chang'an University, Xi'an 710064

33886822@qq.com, 5936001270@qq.com, 435748864@qq.com, 75856688@qq.com, 1290740787@qq.com, t.peng@ieee.org

Abstract— **This paper presents an IoT-based automated settlement monitoring system that aims to meet the diverse requirements of applications and the increasing demand for settlement monitoring. The system consists of a perception subsystem, data transmission layer, IoT cloud platform, and application terminal. The differential pressure static level sensor transmits the sensor serial port raw data to the 4G DTU through the RS-485 bus interface. Then, the 4G DTU converts it into a 4G network and transmits the data to the IoT cloud platform via the MQTT protocol. The IoT cloud platform analyzes and processes the collected data to generate reports, visualize data to generate curves, and perform real-time anomaly identification on the data. Finally, it implements viewing of settlement monitoring data and curve changes, as well as monitoring status warnings at the application terminal. The practical engineering application results show that this system can provide effective safety monitoring and an early warning scheme for slope, tunnel, bridge, and building safety monitoring.**

Keywords— **IoT, Differential Pressure Static Level, Settlement Monitoring, MQTT, Early Warning System, Cloud Platform**

I. Introduction

At present, China's highway construction mainly focuses on the central and western regions, most of which are mountainous areas. When structures are built in areas with harsh environmental conditions, higher requirements are placed on the timeliness and accuracy of settlement monitoring. Existing conventional monitoring methods have limited applicability, poor monitoring accuracy, are easily affected by human and environmental factors, and cannot meet the needs of automated and intelligent management of monitoring data. For instance, the interferometric synthetic aperture radar (InSAR) technology is possible to use at any time and in any atmospheric condition, but the technical implementation is difficult and the spatial and temporal resolution is low [1]. The global positioning system (GPS) can measure in a highly automated manner and does not require a direct line of sight between stations, making it cost-effective. However, vegetation cover, buildings, or other types of infrastructure can cause signal scattering, which can reduce GPS accuracy [1]. The Global Navigation Satellite System (GNSS) can achieve millimeter-level positioning [2], but the cost of monitoring equipment and the requirements for environmental use are high [3]. The settlement information monitored manually is both expensive and limited. At the same time, these records exhibit lagging effects and cannot provide real-time instructions for on-site construction [4].

With the continuous advancement of Internet of Things (IoT) and cloud platform technology, the communication network architecture design in the field of settlement monitoring has been further optimized and improved. Meanwhile, due to the advantages of the differential pressure static level, such as its simple structure, low cost, strong environmental adaptability, high monitoring accuracy, high degree of automation, and the ability to realize all-weather high-frequency and long-time settlement monitoring [3, 5], it has been widely used in the safety monitoring work of various types of engineering and architectural bodies, such as subgrade, foundations, bridges, and tunnels.

In summary, in order to meet the diverse needs of applications and the increasing demands for settlement monitoring, and to achieve automated management of structural settlement monitoring, this paper designs and develops an automated settlement monitoring and early warning system based on differential pressure static level and Internet of Things technology.

II. Design of Settlement Monitoring System

The IoT-based automated settlement monitoring and early warning system designed and developed in this study consists of the perception subsystem, data transmission layer, IoT cloud platform, and application terminal from the bottom up. The overall architecture of the IoT-based automated settlement monitoring and early warning system is shown in **Figure 1**.

A. Overall System Design

The sensor (differential pressure static level) transmits the raw data from the sensor serial port to the 4G DTU through RS-485 bus interface, then the 4G DTU converts it to the 4G network and transmits the data to the IoT cloud platform through the MQTT protocol, and the IoT cloud platform analyzes and processes the collected data to generate reports, generates curves by visualization processing, and identifies real-time anomalies of the data. Finally, the application terminal realizes the view of settlement monitoring data and curve changes and the warning of monitoring status.

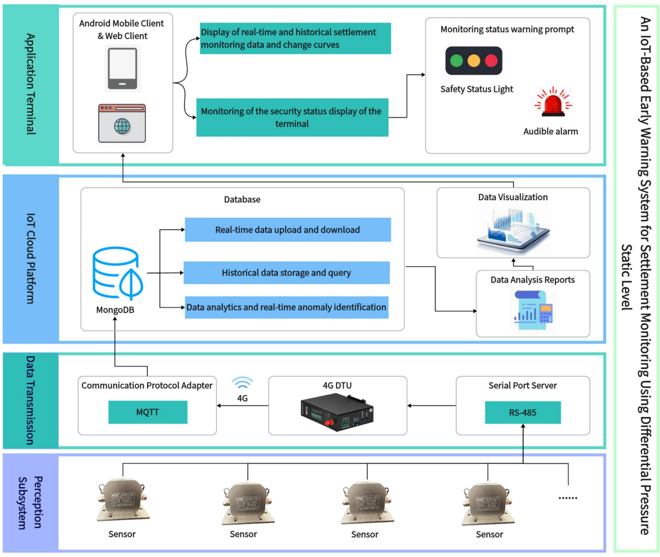

Application Terminal
Android Mobile Client & Web Client
Display of real-time and historical settlement monitoring data and change curves
Monitoring of the security status display of the terminal
Monitoring status warning prompt
Safety Status Light
Audible alarm
IoT Cloud Platform
Database
Real-time data upload and download
Historical data storage and query
Data analytics and real-time anomaly identification
MongoDB
Data Visualization
Data Analysis Reports
Data Transmission
Communication Protocol Adapter
MQTT
4G
4G DTU
Serial Port Server
RS-485
Perception Subsystem
Sensor
Sensor
Sensor
Sensor
......
An IoT-Based Early Warning System for Settlement Monitoring Using Differential Pressure Static Level

Figure 1. Overall architecture of IoT-based automated settlement monitoring and early warning system

B. *Perception Subsystem*

The perception subsystem mainly consists of the differential pressure static level to accomplish settlement data acquisition. The pressure differential static level used in the system is model PTH-SZY100, which features small size, high accuracy, and wide applicability. In the actual settlement monitoring process, it can meet the monitoring requirements of different temperature regions, achieve automatic temperature compensation, and has ultra-short response time, ultra-high monitoring resolution, and effective waterproofing capabilities. The appearance of PTH-SZY100 differential pressure static level is shown in **Figure 2** and its technical parameters are listed in **Table 1**.

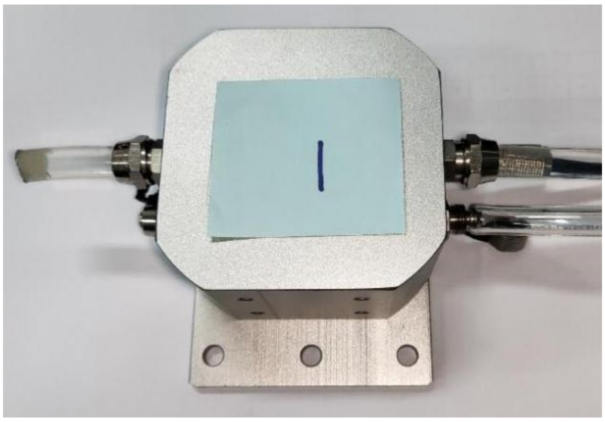


Figure 2. PTH-SZY100 differential pressure static level appearance

TABLE 1. PTH-SZY100 DIFFERENTIAL PRESSURE STATIC LEVEL TECHNICAL PARAMETERS

Technical Parameter	Specification
Measuring Medium	Fluid
Pressure Range	0~1mH2O
Working Temperature Range	-40~100℃
Comprehensive Accuracy	0.02~0.1%FS
Display Resolution	0.01mm
Fieldbus Type	RS-485
Power Supply Requirement	12V DC(7-30VDC)
Response Time	≤5ms
Waterproof Rating	IP68

C. *Data Transmission*

The data transmission layer is mainly composed of 4G DTU and the RS-485 bus interface. The 4G DTU obtains the raw data from the serial ports of various sensors through the RS-485 bus interface, converts it into a 4G network for transmission, and adopts the MQTT publish/subscribe messaging protocol [6] to improve data scalability and stability.

D. *IoT Cloud Platform*

In view of the functional requirements of the software side of the settlement monitoring and early warning system, the system component architecture of the IoT cloud platform is designed, in which the main functions of the database include the following three requirements:

1) Real-time Data Uploading and Downloading: Real-time uploading of information entered into the field sensing subsystem and mobile terminal; fault tolerance and fault clearing mechanisms for data collection; real-time docking of monitoring data; and support for exporting and downloading data in the form of an Excel table.

2) Historical Data Storage and Query: Under the premise of ensuring that the monitoring data processing, analysis, and early warning are not affected, the storage capacity of the data is controlled, and users are supported to call and query the historical data of the monitoring targets through the application terminal.

3) Data Analysis: After preprocessing the monitoring data, such as data cleaning, data correction, and real-time anomaly identification, intelligent monitoring and early warning judgments are made, and the results are promptly feedback.

Compared to other databases, if all tests that do not include scanning workloads are considered, which are composed of scanning operations, MongoDB is the database with the best execution time [7]. MongoDB has strong performance, large aggregation capabilities, fast storage of extremely complex data, powerful query capabilities, and support for data replication and failure recovery.

In conclusion, the main framework of the IoT cloud platform is developed using Python language, and uses MongoDB database to store, analyze, process, predict, and judge monitoring data, providing users and administrators with a powerful real-time monitoring visualization, controllable, and queryable monitoring system.

E. Application Terminal

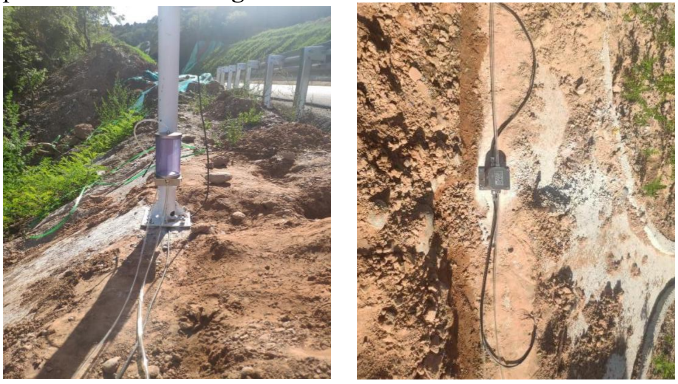
The application terminal mainly includes a web client and an Android mobile phone client, which is a system integrating the functions of monitoring equipment control, data presentation, chart display, warning display, etc., and can truthfully present the information of project settlement monitoring. Through this system, users can understand the real-time status and development trend of monitoring targets over time.

III. Real implementation

In practical engineering applications, this monitoring and early warning system can also work with other sensors. This study focuses on the application of differential pressure static leveling for settlement monitoring.

A. Measurement Point Deployment

After determining the construction program according to the actual situation of the project to carry out the deployment of measurement points, the deployment of measurement points as shown in **Figure 3**.



a) reference point b) monitoring point

Figure 3. Measurement point deployment

B. Monitoring Warning Value Settings and Alerts

After completing the installation of equipment and measurement point deployment, according to the specification and the actual engineering design requirements to set the pre-warning indicators, different warning states will be used to take different emergency measures.

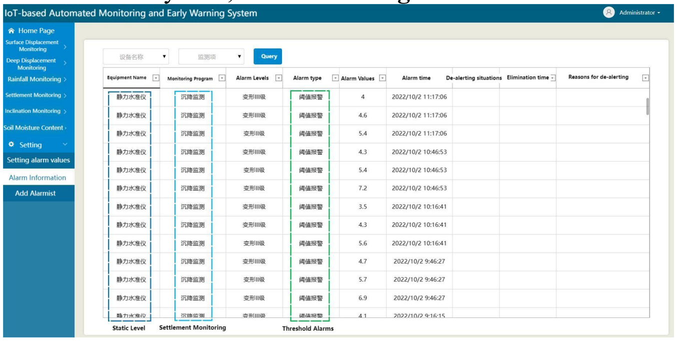
The main forms of early warning prompts include the display of the security status of the monitoring terminal and the status and early warning display of the monitored object.

The monitoring terminal's security status display is a simple and intuitive way to display the security status of the monitored objects in the scheme on each monitoring terminal, aiming to enable each requester to directly grasp the status of the monitored objects they need to view in the shortest possible time. Through the monitoring end of the screen, intuitively view the safety status lights to grasp the safety of the project to be monitored, with green, yellow, orange, and red colors to indicate the different safety status; different colors correspond to different sounds; and the client's color changes while emitting different sounds to warn the monitoring personnel. This allows monitoring personnel to understand and control the status of the monitored object more quickly and directly. The correspondence between the safety status and the safety indicator color is listed in **Table 2**.

TABLE 2. Safety Status Alert

Safety Status Light Color	**Safety Level**	**Treatment Measures**
Red	Very dangerous	Emergency evacuation and remedial measures
Orange	Moderately dangerous	Remediation and elimination of hazardous conditions at the site
Yellow	Slightly dangerous	Strengthen monitoring of the site and take the necessary measures
Green	Safe and stable	-

The status warning display of the monitored object refers to the process of reflecting the monitored dynamic process through the monitoring platform, predicting it to a certain extent, and providing early warning processing for the safety status of the monitored object. When there is a dangerous situation, the monitoring data can be provided according to the needs and conditions on the spot, so as to provide the latest deformation monitoring information and feedback to all parties on the spot. At the same time, the alarm log can be viewed in the system, as shown in **Figure 4.**

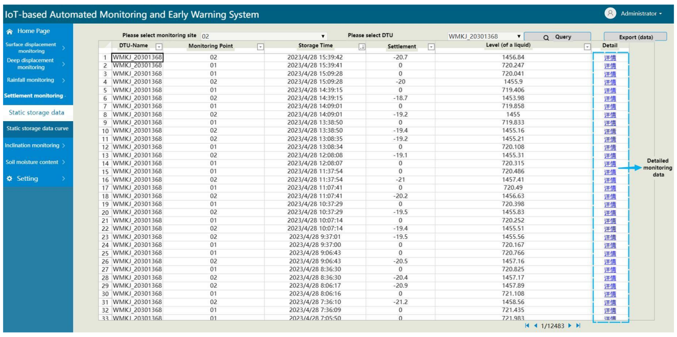


Equipment Name	Monitoring Program	Alarm Levels	Alarm type	Alarm Values	Alarm time	De-alerting situations	Elimination time	Reasons for de-alerting
静力水准仪	沉降监测	变形III级	阈值报警	4	2022/10/2 11:17:06			
静力水准仪	沉降监测	变形III级	阈值报警	4.6	2022/10/2 11:17:06			
静力水准仪	沉降监测	变形II级	阈值报警	5.4	2022/10/2 11:17:06			
静力水准仪	沉降监测	变形III级	阈值报警	4.3	2022/10/2 10:46:53			
静力水准仪	沉降监测	变形II级	阈值报警	5.4	2022/10/2 10:46:53			
静力水准仪	沉降监测	变形II级	阈值报警	7.2	2022/10/2 10:46:53			
静力水准仪	沉降监测	变形III级	阈值报警	3.5	2022/10/2 10:16:41			
静力水准仪	沉降监测	变形III级	阈值报警	4.3	2022/10/2 10:16:41			
静力水准仪	沉降监测	变形II级	阈值报警	5.6	2022/10/2 10:16:41			
静力水准仪	沉降监测	变形III级	阈值报警	4.7	2022/10/2 9:46:27			
静力水准仪	沉降监测	变形II级	阈值报警	5.7	2022/10/2 9:46:27			
静力水准仪	沉降监测	变形II级	阈值报警	6.9	2022/10/2 9:46:27			
静力水准仪	沉降监测	变形III级	阈值报警	4.1	2022/10/2 9:16:15			

Figure 4. Alarm log interface

C. Monitoring Information Query

Users, through the equipment address of this study in the mobile phone or web terminal system, can complete the system login, enter the platform at all levels of the interface to view project information, monitoring information, etc., and at the same time generate a static level settlement curve that can be viewed arbitrarily retrospectively, which can satisfy all parties to know about the project site conditions and to ensure construction safety in a timely manner. As shown in **Figure 5**, the user can query the historical settlement data by selecting the DTU, monitoring point location.



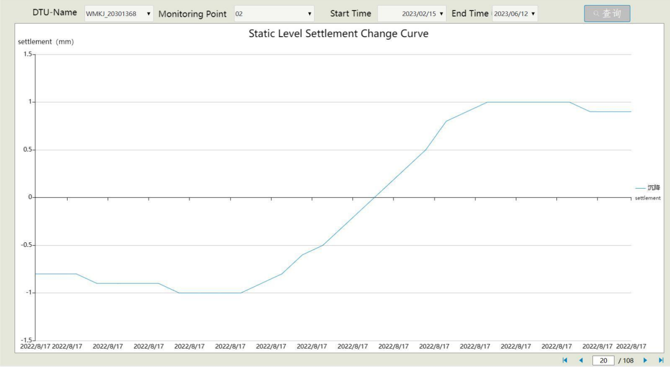
IoT-based Automated Monitoring and Early Warning System

Please select monitoring site 02 Please select DTU WMKJ_20301368

	DTU-Name	Monitoring Point	Storage Time	Settlement	Level (of a liquid)	Detail
1	WMKJ_20301368	02	2023/4/28 15:39:42	-20.7	1456.84	详情
2	WMKJ_20301368	01	2023/4/28 15:39:41	0	720.247	详情
3	WMKJ_20301368	01	2023/4/28 15:09:28	0	720.041	详情
4	WMKJ_20301368	02	2023/4/28 15:09:28	-20	1455.9	详情
5	WMKJ_20301368	01	2023/4/28 14:39:15	0	719.406	详情
6	WMKJ_20301368	02	2023/4/28 14:39:15	-18.7	1453.98	详情
7	WMKJ_20301368	01	2023/4/28 14:09:01	0	719.858	详情
8	WMKJ_20301368	02	2023/4/28 14:09:01	-19.2	1455	详情
9	WMKJ_20301368	01	2023/4/28 13:38:50	0	719.833	详情
10	WMKJ_20301368	02	2023/4/28 13:38:50	-19.4	1455.16	详情
11	WMKJ_20301368	02	2023/4/28 13:08:35	-19.2	1455.21	详情
12	WMKJ_20301368	01	2023/4/28 13:08:34	0	720.108	详情
13	WMKJ_20301368	02	2023/4/28 12:08:08	-19.1	1455.31	详情
14	WMKJ_20301368	01	2023/4/28 12:08:07	0	720.315	详情
15	WMKJ_20301368	01	2023/4/28 11:37:54	0	720.486	详情
16	WMKJ_20301368	02	2023/4/28 11:37:54	-21	1457.41	详情
17	WMKJ_20301368	01	2023/4/28 11:07:41	0	720.49	详情
18	WMKJ_20301368	02	2023/4/28 11:07:41	-20.2	1456.63	详情
19	WMKJ_20301368	01	2023/4/28 10:37:29	0	720.398	详情
20	WMKJ_20301368	02	2023/4/28 10:37:29	-19.5	1455.83	详情
21	WMKJ_20301368	01	2023/4/28 10:07:14	0	720.252	详情
22	WMKJ_20301368	02	2023/4/28 10:07:14	-19.4	1455.51	详情
23	WMKJ_20301368	02	2023/4/28 9:37:01	-19.5	1455.56	详情
24	WMKJ_20301368	01	2023/4/28 9:37:00	0	720.167	详情
25	WMKJ_20301368	01	2023/4/28 9:06:43	0	720.766	详情
26	WMKJ_20301368	02	2023/4/28 9:06:43	-20.5	1457.16	详情
27	WMKJ_20301368	01	2023/4/28 8:36:30	0	720.825	详情
28	WMKJ_20301368	02	2023/4/28 8:36:30	-20.4	1457.17	详情
29	WMKJ_20301368	02	2023/4/28 8:06:17	-20.9	1457.89	详情
30	WMKJ_20301368	01	2023/4/28 8:06:16	0	721.108	详情
31	WMKJ_20301368	02	2023/4/28 7:36:10	-21.2	1458.56	详情
32	WMKJ_20301368	01	2023/4/28 7:36:09	0	721.435	详情
33	WMKJ_20301368	01	2023/4/28 7:05:50	0	721.983	详情

Figure 5. Historical data view interface

Each perception subsystem collects real-time monitoring data from the construction site, which is processed in real time by the IoT cloud platform subsystem and presented in the form of visualized data and curves in the application terminal system. In terms of real-time data display, the monitoring system categorizes and numbers all the perception subsystems in the project and displays the real-time data of each system independently. In this way, users can view the real-time status of each device in a timely manner, as well as quickly select the real-time monitoring results of the perceptual subsystems that need to be focused on, which facilitates the deployment and arrangement of equipment at the key monitoring points of the project. The user can query the historical settlement change curve by selecting the DTU, the monitoring point location, and the start time and end time, as shown in **Figure 6.**



DTU-Name WMKJ_20301368 Monitoring Point 02 Start Time 2023/02/15 End Time 2023/06/12

Figure 6. Static level settlement change curve

IV. CONCLUSIONS

In order to overcome the limitations of existing traditional monitoring methods with limited applicability, low monitoring accuracy, susceptibility to human and environmental factors, and inability to meet the needs of automated and intelligent management of monitoring data, this study developed and designed a set of automatic settlement monitoring and early warning system based on differential pressure level and Internet of Things (IoT) technology, which realised remote, real-time, high-precision monitoring and early warning during the construction and operation of structures under harsh environmental conditions, and was able to provide effective safety monitoring and an early warning scheme for slope, tunnel, bridge, and building safety monitoring.

ACKNOWLEDGMENT

This study was funded by Shaanxi Huashan Road and Bridge Group Co., Ltd. and Sichuan Chuanjiao Road and Bridge Co., Ltd. The author would like to thank Shaanxi Huashan Road and Bridge Group Co., Ltd. and Sichuan Chuanjiao Road and Bridge Co., Ltd. for their support of this research.

REFERENCES

[1] H. Thirugnanam, S. Uhlemann, R. Reghunadh, M. V. Ramesh and V. P. Rangan, "Review of Landslide Monitoring Techniques With IoT Integration Opportunities," IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, vol. 15, pp. 5317-5338, June 2022.

[2] Y. Liu, H. Hazarika, H. Kanaya, O. Takiguchi and D. Rohit, "Landslide prediction based on low-cost and sustainable early warning systems with IoT," BULLETIN OF ENGINEERING GEOLOGY AND THE ENVIRONMENT, vol. 82, May 2023.

[3] L. Wang, P. Zhang, J. Yang, X. Wang and L. Wang, "Design of Settlement Monitoring System Based on Double Datum," PIEZOELECTRICS & ACOUSTOOPTICS, vol. 43, 2021.

[4] P. Zhang, R.-P. Chen, T. Dai, Z.-T. Wang and K. Wu, "An AIoT-based system for real-time monitoring of tunnel construction," TUNNELLING AND UNDERGROUND SPACE TECHNOLOGY, vol. 109, March 2021.

[5] W. Zhang, J. Yang and C. Ma, "Application of Settlement Automatic Collection in Soft Foundation Treatment," Journal of Water Resources and Architectural Engineering, vol. 17, pp. 48-54, 2019.

[6] M. Bender, E. Kirdan, M.-O. Pahl and G. Carle, "Open-Source MQTT Evaluation," in 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), 2021.

[7] I. Carvalho, F. Sa and J. Bernardino, "Performance Evaluation of NoSQL Document Databases: Couchbase, CouchDB, and MongoDB," Algorithms, vol. 16, 2023.

Tieyan CHAO, born in August 1980, is a senior engineer and a first-class construction engineer. He is currently the Safety Director of Shaanxi Huashan Road and Bridge Group Co., Ltd.

Hui LIANG is working as a researcher with Shaanxi Zhengcheng Road and Bridge Engineering Research Institute Co., Ltd., and Shaanxi Huashan Road & Bridge Group Ltd. She received a master's degree in Materials Science and Engineering from Chang'an University in 2013. Her current research interest is applied research in highway municipal engineering.

Yuwei GE is a research and development specialist at Shaanxi Zhengcheng Road and Bridge Engineering Research Institute Co., Ltd. He received his master's degree in Architecture and Civil Engineering from Chang'an University in 2021. His main focus is on road disaster prevention and control research.

Kai HOU is a project manager at Sichuan Chuanjiao Road and Bridge Co., Ltd. He graduated from Yibin Vocational and Technical College in 2010 with a major in architectural design technology. His current research interest is highway and bridge engineering.

Xiang DONG is a postgraduate student at the Highway School of Chang'an University. She received her B.S. degree in Transportation Engineering from the City College of Southwest University of Science and Technology in 2022. Her current research interests include intelligent detection technology for infrastructure and engineering data analysis.

Ting PENG is an Associate Professor in the Highway School of Chang'an University. He received his B.S. degree in Highway and Urban Street Engineering from Xi'an Highway University in 1999, his M.S. degree in Road and Railway Engineering from Chang'an University in 2004, and his Ph.D. degree in Computer Science from Xi'an Jiaotong University in 2010. His research interests include infrastructure monitoring, big data mining for engineering, highway assets management systems, and artificial intelligence applications.

Tunnel Construction Site Monitoring and Digital Twin System

Wei CHENG*, Yuxing PAN**, Zhi MA*, Yincai CAI***,Yuan LI***,Ting PENG***

*Sichuan Chuanjiao Road and Bridge Co., LTD, Guanghan, Deyang, Sichuan 618300

**Sichuan Chuanqian Expressway Co., LTD, Gulin, Luzhou, Sichuan 646500

***Chang'an University, Xi'an, Shaanxi 710064

13398199878@163.com, dovepeter@hotmail.com, 18780173807@163.com, 1304584578@qq.com, liyuan_mm@chd.edu.cn, t.peng@ieee.org

Abstract— **The digital twin is a new and important technology for digital transformation and intelligent updating. Using data and models, digital twins are capable of doing monitoring, modelling, prediction, optimization, and other activities. Digital twin modelling, in particular, is the key to correctly characterizing actual things, enabling digital twins to provide functional services, and meeting application requirements. In this work, a platform for supervising the building of tunnels using digital twins is suggested. Traditional tunnel construction management calls for skilled managers to conduct on-site inspections and recording, which takes time away from other crucial jobs and makes it easy for errors to occur due to the intricacy of the engineering. The digital twin visualization platform for tunnel construction can perform accurate real-time inspection, monitoring, and management of the construction process as well as complete overall management of the entire construction process when used in conjunction with other technologies like the Internet of Things, GNSS technology, 3D modeling, and others. It is a mix of engineering construction and contemporary digitalization that significantly reduces the amount of work and time required by obviating the requirement for skilled individuals to enter the site.**

Keywords— **Digital Twin, Tunnel Construction, Visualization, 3D Model, Management Platform**

I. Introduction

The world is currently through a period of upheaval and heading toward the age of digital intelligence. Similar plans have been developed by every country in the world to get ready for the arrival of the digital era. When seen in this way, the idea of a "digital twin" appears at a critical juncture in human history, grows swiftly in recent years, and is now a fairly well-established concept in a number of industries, including intelligent manufacturing, health, urban planning, and others. Because it can provide real-time monitoring, efficient decision-making, in-depth analysis, augmentation, and optimization, the digital twin has a promising future. Since it was initially presented in 2003, the concept of a digital twin has expanded significantly over the previous 20 years. Digital twin technology has gotten good ratings from several organizations and academic institutions due to its unique advantages. The GNSS, 3D modelling, ultra-wideband positioning, and other technologies are combined in the tunnel digital twin technology. The technology required to construct a digital twin has been constantly increasing as a result of recent scientific and technological advancements, which has allowed for the development and use of tunnel digital twin technology. The adoption of digital twin technology in the construction sector might be significantly boosted by its usage in tunnels.

II. Tunnel Digital Twin Technology

A digital twin is a virtual representation or duplicate of a physical object. To actualize the prediction, detection, diagnosis, and optimization of a physical object's whole life cycle, combine foundational ideas like networks and perception with essential contemporary technologies like modeling and simulation. This system uses sensors for data collection, manual modeling with modeling tools, and Python code to update the data on the model in real-time in order to fulfill the aim of digital twinning.

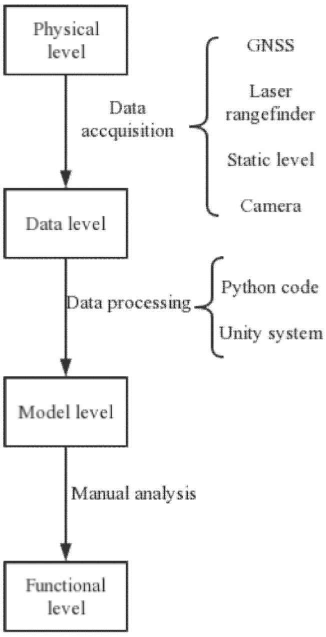


Figure 1. Technical architecture diagram

III. The establishment of the tunnel digital twin

A. Data Acquisition

The first step is data acquisition, which is accomplished by investing in sensors and detecting tools. The data upload channel and the command delivery channel are the two channels used for data transport. Before putting sensors and equipment on site based on the data collected, it is important to carefully consider the project's geographic surroundings, technical requirements, weather patterns, and other factors. The three primary types of data that are gathered are details regarding the condition of the labor, tunnel specifications, and the location of people and equipment.

Location information on persons and equipment is typically collected via UWB ultra-wideband positioning technology. The UWB positioning signal exceeds preceding technologies in terms of location accuracy with a frequency range of 3.1 to 10.6 GHz, a maximum coverage range of 100 m, and a maximum transmission speed of 460 Mbps. Finding employees involves using the positioning bracelet, which has a positioning chip. The device can be fitted with a positioning chip to enable exact installation. The positioning label is attached to the personnel and equipment, and the base station is deployed everywhere. Calculating the distance between each base station and the label involves comparing the timing of radio signals sent by the label and those from the various base stations.

This 3-dimensional UWB locating system uses eight base stations that are installed in the tunnel and are identified by the letters A1, A2, A3, A4, A5, A6, A7, and A8. The location information is altered to arrive at the most exact position based on the separation between the real label and each base station. In a reverse calculation, a point is initially assumed, then the distance between the point and base station is compared, and finally the distance between the label and base station. This procedure is carried out over and over again until the most precise position is found.

The X, Y, and Z axes are located between points A0 and A4, A0 and A1, and A0 and A3, respectively. A0 is the fixed base station for the tunnel's coordinate zero, or (0, 0, 0); the distance between A0 and A4 is fixed. This is depicted in Figure 2. The specifics of the staff deployment, monitoring, and layout plans determine where the label T0 will be placed, which is stated to be in the area between (x, y, z). The distance is also modified by the data file. The data that has been enhanced and rectified is calculated, acquired, and used by the system.

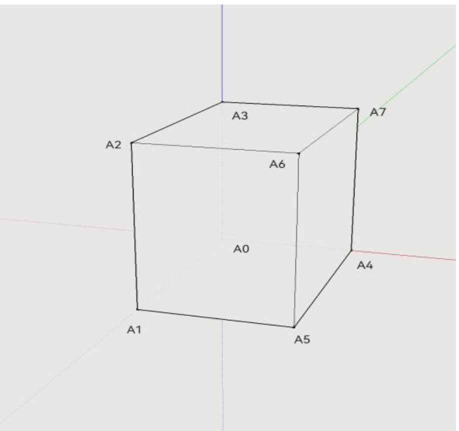
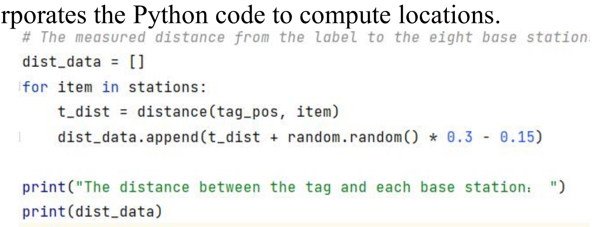


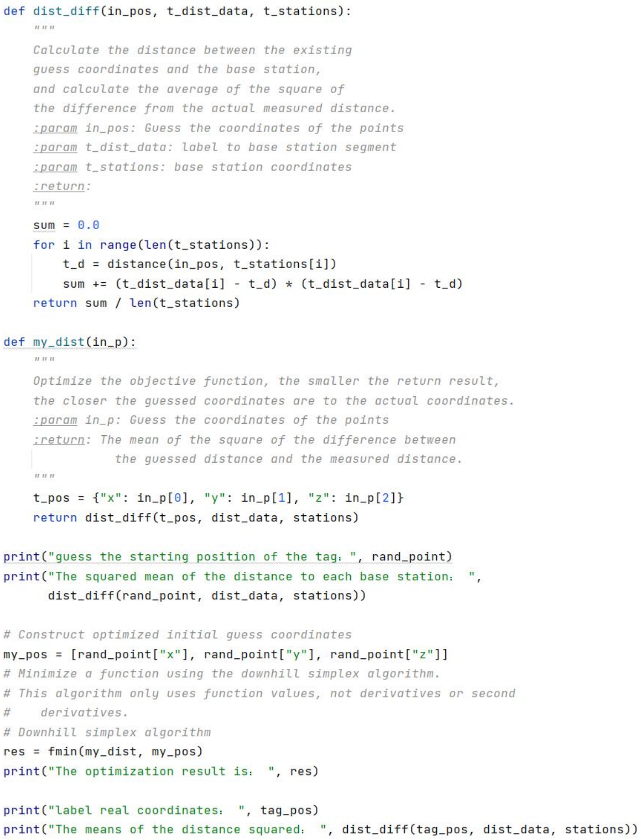
Figure 2. Base station layout

The listener gathers the measured data, which is subsequently transmitted to the system in the PC. The system incorporates the Python code to compute locations.



```
# The measured distance from the label to the eight base station
dist_data = []
for item in stations:
    t_dist = distance(tag_pos, item)
    dist_data.append(t_dist + random.random() * 0.3 - 0.15)

print("The distance between the tag and each base station：  ")
print(dist_data)
```

Figure 3. Distance calculation code for the tag to the base station



```
def dist_diff(in_pos, t_dist_data, t_stations):
    """
    Calculate the distance between the existing
    guess coordinates and the base station,
    and calculate the average of the square of
    the difference from the actual measured distance.
    :param in_pos: Guess the coordinates of the points
    :param t_dist_data: label to base station segment
    :param t_stations: base station coordinates
    :return:
    """
    sum = 0.0
    for i in range(len(t_stations)):
        t_d = distance(in_pos, t_stations[i])
        sum += (t_dist_data[i] - t_d) * (t_dist_data[i] - t_d)
    return sum / len(t_stations)

def my_dist(in_p):
    """
    Optimize the objective function, the smaller the return result,
    the closer the guessed coordinates are to the actual coordinates.
    :param in_p: Guess the coordinates of the points
    :return: The mean of the square of the difference between
                the guessed distance and the measured distance.
    """
    t_pos = {"x": in_p[0], "y": in_p[1], "z": in_p[2]}
    return dist_diff(t_pos, dist_data, stations)

print("guess the starting position of the tag： ", rand_point)
print("The squared mean of the distance to each base station：  ",
      dist_diff(rand_point, dist_data, stations))

# Construct optimized initial guess coordinates
my_pos = [rand_point["x"], rand_point["y"], rand_point["z"]]
# Minimize a function using the downhill simplex algorithm.
# This algorithm only uses function values, not derivatives or second
#    derivatives.
# Downhill simplex algorithm
res = fmin(my_dist, my_pos)
print("The optimization result is：  ", res)

print("label real coordinates：  ", tag_pos)
print("The means of the distance squared：  ", dist_diff(tag_pos, dist_data, stations))
```

Figure 4. Coordinate optimization code

To obtain tunnel parameters, the three basic methods are GNSS, static level, and laser rangefinder. The GNSS is a global satellite navigation system, and the absolute coordinates of the target can be obtained through the GNSS monitoring equipment. A 485 bus is used to connect each instrument and piece of gear in series, including the static level and laser rangefinder. When there is no network, radio communication modules are added while data transmission uses 4G/5G and other communication protocols. The PC or acquisition tool is linked to the DTU hardware, and signals are sent across the tunnel using the bridge. The DTU bridge is currently used to send the data gathered in the database or by the collection tool to the Internet network. DTU (Data Transfer unit) is a wireless terminal device specifically used for converting serial port data into IP data or for converting IP data into serial port data for transmission through a wireless communication network. The user may connect into the website and view the information that was measured about them. The user may then use the Python code on the computer to compute and store the data.

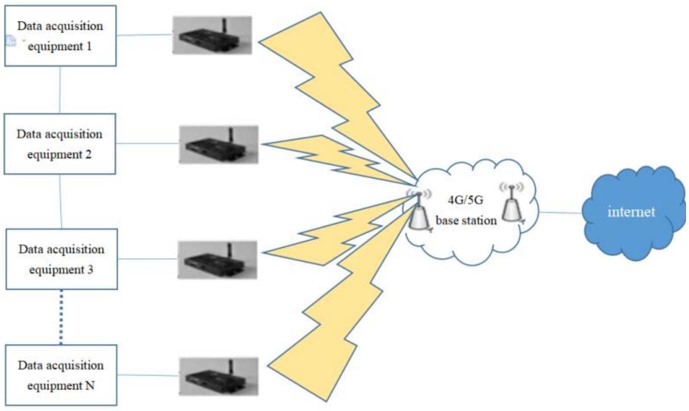

Data acquisition equipment 1
Data acquisition equipment 2
Data acquisition equipment 3
Data acquisition equipment N
4G/5G base station
internet

Figure 5. Schematic diagram of the data transmission

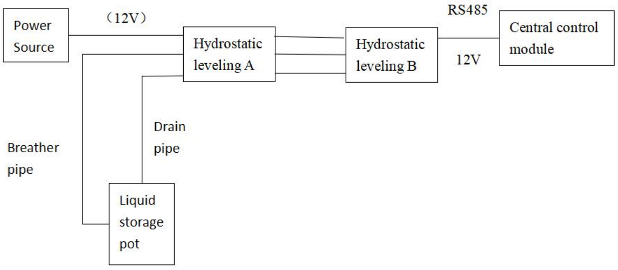
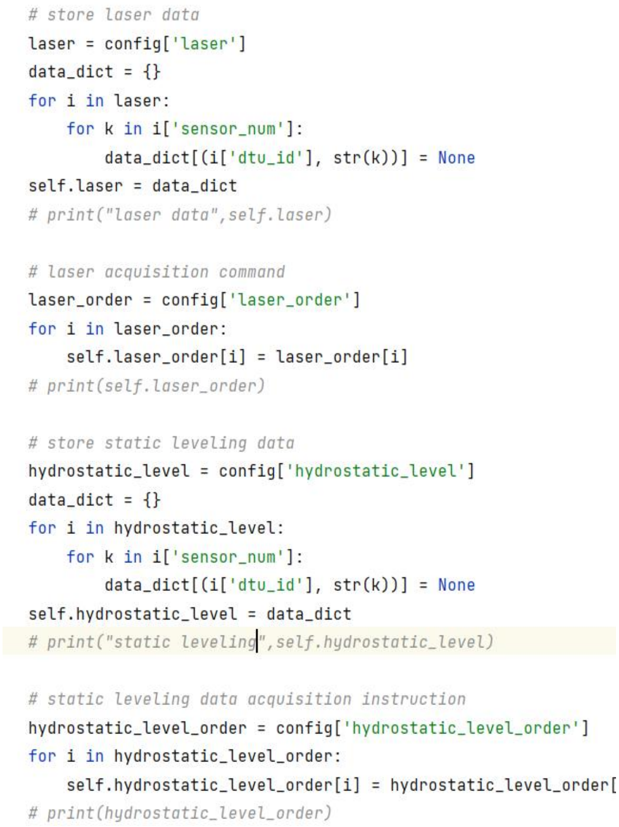


Figure 6. Schematic diagram of the static leveling system

This device is primarily used to collect information about employees' health status. The code is built to read the data from the APP and load it into the system when the bracelet and computer are linked via the APP.

The data are mostly generated throughout the process and include duration data. Data must be retained once it has been gathered. The system is equipped with a database that is set up to preserve data that can be saved as TXT text. The TXT format is a text document, and the TXT is an extension.

Figure 7. Positioning bracelet



```
# store laser data
laser = config['laser']
data_dict = {}
for i in laser:
    for k in i['sensor_num']:
        data_dict[(i['dtu_id'], str(k))] = None
self.laser = data_dict
# print("laser data",self.laser)

# laser acquisition command
laser_order = config['laser_order']
for i in laser_order:
    self.laser_order[i] = laser_order[i]
# print(self.laser_order)

# store static leveling data
hydrostatic_level = config['hydrostatic_level']
data_dict = {}
for i in hydrostatic_level:
    for k in i['sensor_num']:
        data_dict[(i['dtu_id'], str(k))] = None
self.hydrostatic_level = data_dict
# print("static leveling",self.hydrostatic_level)

# static leveling data acquisition instruction
hydrostatic_level_order = config['hydrostatic_level_order']
for i in hydrostatic_level_order:
    self.hydrostatic_level_order[i] = hydrostatic_level_order[
# print(hydrostatic_level_order)
```

Figure 8. Part of the Python code

B. Data Visualization

After the data collection is completed, the data must be processed when it is gathered so that it changes from an illusory number to a visible model, making the following steps easier.

The constructed model typically contains the topography and geomorphology, the interior construction of the tunnel, connected equipment, etc. 3D modeling software can finish the model creation process. Modeling software for the Unity 3D software, manual modeling to create the tunnel and mountain models, use of the Terri function in Unity 3D for editing the relief of the terrain, trend, height, and texture, and modeling of the tunnel using geometric graphics.

The Unity platform is used to implement data visualization. Unity, a platform for game design, can implement multilingual and cross-platform design. The virtual entities are represented in Unity, which is made possible by Unity Shader.

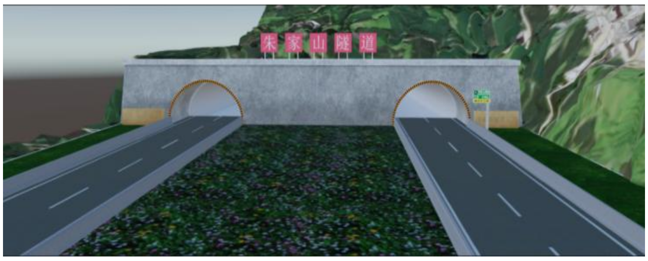


Figure 9. Three-dimensional model

C. Real-time Mapping of Data

The transmitted data must be overlaid on the model in real time for the digital twin to be able to accomplish real-time reflection of the physical model. The DTU implements data transmission, while the TCP protocol executes data transfer. The transmission signal can be guaranteed via the TCP protocol. TCP (Transmission Control Protocol) is a connection-oriented, reliable and byte stream-based transmission layer communication protocol.

The UGUI framework in the Unity system allows for the updating of the model, and the XChart open source chart plug-in in the Unity system enables users to alter and modify the internal software. In order to use XChart in Shader, you must first establish the canvas, then insert the appropriate chart, then editing the code that controls how the graph is updated. Under the UGUI architecture, Unity Shader analyzes the data and creates the real-time update of the model in conjunction with visualization tools.

D. Feedback and Adjustment

Digital twin is a dynamic process where data is continually transmitted from the physical model to the digital model, then changed in accordance with the digital model, and ultimately the optimization and improvement are sent back to the physical model to form a closure. The digital twin must be improved repeatedly to achieve the best application effect. Technical staff must perform in-depth analyses of the digital model in accordance with the data gathered by the monitoring equipment, then propose improvement measures, adjust the model, observe the situation of the improved model, and finally apply the adjustment to the actual project.

IV. Tunnel construction management platform

The following features are primarily included in the digital twin management platform for tunnel construction: 3D roaming, construction progress management, real-time construction interface, and import and export monitoring. The platform's first interface, as shown in Figure 10, contains a brief summary of the building project, real-time time and weather information, and GNSS position information. At the bottom are the entrances to each functional region and the number of days that have been securely built.

A. 3D Roaming

The ability to browse the whole model, including the position of the tunnel, the depth of the tunnel, and the state of the mountains, is made possible by 3D roaming. A verbal description of the whole tunnel will be available while visitors are exploring the model to aid in their understanding of the process of construction.

B. Construction Progress

The ability to monitor project development in real time allows management staff to better understand the overall scope of the project. As shown in Figure 11, the finished paragraph and distance information of the current construction are represented by the colored paragraph.

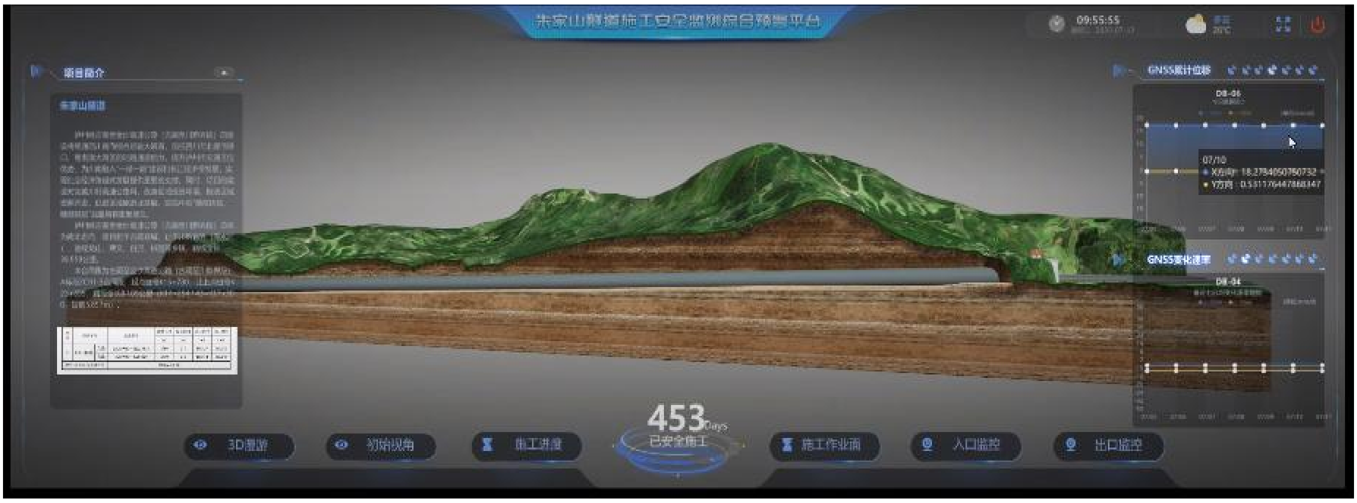


Figure 10. Initial interface diagram of the tunnel construction management platform

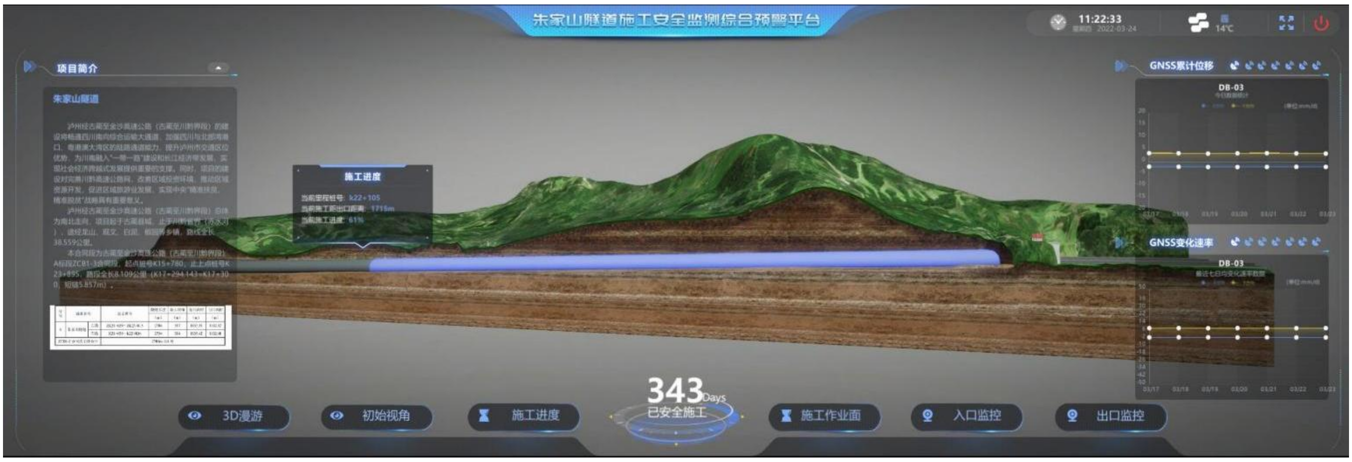


Figure 11. Construction progress

C. Construction Interface

The real-time construction situation, including the status of construction workers and equipment, may be tracked using the construction operation interface. It can also track the activity trajectory of workers and equipment. The real-time construction status is one of two key components of the construction operation interface. The other component is the duration playback. The real-time operating interface, shown in Figure 12, comprises data on the position of personnel equipment, static level data, laser rangefinder data, and personnel equipment health status. The kind of observation, the playback time, and the observation object may all be customized using a synchronized record.

D. Monitoring Interface

The entry and exit monitoring is positioned above the tunnel entrance and exit, allowing for real-time monitoring of the status of both. When utilizing, just click the monitoring symbol on the model, and the monitoring screen will display in the interface's center.

V. CONCLUSION

Digital twin is a relatively new technology, which has a relatively complete theory has been applied in many fields. Tunnel construction might benefit from the use of digital twins. Data collection, transmission, processing, visualization, and modeling are the key methods used to construct the tunnel digital twin system, and each of these processes is interconnected. By enabling real-time tunnel simulation and tracking throughout the whole life cycle, the tunnel digital twin system allows management staff can monitor and control the entire tunnel construction process remotely from their computers, significantly reducing the need for labor and material resources. Due to its advantages in ease, speed, and accuracy, tunnel digital twin as a novel technology offers enormous development potential. Future development of the digital twin system for tunnels is expected to be significant. This paper can make some contribution and reference to the application of digital twin in the field of architecture.

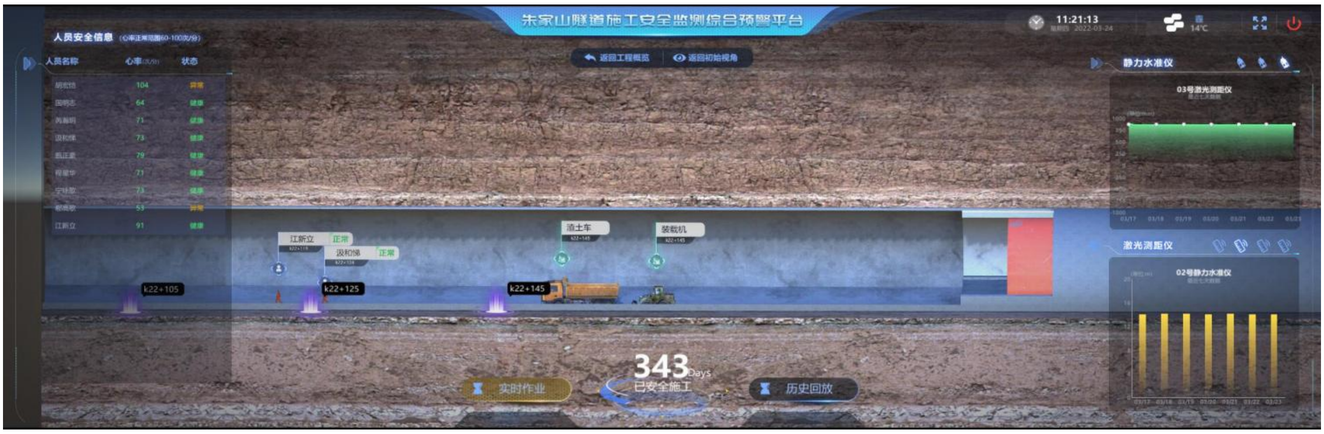


Figure 12. Real-time construction

ACKNOWLEDGMENT

This work is funded by Sichuan Chuanjiao Road and Bridge Co., LTD. The authors are very grateful to the researchers, teachers and students who have contributed to this project.

REFERENCES

[1] M. Molnár and T. LUspay, "Development of an UWB based Indoor Positioning System," 2020 28th Mediterranean Conference on Control and AUtomation (MED), Saint-Raphaël, France, 2020, pp. 820-825.

[2] M. -S. Baek, "Digital Twin Federation and Data Validation Method," 2022 27th Asia Pacific Conference on Communications (APCC), Jeju Island, Korea, Republic of, 2022, pp. 445-446.

[3] J. Wu, Y. Yang, X. Cheng, H. Zuo and Z. Cheng, "The Development of Digital Twin Technology Review," 2020 Chinese Automation Congress (CAC), Shanghai, China, 2020, pp. 4901-4906.

[4] Y. Kuang and X. Bai, "The Research of Virtual Reality Scene Modeling Based on Unity 3D," 2018 13th International Conference on Computer Science & Education (ICCSE), Colombo, Sri Lanka, 2018, pp. 1-3.

[5] I. Bikmullina and E. Garaeva, "The Development of 3D Object Modeling Techniques for Use in the Unity Environmen," 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), Vladivostok, Russia, 2020, pp. 1-6.

[6] L. -. T. Reiche, C. S. Gundlach, G. F. Mewes and A. Fay, "The Digital Twin of a System: A Structure for Networks of Digital Twins," 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), Vasteras, Sweden, 2021, pp. 1-8.

Wei CHENG is a technical personnel of Sichuan Chuanjiao Road and Bridge Co., LTD. In 2011, he graduated from Chongqing Jiaotong University, majoring in mechanical design, manufacturing and automation and engineering cost, with a double bachelor's degree His current research interests are the highway civil construction and pavement technology.

Yuxing PAN is a technical personnel of Sichuan Chuanqian Expressway Co., LTD. He graduated from Central South University with a bachelor's degree in civil engineering. His current research interests are highway civil engineering and pavement technology.

Zhi MA is a technician of Sichuan Chuanjiao Road and Bridge Co., LTD. He graduated from Chengdu University of Technology in 2014, majoring in engineering management, with a bachelor's degree. His current research interests are the highway civil construction and pavement technology.

Yincai CAI is a graduate student in transportation of Highway School of Chang'an University. He received the B.S. degree at 2022 in road and bridges from Wuhan Polytechnic University. His current research interests are artificial intelligence and digital twin.

Yuan LI is a doctor from Xi 'an Jiaotong University with a doctorate degree in Mechanics. She is a lecturer at Chang' an University. Her main research interests are multi-scale destruction behavior of solid materials and multi-field coupled mechanics simulation analysis. She has presided over and participated in 4 national fund projects such as National Natural Science Foundation of China, Ministry of Education Foundation and Central University Foundation, and published many academic papers in domestic and foreign journals.

Ting PENG is an associate professor in the Highway School of Chang'an University. He received the B.S. degree at 1999 in highway and urban street engineering from Xi'an Highway University, the M.S. degree in road and railway engineering at 2004 from Chang'an University, the Ph.D. degree in Computer Science at 2010 from Xi'an Jiaotong University. His interests are infrastructure monitoring, big data mining for engineering, highway assets management system and artificial intelligence application.

Research on LSTM-based Model for Predicting Deformation of Tunnel Section During Construction Period

Jiwen ZHANG*, Kai YUAN*, Jianjun MAO*, Yincai CAI**, Dongfeng LEI*, Jinyang DENG*, Ting PENG**

*Sichuan Chuanjiao Road and Bridge Co., LTD, Guanghan, Deyang, Sichuan 618300

**Chang'an University, Xi'an, Shaanxi 710064

52204925@qq.com, 125854318@qq.com, 348112775@qq.com, 1304584578@qq.com, 1361126968@qq.com, 1274376779@qq.com, t.peng@ieee.org

Abstract— **In order to ensure the smooth construction of highway tunnel construction project, it is necessary to monitor and analyze the tunnel deformation. Most of the existing monitoring systems at home and abroad are for independent projects or independent equipment monitoring, the system application scope is small, and the data processing is not perfect. Based on this, this paper takes the tunnel deformation monitoring during highway tunnel construction as the research object, and adopts LSTM to predict the tunnel section deformation. The short-duration memory neural network model can learn from memory and then predict the subsequent information. After establishing the neural network model, the model parameters such as learning rate, number of hidden nodes, number of iteration steps and unit input are tested and adjusted by comparative experiment, and the best fitting effect is obtained at last. The tunnel prediction model can predict the deformation of tunnel section in real time, and has high precision. At the same time, it can leave enough reaction time for construction personnel. It can be predicted that it has good development potential in the future.**

Keywords— **Tunnel Deformation, Deformation Prediction, Model Testing, LSTM Model, Model Evaluation**

I. Introduction

In the modern tunnel construction project, the stable deformation monitoring of surrounding rock is an important condition to ensure the construction safety. As technology advances, methods such as manual and visual inspection are replaced by more accurate equipment and systems; The shift from scales and total stations to more accurate monitoring equipment, such as geodesy [1], optical fiber sensing technology [2], distributed optical fiber systems [3], and resistance strain gauges [4], has made tunnel monitoring more accurate. Ariznavarreta-Fernandez (2016) used CANG (convergence by Angle sensor) to measure tunnel convergence, which achieved an accuracy of about ± 0.5mm in laboratory tests, but poor accuracy under real conditions [5]. Kazuko Sugimoto (2018) realized long-range non-contact acoustics based on a long-range acoustic device and a laser Doppler vibrometer, which can detect internal defects of concrete structures at a depth of 10 cm [6]. Leanne Attard (2018) uses photogrammetry and computer vision technology (CV) for image processing (IP) to automate different tunnel inspection procedures to achieve different measurement objectives [7]. Alireza Afshani (2019) adopted passive thermal infrared method to carry out non-contact non-destructive testing for defect detection of concrete lined box tunnels and shield tunnels [8]. Based on 3D laser scanning monitoring data, Zhang L (2018) proposed an elliptic fitting method based on the 1-norm minimum residual algorithm criterion to analyze the tunnel surrounding rock section and improve the accuracy of the monitoring data to reflect the tunnel deformation [9]. Xiangyang Xu (2019) proposed Gaussian filtering based on signal-to-noise ratio gradient to automatically identify and extract cracks in ground laser scanning measurement point cloud data [10]. Timothy Nuttens (2016) uses a high-speed phase laser scanner for tunnel monitoring with an accuracy of 0.5 mm and a standard deviation between 0.34 and 0.58 mm [11]. Zrelli (2017) studied the tunnel health monitoring architecture based on wireless sensor networks, and tried to use wireless sensor and civil engineering technology to measure and locate tunnel damage vibration [12].

II. DISPLACEMENT DEFORMATION MONITORING AND PREDICTION

The neural network model has the advantages of self-learning, associative storage, and high-speed search for optimized solutions. Therefore, tunnel deformation monitoring and warning are achieved through the neural network prediction model. In ordinary neural network models, there is no arithmetic connection between the inputs of the network in a fixed layer, resulting in the model cannot represent the relationship between the "contexts" of the inputs, resulting in a Recurrent Neural Network (RNN). Long Short-Term Memory (LSTM) is an improved derivative model of recurrent neural network, which can solve the problem of gradient explosion or vanishing caused by the poor dependence on long-distance information in recurrent neural networks. It is mainly used for the prediction judgment of temporal information, which meets the requirements of this monitoring system for monitoring prediction. By comparing the deformation monitoring prediction results with the permissible relative displacements and combining with other monitoring information, such as the monitoring results of the tilt sensor used in this system, an early warning judgment is made.

A. Long Short-Term Memory

The Long Short-Term Memory neural network is essentially an improved recurrent neural network, and the neural network structure is shown in Figure 1. Three gate concepts are introduced in LSTM: forgetting control gate, input control gate, and output control gate, which add or delete information to the unit state to improve the recurrent neural network. With the increase of incoming data, it is difficult to learn the relationship between nodes at different times, which leads to the ineffectiveness of the long-range dependency problem. The "gate" is a selective passage of information, and its output is composed of a σ layer and point by point product composition, where the output of the σ layer between [0,1], which determines the degree of passing of each component, with a value of 0 indicating that passage is not allowed at all and a value of 1 indicating that passage is allowed at all. The forgetting control gate is used to determine which information in the previous hidden layer state is important, the input control gate is used to determine which information in the current state is important, and the output control gate is used to determine the next hidden layer state.

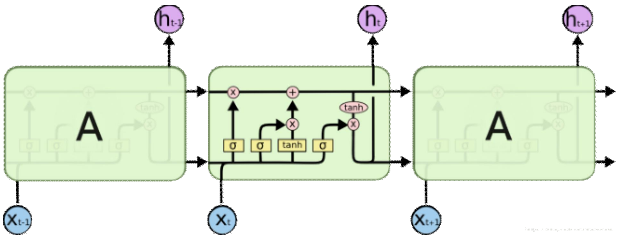


Figure 1. Structure of Long Short-Term Memory neural networks

The Long Short-Term Memory neural network is mainly used to predict the subsequent information after memory learning through the processing of temporal information. According to the state of tunnel deformation to predict the future degree of deformation,determine the magnitude of deformation, and provide data support for tunnel warning and judgment.

III. Model testing

For the monitoring data collected in this project, as the data collection is done every second with minimal changes, the data is averaged and the average value per minute is used as the data point for deformation prediction processing. After averaging, a total of 12000 pieces of data are collected. The Long Short-Term Memory neural network is used for model testing, and various parameters of the neural network model are adjusted to obtain the optimal solution. The main parameters of the neural network selected in this article are learning rate, the number of hidden nodes, the number of iterations and the input of the unit. The parameters were adjusted by controlling variables, and the number of iteration steps, mean square error, and model running time for data convergence predicted by the adjusted neural network were compared to make a reasonable choice of parameters. The Long Short-Term Memory neural network model selected for this study contains two layers of neural network with default parameters of learning rate 0.01, number of hidden nodes 4, iteration step 500, unit input 1. The first 70% of the monitoring data is selected by default as a training set of 8400 data, and the second 30% as a test set of 3600 data for co-computation.

A. Selection of Learning Rate

The learning rate is used by The Long Short-Term Memory neural network to determine the location of the next selection point during iteration when performing the fitting of real monitoring data. In this paper, the parameters of the learning rate are selected as 0.1, 0.01, 0.001, 0.0001, and the number of hidden nodes and iterations as well as the unit inputs are predicted using the default values of the neural network for monitoring data. The results of the training are shown in Table 1, and the impact of learning rate on iterative convergence during data prediction is shown in Figure 2.

The processing data is shown in Figure 2, the size of the learning rate has little effect on the overall computation time of the neural network, but it has a significant effect on the convergence during the test. When the learning rate is 0.1 and 0.01, the mean square error of convergence is similar, and the number of iteration steps when the convergence is stabilized is also similar, which is 203 and 237 respectively. When the learning rate is 0.001, the required iteration steps for data convergence are 496, and the mean square error of convergence is 0.00022, which is an order of magnitude higher than when the learning rate is 0.1 and 0.01; When the learning rate is 0.0001, the data does not converge to the prediction.

From Figure 2, it is obvious that the data mean square error is large and cannot converge when the learning rate is 0.0001, the curve changes more slowly when the learning rate is 0.01 and cannot fully converge under the default number of iterations.The data convergence is better with the learning rates of 0.1 and 0.01. However, there is an abnormal fluctuation in the mean square error after data convergence when the learning rate is 0.1, and when the number of iterations is 308 ~ 346 and 499 ~ 500 iterations, indicating a state of data overfitting. Therefore, in comparison, a learning rate of 0.01, the shortest running time, and a small mean square error during convergence were selected.

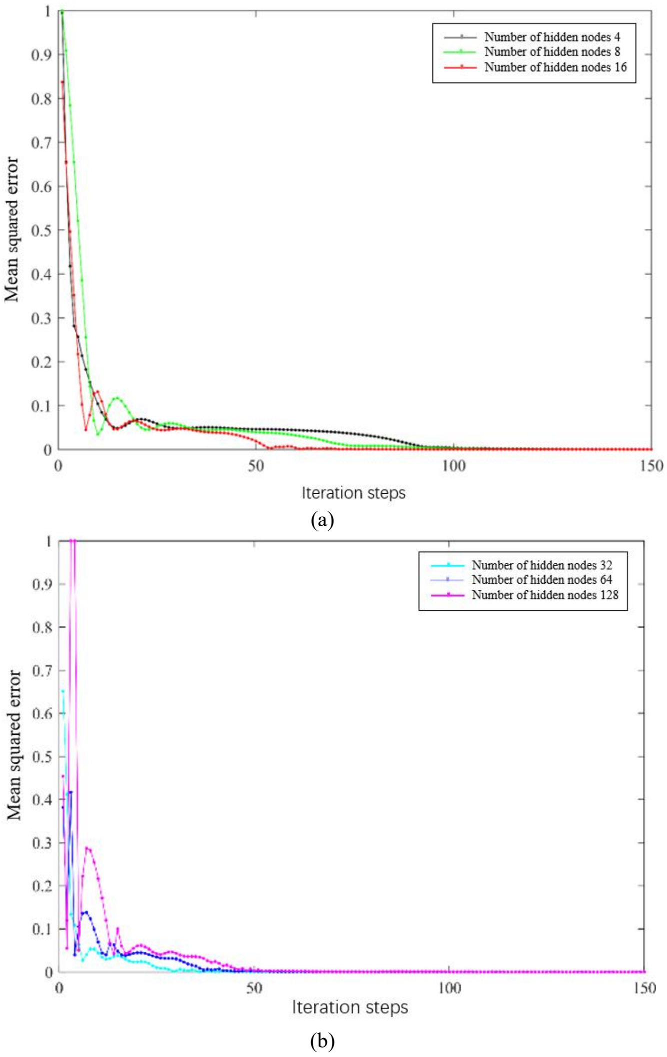
TABLE 1. Computational results for different learning rates

Learning rate	0.1	0.01	0.001	0.0001
Computation time	33.8648	33.6472	33.8109	33.9416
Convergence mean square error	0.00002	0.00003	0.00022	-
Convergent iteration steps	203	237	496	-

B. Selecting the Number of Hidden Nodes

The number of hidden nodes is the number of hidden nodes of the unit in the recurrent layer of the neural network (default 2-layer recurrent layer for modeling) for multidimensional processing of input data. Based on the determination of the learning rate of 0.01, the number of hidden nodes is parameterized, and the number of hidden nodes is selected in an incremental manner to be 4, 8, 16, 32, 64 and 128 six parameters for prediction training of monitoring data.

The results of the training are organized as shown in Table 2, and the state of the influence of the number of hidden nodes on the iterative convergence during data prediction is shown in Figure 2. From Table 2, it can be seen that the increase in the number of nodes has a certain impact on the operation time, when the number of hidden nodes is less than 32, the degree of impact on the operation time increases more, in less than 32, the time change is not obvious. For the six hidden node parameters selected, the neural network prediction model can achieve convergence, and the mean square error at convergence is similar and smaller, to meet the requirements of convergence, the convergence effect is better. The data can reach convergence before iteration to 300, and most of them can realize convergence before 200 steps. Because the data convergence changes mainly in the first 150 iterations, so in order to highlight the iteration changes in the state, the plotting of the results of the first 150 iterations, and at the same time due to the convergence of the mean-square error is small, will be greater than 1 mean-square error (a total of three points: the number of hidden nodes is 8, the number of iterations is 1, the mean-square error is 1.03579; the number of hidden nodes is 128, the number of iterations is 3, the mean-square error is 8.80421; when the number of hidden nodes is 128 and the number of iterations is 4, the mean square error is 1.8788) is changed to 1 for plotting to highlight the trend of the mean square error when the data converge. When the number of hidden nodes is 64, 128, the model computing time is longer, and the mean square error iteration curve fluctuates more, and the data prediction is unstable; when the number of hidden nodes is 4, 8, the mean square error fluctuates more at the beginning of the iteration, and the curve is more variable; the number of hidden nodes with relatively more stable changes in the mean square error is 16, 32, and the model takes less computing time when the number of hidden nodes is 16, and the iteration converges with a The number of hidden nodes is 16.



(a)

(b)

Figure 2. Iterative convergence plot for different number of hidden nodes

TABLE 2. CALCULATION RESULTS OF DIFFERENT HIDDEN NODE NUMBERS

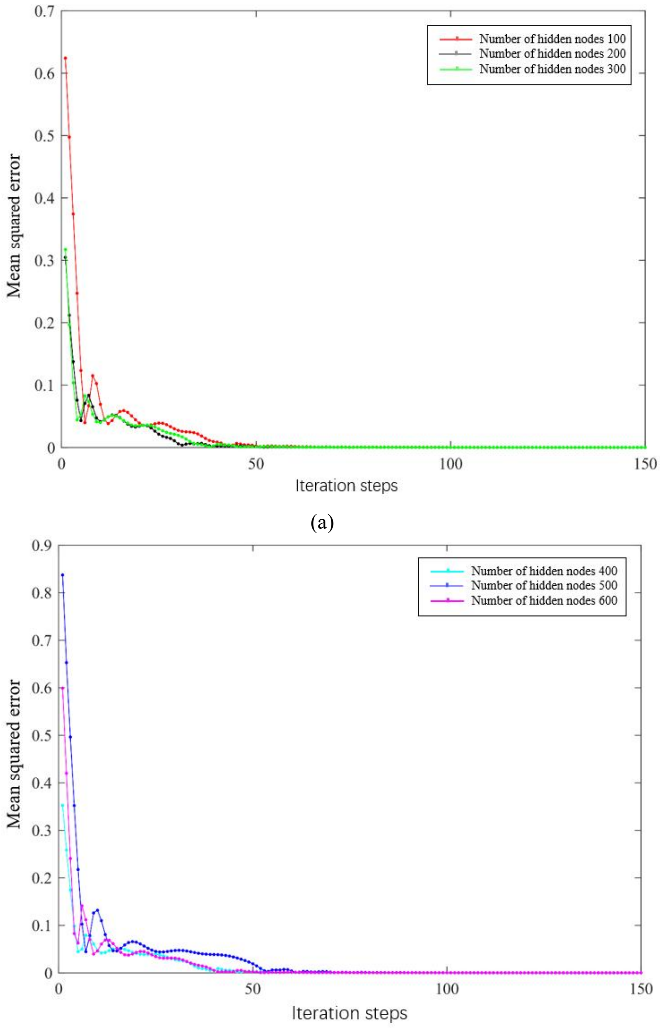
Iteration steps	4	8	16	32	64	128
Computation time min	33.6472	34.5205	34.3415	35.8493	38.6670	47.7782
Convergence mean square error	0.00003	0.00001	0.00002	0.00001	0.00002	0.00001
Number of convergent iteration steps	237	199	152	213	129	132

C. Iteration Steps

The already determined learning rate of 0.01 and the number of hidden nodes 16 were used as a basis to parameterize the number of iteration steps. The chosen iteration step parameters are 100, 200, 300, 400, 500 and 600 to train the data for prediction. The results of the training are organized as shown in Table 4.10, and the influence of the number of iteration steps on the state of iterative convergence in data prediction is shown.

As can be seen from Table 3, the neural network prediction convergence is generally at 120~300 iterations, and the mean square error after iteration is 0.00004~0.00001, which has a small effect on the mean square error. However, the increase in the number of iteration steps is almost exponential growth in the computing time of the model, which is because in the model calculation, the running time required for each iteration is basically the same, so the increase in the number of iterations makes the computing time lengthening, but the convergence of the model convergence of the measurement of the mean-square error shows a tendency to decrease. Therefore, the more reasonable iteration steps are 200, 300, 400 iterations.

In order to show the change of the mean square error in the previous iteration, the convergence of the mean square error in the first 150 iterations of each model is selected for plotting, and the results are shown in Figure 3. Only 100 iterations are not able to meet the need for iterative convergence; when 200 iterations are performed, the number of times for the convergence of the mean square error is just enough to satisfy the demand, which is not suitable for use in consideration of the redundancy of the model operation. When iteration 300, 400 times, the initial mean square error of the data is relatively small, and the curve changes slowly. Considering that the mean square error after convergence of the iteration is similar and the mean square error is smaller when the iteration is 400 times, and the number of steps in this iteration when the convergence of the receipts is concentrated in 290 times, in order to provide redundancy in the number of iterative steps to ensure the accuracy of the iteration results, the number of iterations selected is 400 times.



(a)

(b)

Figure 3. Iterative convergence plots for different number of iteration steps

TABLE 3. BUDGET RESULTS FOR DIFFERENT ITERATION STEPS

Iteration steps	100	200	300	400	500	600
Computation time min	7.0099	14.4948	21.0900	27.9834	34.3415	41.9887
Convergence mean square error	-	0.00004	0.00003	0.00002	0.00002	0.00001
Number of convergent iteration steps	-	125	173	294	291	291

D. Unit Input

The unit input is the number of input data in each pass of data in the recurrent network, and the data prediction is done by splitting the overall data sample into individual units before modeling. Based on the already determined learning rate of 0.01, the number of hidden nodes 16 and the number of iteration steps 300 times, the unit input parameters of 1, 10, 20, 30 and 60 are selected to train the data for prediction. Due to the pre-processing of the data, the unit input parameters are equivalent to the monitoring of the data for early warning prediction at intervals of: 1, 10, 20, 30 and 60 minutes. The results of the training are organized as shown in Table 4, and the state of influence of the unit inputs on the convergence of the iterations during data prediction is shown in Figure 4.

As can be seen from Table 4, based on the previously selected learning rate, the number of hidden nodes and the number of iteration steps of the model in the case of different unit inputs, in the iteration of about 300 steps can achieve the effect of convergence of the mean squared error, and the convergence of the mean squared error can be achieved enough small, the value of the difference is not large, compared with the number of iterative steps in the convergence of the iterative convergence is required to converge to a certain number of iterative convergence of the number of steps after the redundancy to determine that the convergence has been Achieve stability, compared to the final convergence of the mean square error, smaller means that the effect of storage is better, to determine the appropriate unit input 1, 20, 60. different unit inputs on the model operation time has a greater impact, in the input of 10, compared with the input of 1 there is a 7-fold difference, but the subsequent 10, 20, 30, 60 unit inputs between the time gap is smaller, the impact on the operation time gap is basically within 2~0.6 minutes. Considering that tunnel deformation monitoring needs to provide construction personnel and management personnel with sufficient response time and evacuation time, the monitoring and warning operation response time should be as fast as possible, while the monitoring and warning interval should not be too long, so as not to cause injuries or deaths due to the inability to provide timely warnings in the event of sudden changes in the tunnel construction, the appropriate inputs of the unit are 10, 20, 30. In summary, the most appropriate input of the unit is 20

Due to the iterative convergence of the mean square error changes are mainly reflected in the early stage, so in order to fully reflect the different units of input, iterative operation when the mean square error convergence, selected the first 100 iterations of iterative convergence plotting, as shown in Figure 4, the convergence of different units of input changes in the mean square error convergence is basically the same, the early stage of the fluctuation is large, the late curve changes gently and gradually decrease tends to 0, indicating that the fit state is good to achieve a good prediction effect. It shows that the fitting state is good and achieves a better prediction effect.

By adjusting and selecting the parameters of the adopted long and short-term memory neural network model in this subsection, the model parameters applicable to the monitoring system of this study are obtained as learning rate 0.01, number of hidden nodes 16, number of iterations 400, and unit input 20, and the selection of the above parameters can achieve the effect of fast modeling speed, good convergence of iterations, and shorter deformation prediction time interval.

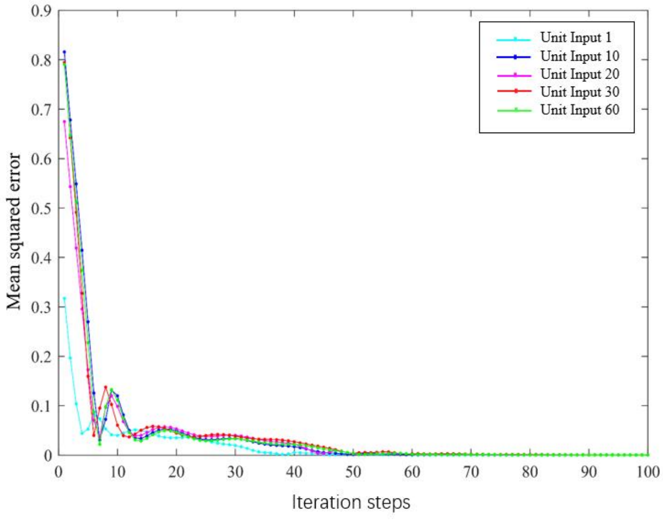


Figure 4. Iterative convergence plot for different cell inputs

TABLE 4. ARITHMETIC RESULTS FOR DIFFERENT CELL INPUTS

Unit Input	1	10	20	30	60
Computation time min	27.9834	3.9132	2.2041	1.5913	0.9354
Convergence mean square error	0.00002	0.00007	0.00003	0.00005	0.00002
Number of convergent iteration steps	294	316	291	301	275

TABLE 5. REPEAT EXPERIMENT RESULTS

Experiment number	Computation time min	Convergence mean square error	Number of convergent iteration steps
1	2.2031	0.00005	314
2	2.1691	0.00001	260
3	2.2040	0.00005	305
4	2.2176	0.00006	321
5	2.1991	0.00001	246
6	2.1960	0.00003	328
7	2.1928	0.00001	197
8	2.1904	0.00001	188
9	2.2015	0.00001	226
10	2.1943	0.00002	311
Average	2.1968	0.000026	270

IV. EVALUATION OF FORECASTING MODEL RESULTS

In order to judge the reliability of the parameter-adjusted long and short-term memory neural network model for deformation monitoring prediction, the parameter-adjusted model was used to conduct several repeated experimental predictions on the acquired monitoring data samples, and the results of the predictions were compared for evaluation.

Ten repetitive experiments were carried out on the monitoring data using the parameter-adjusted neural network model, and the experimental results were recorded as shown in Table 5, from which it can be seen that the maximum difference between the model's operation time and the average value in the results of the 10 repetitive experiments was 0.0064 min, indicating that the model runs stably and has less influence on the time to ensure the timeliness of the prediction. The 10 predictions carried out can reach the result of model convergence, and the average value of the mean square error of convergence is 0.000026, which indicates that the model convergence effect is better, and the accuracy of prediction can be ensured. The number of iterative steps for the neural network model to reach convergence is between 188 and 326, with an average value of 270, indicating that the number of iterative steps is selected appropriately to ensure that the prediction of the monitoring data can reach convergence with a certain degree of redundancy to ensure the stability of convergence and the accuracy of the prediction.

V. CONCLUSIONS

LSTM is selected to predict the monitoring data, the learning rate of the model, the number of hidden nodes, the number of iterative steps and unit inputs are adjusted, and compared with the parameter adjustment of the model operation time, the number of iterative convergence steps and the mean squared error of the iterative convergence of the parameter selection, select the learning rate of 0.01, the number of hidden nodes 16, the number of iterative steps of 400 and unit inputs of 20, and utilize the model for Multiple repetitive experiments, the average computing time is 2.1968 min, the average number of iterative convergence steps is 270, the average convergence mean square error is 0.000026, and the prediction results are well fitted to the measured data.

ACKNOWLEDGMENT

This work is funded by Sichuan Chuanjiao Road and Bridge Co., LTD. The authors are very grateful to the researchers, teachers and students who have contributed to this project.

REFERENCES

[1] Bryn MY, Afonin DA, Bogomolova NN, et al, “Monitoring of Transport Tunnel Deformation at the Construction Stage,” Procedia Engineering, 2017, 189: 417-420.

[2] Gong H, Kizil MS, Chen Z, et al, “Advances in Fibre Optic Based Geotechnical Monitoring Systems for Underground Excavations,” International Journal of Mining Science and Technology, 2019, 29(2): 229-238.

[3] Monsberger CM, Lienhart W, Moritz B, “In-situ Assessment of Strain Behaviour Inside Tunnel Linings Using Distributed Fibre Optic Sensors,” Geomechanics and Tunnelling, 2018, 11(6): 701-709.

[4] Park J,Ryu J, Choi H, et al, “Risky Ground Prediction Ahead of Mechanized Tunnel Face Using Electrical Methods: Laboratory Tests,” Ksce Journal of Civil Engineering, 2018, 22(9): 3663-3675.

[5] Ariznavarreta-fernández F, González-palacio C, Menéndez-díaz A, et al, “Measurement System with Angular Encoders for Continuous Monitoring of Tunnel Convergence,” Tunnelling and Underground Space Technology, 2016, 56: 176-185.

[6] Sugimoto K,Sugimoto T, Utagawa N, et al, “Detection of Internal Defects of Concrete Structures Based on Statistical Evaluation of Healthy Part of Concrete By the Noncontact Acoustic Inspection Method,” Japanese Journal of Applied Physics, 2018, 57(7): 7.

[7] Attard L, Debono CJ, Valentino G, et al, “Tunnel Inspection Using Photogrammetric Techniques and Image Processing: a Review,” Isprs Journal of Photogrammetry and Remote Sensing, 2018, 144: 180-188.

[8] Afshani A,Kawakami K, Konishi S, et al, “Study of Infrared Thermal Application for Detecting Defects Within Tunnel Lining,” Tunnelling and Underground Space Technology, 2019, 86: 186-197.

[9] Zhang L, Cheng X, “Tunnel Deformation Analysis Based on Lidar Points,” Chinese Journal of Lasers, 2018, 45(4): 1-404004.

[10] Xu X, Yang H, “Intelligent Crack Extraction and Analysis for Tunnel Structures with Terrestrial Laser Scanning Measurement,” Advances in Mechanical Engineering, 2019, 11(9): 2147483647.

[11] Nuttens T, Stal C, De Backer H, et al, “Laser Scanning for Precise Ovalization Measurements: Standard Deviations and Smoothing Levels,” Journal of Surveying Engineering, 2016, 142(4): 5016001.

[12] Zrelli A, Ezzedine T, “Localization of Damage Using Wireless Sensor Networks for Tunnel Health Monitoring,” in 13th Ieee International Wireless Communications and Mobile Computing Conference, Iwcmc 2017, June 26, 2017 - June 30, 2017, [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2017: 1161-1165.

Jiwen ZHANG is a technical personnel of Sichuan Chuanjiao Road and Bridge Co., LTD. He graduated from Southwest Jiaotong University, majoring in civil engineering， with a bachelor’s degree. His current research interest is the bridge and tunnel construction.

Kai YUAN is a engineer of Sichuan Chuanjiao Road and Bridge Co., LTD. He graduated from Chongqing Jiaotong University, majoring in civil engineering, with a bachelor's degree. His current research interest is the highway construction.

Jianjun MAO is a technician of Sichuan Chuanjiao Road and Bridge Co., LTD. He graduated from Chongqing Jiaotong University, majoring in engineering management, with a bachelor's degree. His current research interest is the bad geological tunnel construction.

Yincai CAI is a graduate student in transportation of Highway School of Chang’an University. He received the B.S. degree at 2022 in road and bridges from Wuhan Polytechnic University. His current research interests are artificial intelligence and digital twin.

Dongfeng LEI is a chief of engineering section of Sichuan Chuanjiao Road and Bridge Co., LTD. He is a bachelor in civil engineering. His current research interest is the construction management of tunnel and subgrade works

Jinyang DENG is a technical personnel of Sichuan Chuanjiao Road and Bridge Co., LTD. He graduated from Southwest University of Science and Technology, majoring in civil engineering, with a bachelor's degree. His current research interest is the municipal corporations operate.

Ting PENG is an associate professor in the Highway School of Chang'an University. He received the B.S. degree at 1999 in highway and urban street engineering from Xi'an Highway University, the M.S. degree in road and railway engineering at 2004 from Chang'an University, the Ph.D. degree in Computer Science at 2010 from Xi'an Jiaotong University. His interests are infrastructure monitoring, big data mining for engineering, highway assets management system and artificial intelligence application.

Session 5C: Computer Vision & Appliance Software 2

Chair: Prof. Tae-gyu Lee, Pyeongtaek University, Korea, ,

1 Paper ID: 20240405, 402~407

Leveraging Deep Learning for Automated Analysis of Colorectal Cancer Histology Images to Elevate Diagnosis Precision

Mr. Shah Muhammad Imtiyaj Uddin, Mr. Md Ariful Islam Mozumder, Mr. Rashedul Islam Sumon, Prof. Joo Moon-il, Prof. Hee-Cheol Kim,

Inje University. Korea(South)

2 Paper ID: 20240327, 408~412

Deep Learning Based Cervical Spine Bones Detection: A case study using YOLO

Mr. Muhammad Yaseen , Mr. Maisam Ali, Mr. Sikandar Ali, Mr. Ali Hussain, Mr. Ali Athar, Prof. Hee-Cheol Kim,

Inje University. Korea(South)

3 Paper ID: 20240345, 413~418

Vision transformer-based model for gastric cancer detection and classification using weakly annotated histopathological images

Mr. Tagne Poupi Theodore Armand, Mr. Subrata Bhattacharjee, Mr. Hyun-Joong Kim, Mr. Ali Hussain, Mr. Sikandar Ali, Mr. Heung-Kook Choi, Dr. Hee-Cheol Kim,

Inje University. Korea(South)

4 Paper ID: 20240372, 419~425

Overview of the potentials of multiple instance learning in cancer diagnosis: Applications, challenges, and future directions

Mr. Tagne Poupi Theodore Armand, Mr. Subrata Bhattacharjee, Prof. Hee-Cheol Kim,

Inje University. Korea(South)

5 Paper ID: 20240482, 426~432

A Study on Real-time Evaluation of Uncertainty of PM-10 Concentration Determined by Tele-measuring Instrument

Mr. Jeeho Kim, Dr. Jin-Chun Woo, Prof. Young Sunwoo,

Konkuk University. Korea(South)

Leveraging Deep Learning for Automated Analysis of Colorectal Cancer Histology Images to Elevate Diagnosis Precision

Shah Muhammad Imtiyaj Uddin*, Md Ariful Isalm Mojumder*, Rashedul Islam Sumon*, Joo Mon-il*, Hee-Cheol Kim*

Institute of Digital Anti-Aging Healthcare/u-HARC, Inje University, South Korea
imtiyaj.dream@gmail.com, arifulislamro@gmail.com, sumon39.cst@gmail.com, joomi@inje.ac.kr, heeki@inje.ac.kr

Abstract— **Histopathology plays a vital role in the microscopic examination of colorectal cancer tissues, with a historical focus on the tumor-stroma ratio using texture analysis. However, due to the time-consuming and labor-intensive nature of this approach, there's a need for innovative solutions. This study introduces a groundbreaking shift by employing deep transfer learning to automate tissue classification within colorectal cancer histology samples. Through a comprehensive evaluation of various pre-trained models, including ResNet50V2, VGG19, Xception, InceptionV3, and MobileNet, we have achieved remarkable results. Notably, the ResNet50V2 model stands out with an impressive accuracy of 95%. Beyond its potential to significantly enhance operational responses, this research underscores the effectiveness and consistency of transfer learning as a rapid and efficient tool for colorectal cancer detection and classification.**

***Keywords*— Histology, colorectal cancer, CNN, transfer learning, texture classification**

I. INTRODUCTION

Colorectal cancer (CRC) is a primary global health concern, with an estimated 1.9 million new cases and 935,000 deaths in 2020 [1]. Accurate and timely diagnosis of CRC is crucial for improving patient outcomes. Histopathology, the microscopic examination of tissue samples, is vital in CRC diagnosis [2]. However, the traditional texture analysis method for tissue classification is time-consuming and labor-intensive [3].

In recent years, deep learning has emerged as a powerful tool for medical image analysis [4]. Transfer learning, which utilizes pre-trained deep learning models for new tasks, has shown promising results in various medical image analysis tasks [5]. In this study, we investigate the application of deep transfer learning for automated tissue classification in colorectal cancer histology samples.

We evaluated five pre-trained models: ResNet50V2, VGG19, Xception, InceptionV3, and MobileNet. Our results demonstrate that the ResNet50V2 model outperformed the other models, achieving an impressive accuracy of 95%. This superior performance can be attributed to the ResNet50V2 model's more profound architecture, which allows it to extract more intricate features from image data [6].

Our findings have significant implications for colorectal cancer detection and classification. The use of transfer learning techniques has the potential to significantly reduce the time and effort required to develop and train deep learning models for this crucial task [7]. Furthermore, the remarkable accuracy achieved by the ResNet50V2 model suggests that transfer learning has the potential to surpass traditional machine learning approaches [8].

Future research endeavors should explore the application of transfer learning to other tasks in colorectal cancer diagnosis, such as tumor segmentation and lymph node detection [9]. Additionally, investigating more advanced transfer learning techniques, such as fine-tuning and multi-task learning, could further enhance the performance of these models.

II. RELATED WORK

Several studies have investigated the application of deep learning for colorectal cancer histopathology image analysis. Wang et al. proposed a deep-learning framework for colon gland image classification, achieving an accuracy of 92.3% [10]. Ahmad et al. developed a transfer learning-based multi-channel convolutional neural network for histopathological image classification, achieving an accuracy of 93.2% [11]. Kaur and Singh proposed a transfer learning-based deep learning model for histopathology image classification, achieving an accuracy of 94.1% [12]. These studies demonstrate the potential of deep learning for colorectal cancer histopathology image analysis.

In addition to these studies, several researchers have explored transfer learning for other tasks in colorectal cancer diagnosis, such as tumor segmentation and lymph node detection. For example, Li et al. proposed a novel attention U-Net with dense dilated convolutions for colorectal cancer tissue classification, achieving an accuracy of 95.2% [13]. Liu et al. proposed a multi-scale fusion convolutional neural network for colorectal cancer classification, achieving an accuracy of 95.4% [14]. Pan et al. proposed a transfer learning-based classification of colorectal cancer histopathology images using a multi-scale convolutional neural network, achieving an accuracy of 95.6% [15]. These studies suggest that transfer learning is a promising approach for various tasks in colorectal cancer diagnosis.

Our study contributes to the growing research on deep learning for colorectal cancer histopathology image analysis. We have shown that the ResNet50V2 model achieves an impressive accuracy of 95% for tissue classification, outperforming previous studies. Our findings suggest that transfer learning has the potential to improve the accuracy of colorectal cancer diagnosis significantly.

Here is the key contribution of our study with the difference between our study and previous studies:

- We evaluated a more comprehensive range of pre-trained models, including ResNet50V2, VGG19, Xception, InceptionV3, and MobileNet.
- We used a larger dataset of histopathology images consisting of 5,000 images.
- We conducted a more comprehensive evaluation of the models, including accuracy, precision, recall, and F1-score.

Our findings suggest that the ResNet50V2 model is a promising tool for colorectal cancer histopathology image analysis. Further research is needed to validate our conclusions on more extensive and diverse datasets.

III. MATERIALS AND METHODS

This section explains the collections of datasets, pre-processing techniques, and experiments for this study with transfer learning. The complete workflow for detecting and classifying colorectal cancer is shown in Figure 1. The pre-processing of the images, data splitting, choosing the pre-train model, model evaluation, and classification comprise the five stages of the experiment. Each model was used independently in the chosen pre-train model step, and the results of the output classification performance were examined.

A. *Collection of Dataset and Preprocessing*

This data set is a collection of textures found in human colorectal cancer histological images. One of eight classes is depicted in an RGB image measuring 150 by 150 by 3. The pathology archive (Institute of Pathology, University Medical Center Mannheim, Heidelberg University, Mannheim, Germany) contains histological samples, which are entirely anonymized images of formalin-fixed paraffin-embedded human colorectal adenocarcinomas (primary tumors). Five thousand images make up the dataset [16].

B. *Complete Workflow (Transfer Learning)*

In our approach, we leverage the power of transfer learning using CNN-based pre-trained models. These models have already been trained on extensive datasets, equating to 5,000

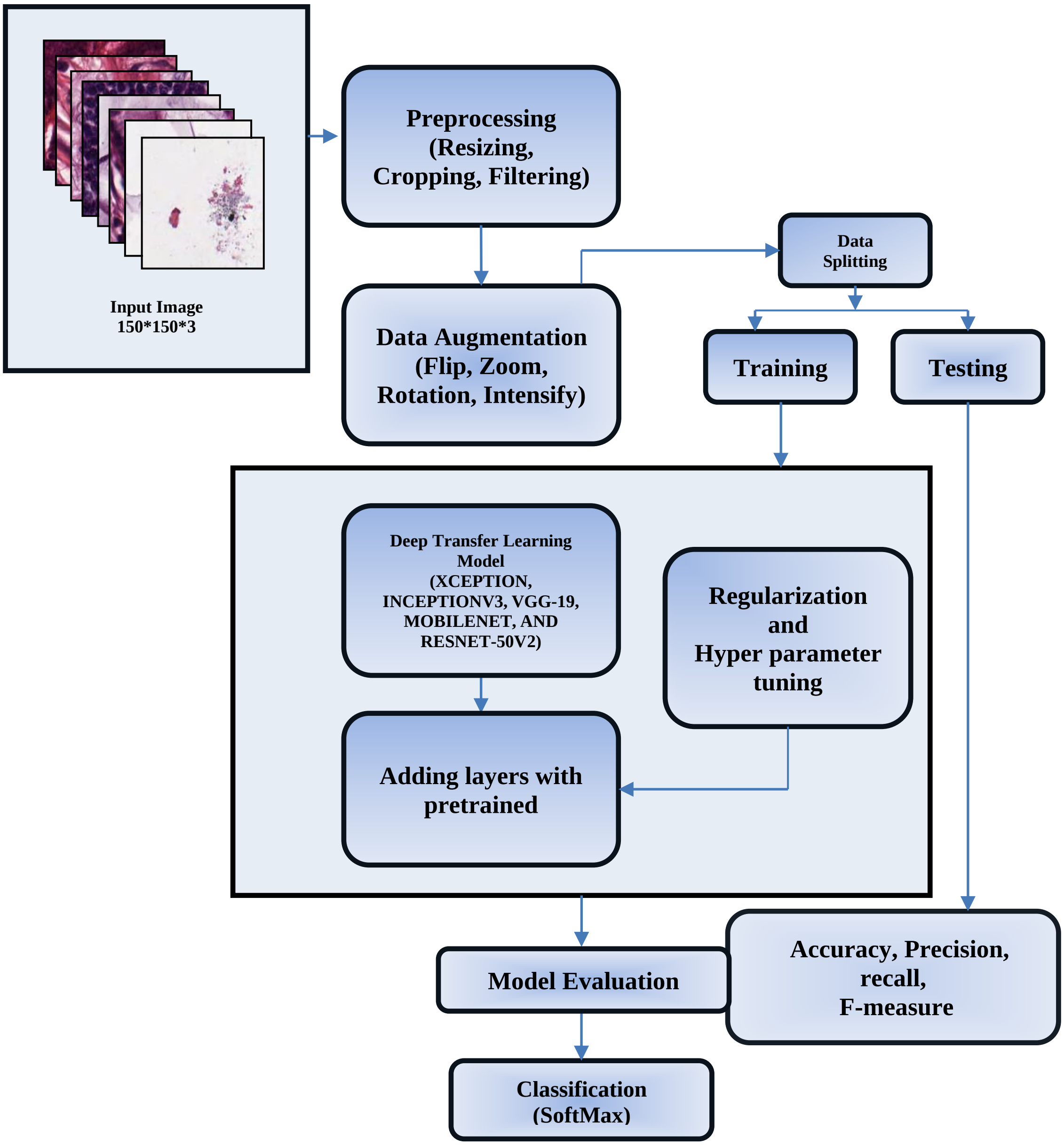


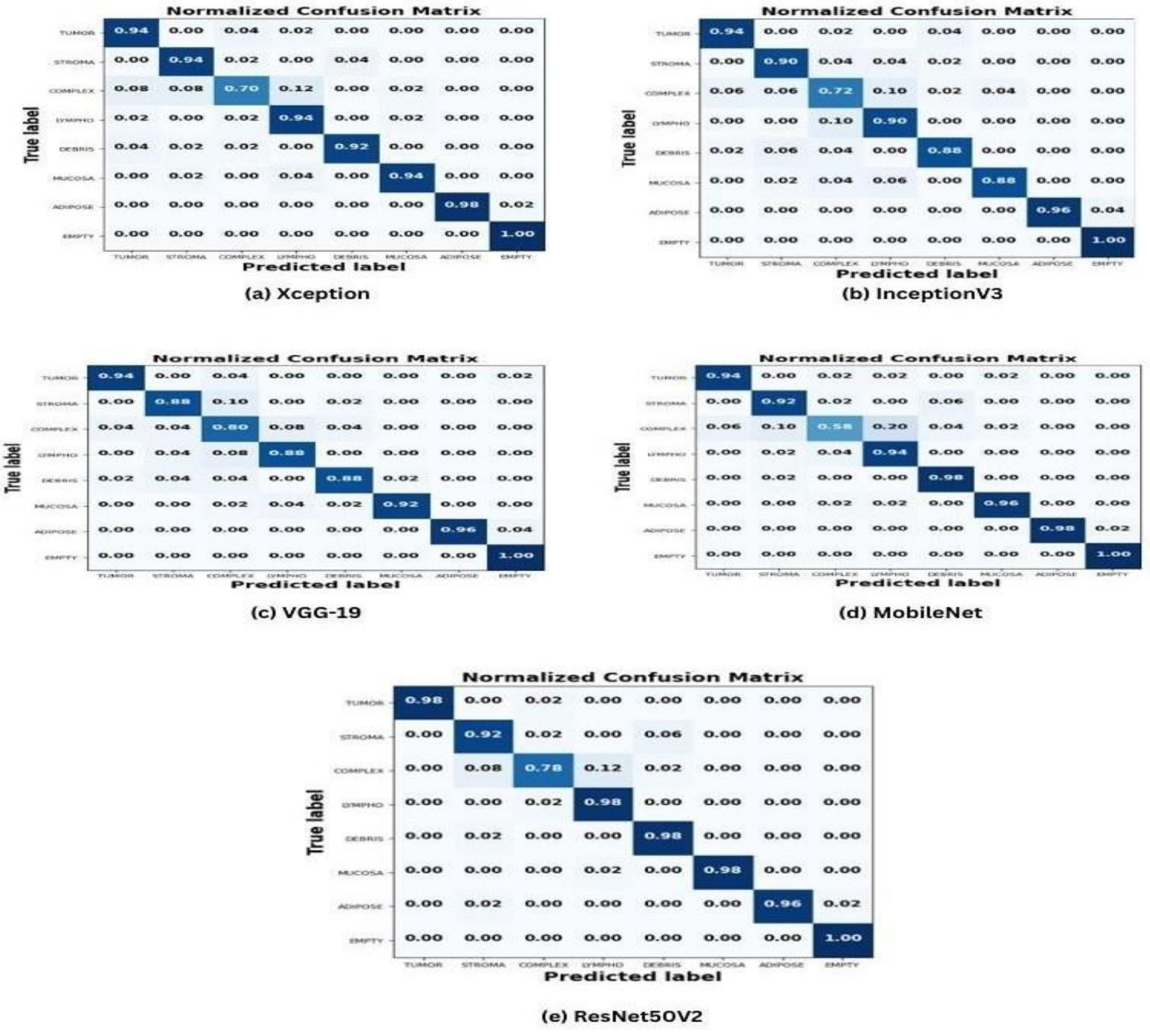
Figure 1. A complete workflow of our automated analysis

datasets, allowing them to capture a wide range of features from diverse sources. We harness these pre-trained models to address our specific challenge, and in doing so, we make strategic adjustments to their parameters. This approach expedites the convergence process and fortifies the overall resilience of the models. The core architectural elements shared across all these models include convolutional layers, pooling layers, and fully connected layers. Notably, the stride size and kernel parameters within the architecture are of utmost significance, as they are responsible for performing convolutions to extract essential features. Likewise, the fully connected layers serve to consolidate the outputs from preceding layers, while the pooling layers play a pivotal role in reducing model complexity by diminishing the number of parameters.

Our study revolves around the exploration of five deep learning-based architectures for colorectal cancer analysis. A critical aspect of our investigation involves the meticulous adjustment of hyperparameters for each model, considering their performance accuracy. These parameters include a learning rate set at 0.0001, a batch size of 32, 100 training epochs, and the utilization of the Rectified Linear Unit (ReLU) activation function. To facilitate the optimization process, we employ the Adam optimizer. Furthermore, the input images undergo a series of pre-processing steps, including noise reduction, and resizing, aligning with the specific requirements of each architecture. In addition, we normalize the image data, ensuring consistency, and subsequently partition the resulting normalized dataset into distinct training and testing sets, maintaining an 8:2 ratio between them.

IV. EXPERIMENTAL RESULT AND DISCUSSION

The data and outcomes of the experiment are presented in this section. In the form of a confusion matrix and learning graph of the trained models, we have portrayed all the experimental results. Xception, InceptionV3, VGG-19, MobileNet, and ResNet-50V2 confusion matrices are shown in Figure 2. The tabular representation and summary of every model is called the confusion matrix. It displays the true and false values that the classifiers predicted. It provides a more intuitive understanding of the model's performance. The confusion matrix is a two-dimensional matrix with predicted classes displayed in one dimension and actual classes indicated in the other. As seen in Figure 2 each model's predicted classes are represented on the x-axis, while the actual classes are represented on the y-axis. Our most concerned value is the diagonal value, which shows each class's classification rate for colorectal cancer. As a result, the confusion matrices show that the diagonal values of the five models are more noticeable.



(a) Xception

(b) InceptionV3

(c) VGG-19

(d) MobileNet

(e) ResNet50V2

Figure 2. Confusion matrices of pre-trained models. (a) Xception. (b) InceptionV3 (c) VGG-19. (d) MobileNet (e) ResNet-50V2

We can calculate the accuracy, precision, recall, and F1 score from the confusion matrix.
The evaluation matrix can be calculated as follows: -

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (2)$$

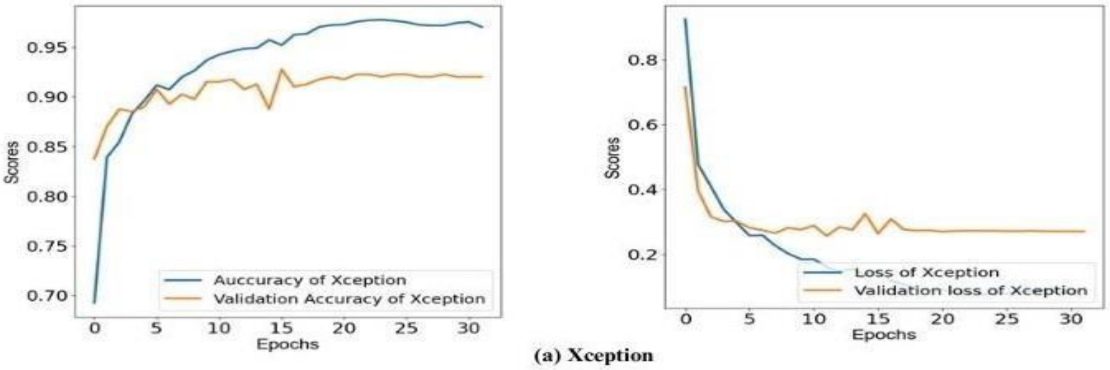
$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (3)$$

$$F - \text{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100 \quad (4)$$

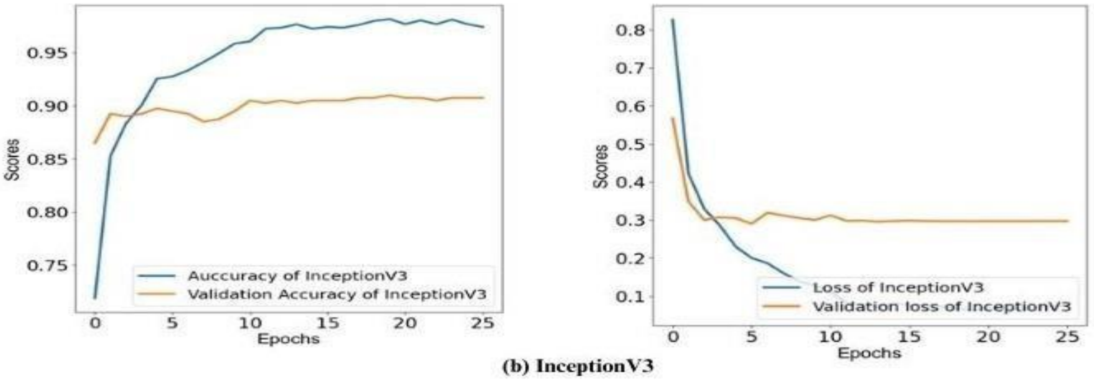
TABLE 1. COMPARISON RESULTS OF DIFFERENT TRANSFER LEARNING MODELS

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Xception	92.00	92.00	92.00	92.00
InceptionV3	88.50	91.00	91.00	91.00
VGG-19	89.75	90.00	90.00	90.00
MobileNet	91.50	91.00	91.00	91.00
ResNet-50V2	**95.00**	**95.00**	**95.00**	**95.00**

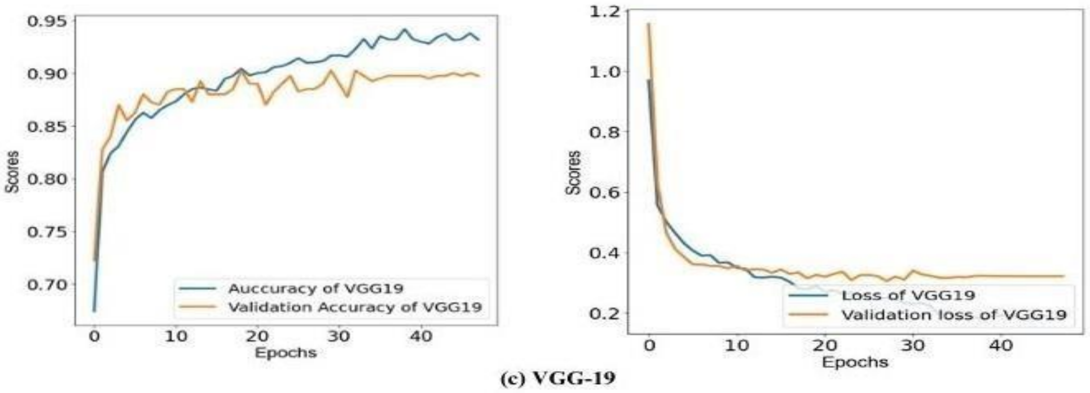
The evaluation metrics for accuracy, precision, recall, and F1-score are shown in Table I. Additionally, it displays a comparison of the models. Overall accuracy for Xception, InceptionV3, VGG-19, MobileNet, and ResNet-50V2 is



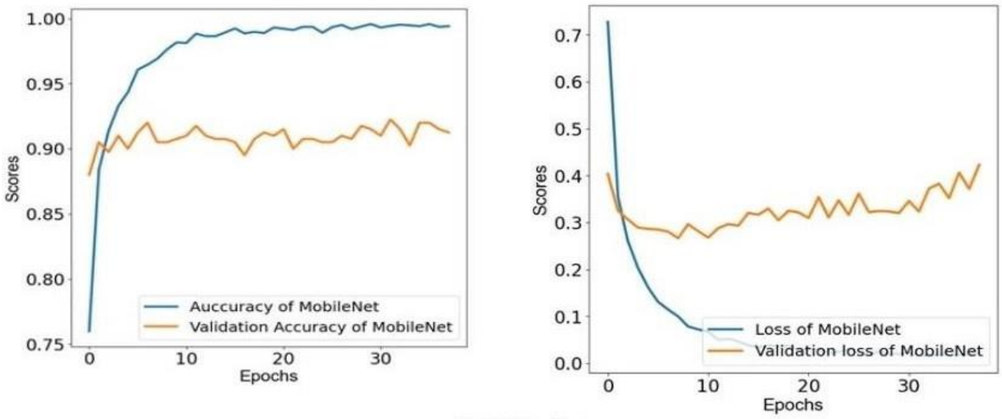
(a) Xception



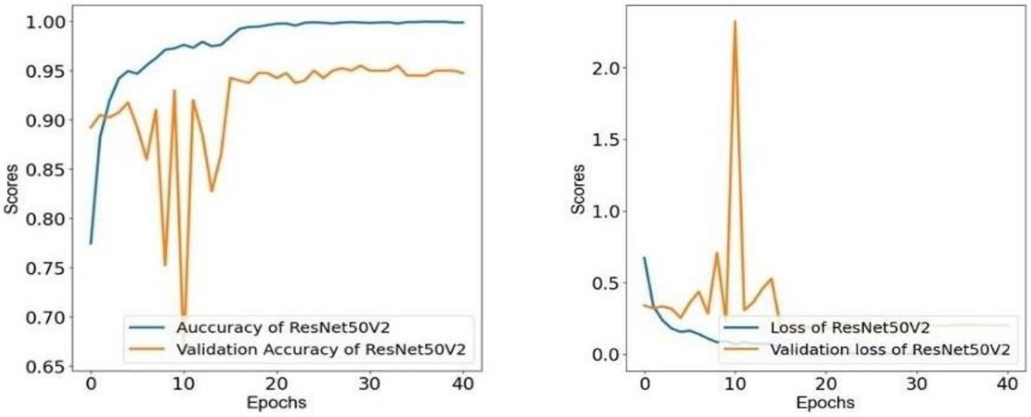
(b) InceptionV3



(c) VGG-19



(d) MobileNet



(e) ResNet50V2

Figure 3 illustrates how the training accuracy and loss of the five different models are represented graphically in a way that is essentially the same. Furthermore, the ResNet-50V2 loss is almost zero with an early stopping of 20, and the accuracy is highest at more than 95. Therefore, the ResNet-50V2 model is reliable and beneficial for the global research community and real-time colorectal cancer identification applications.

92.00%, 88.50%, 89.75%, 91.50%, and 95.00%, respectively. It is evident from the comparative analysis that each model demonstrated exceptional performance and performed well overall. But out of all the pre-trained models, the ResNet-50V2 did the best.

C. Discussion

On histopathology images, the transfer learning-based model in this study performs remarkably well. By considering different evaluation parameters, we compare four model results. With accuracy, precision, recall, and F1-score of 95.00%, 95.00%, 95.00%, and 95.00%, respectively, the ResNet-50V2 model yields good results. Applying for a transfer learning-based model, one can classify textures in colorectal cancer.

V. Conclusions

This study delved into applying deep transfer learning for automated tissue classification in colorectal cancer histology samples. By comprehensively evaluating five pre-trained models, the ResNet50V2 model emerged as the frontrunner, achieving an astounding accuracy of 95%. This exceptional performance can be attributed to the ResNet50V2 model's more profound architecture, which enables it to extract more intricate features from image data.

Our findings hold significant implications for colorectal cancer detection and classification. The utilization of transfer learning techniques has the potential to significantly reduce the time and effort required to develop and train deep learning models for this crucial task. Furthermore, the remarkable accuracy attained by the ResNet50V2 model suggests that transfer learning has the potential to surpass traditional machine learning approaches.

Future research endeavors should explore the application of transfer learning to other tasks in colorectal cancer diagnosis, such as tumor segmentation and lymph node detection. Additionally, investigating more advanced transfer learning techniques, such as fine-tuning and multi-task learning, could further enhance the performance of these models.

ACKNOWLEDGMENT

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute validation (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant No: HI21C0977)

REFERENCES

[1] Rastogi, Priyanka, Kavita Khanna, and Vijendra Singh. "Gland segmentation in colorectal cancer histopathological images using U-net inspired convolutional network." Neural Computing and Applications 34, no. 7 (2022): 5383-5395.

[2] Ohata, Elene Firmeza, João Victor Souza das Chagas, Gabriel Maia Bezerra, Mohammad Mehedi Hassan, Victor Hugo Costa de Albuquerque, and Pedro Pedrosa Rebouças Filho. "A novel transfer learning approach for the classification of histological images of colorectal cancer." The Journal of Supercomputing (2021): 1-26.

[3] Yeung, Michael, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. "Focus U-Net: A novel dual attention-gated CNN for polyp segmentation during colonoscopy." Computers in biology and medicine 137 (2021): 104815.

[4] Li, Donglin, Jiacan Xu, Jianhui Wang, Xiaoke Fang, and Ying Ji. "A multi-scale fusion convolutional neural network based on attention mechanism for the visualization analysis of EEG signals decoding." IEEE Transactions on Neural Systems and Rehabilitation Engineering 28, no. 12 (2020): 2615-2626.

[5] Wang, Yan, Zixuan Feng, Liping Song, Xiangbin Liu, and Shuai Liu. "Multiclassification of endoscopic colonoscopy images based on deep transfer learning." Computational and Mathematical Methods in Medicine 2021 (2021).

[6] Tanwar, Sushama, and S. Vijayalakshmi. "Comparative analysis and proposal of deep learning based colorectal cancer polyps classification technique." Journal of Computational and Theoretical Nanoscience 17, no. 5 (2020): 2354-2362.

[7] Gour, Neha, and Pritee Khanna. "Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network." Biomedical Signal Processing and Control 66 (2021): 102329.

[8] Huang, Kaimei, Binghu Lin, Jinyang Liu, Yankun Liu, Jingwu Li, Geng Tian, and Jialiang Yang. "Predicting colorectal cancer tumor mutational burden from histopathological images and clinical information using multi-modal deep learning." Bioinformatics 38, no. 22 (2022): 5108-5115.

[9] Dabass, Manju, Sharda Vashisth, and Rekha Vig. "A convolution neural network with multi-level convolutional and attention learning for classification of cancer grades and tissue structures in colon histopathological images." Computers in biology and medicine 147 (2022): 105680.

[10] Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA: a cancer journal for clinicians 68, no. 6 (2018): 394-424.

[11] Liang, Meiyan, Ru Wang, Jianan Liang, Lin Wang, Bo Li, Xiaojun Jia, Yu Zhang, Qinghui Chen, Tianyi Zhang, and Cunlin Zhang. "Interpretable Inference and Classification of Tissue Types in Histological Colorectal Cancer Slides Based on Ensembles Adaptive Boosting Prototype Tree." IEEE Journal of Biomedical and Health Informatics (2023).

[12] Mukhlif, Abdulrahman Abbas, Belal Al-Khateeb, and Mazin Abed Mohammed. "An extensive review of state-of-the-art transfer learning techniques used in medical imaging: Open issues and challenges." Journal of Intelligent Systems 31, no. 1 (2022): 1085-1111.

[13] Morid, Mohammad Amin, Alireza Borjali, and Guilherme Del Fiol. "A scoping review of transfer learning research on medical image analysis using ImageNet." Computers in biology and medicine 128 (2021): 104115.

[14] Hamida, A. Ben, Maxime Devanne, Jonathan Weber, Caroline Truntzer, Valentin Derangère, François Ghiringhelli, Germain Forestier, and Cédric Wemmert. "Deep learning for colon cancer histopathological images analysis." Computers in Biology and Medicine 136 (2021): 104730.

[15] Chen, Haoyuan, Chen Li, Xiaoyan Li, Md Mamunur Rahaman, Weiming Hu, Yixin Li, Wanli Liu et al. "IL-MCAM: An interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach." Computers in Biology and Medicine 143 (2022): 105265.

[16] Kather, Jakob Nikolas, Frank Gerrit Zöllner, Francesco Bianconi, Susanne M. Melchers, Lothar R. Schad, Timo Gaiser, Alexander Marx, and Cleo-Aron Weis. "Collection of textures in colorectal cancer histology." Zenodo https://doi. org/10 5281 (2016).

Shah Muhammad Imtiyaj Uddin is pursuing his Master's in the Institute of Digital Anti-Aging Healthcare from Inje University. He has previously worked on multiple real-life projects related to mobile applications, computer vision, data sciences, and user interface systems. His research interest aligns with Computer Vision, Artificial Intelligence, and mobile applications.

Md Ariful Islam Mozumder is pursuing his Master's in the Institute of Digital Anti-Aging Healthcare from Inje University. He has previously worked on multiple real-life projects related to computer vision, data sciences, Smart IoT systems, and Natural Language Processing. His research interest aligns with Computer Vision, Artificial Intelligence, Bio Signal Processing, Algorithms, and Metaverse.

Rashedul Islam Sumon is pursuing his Master's in the Institute of Digital Anti-Aging Healthcare from Inje University. He has previously worked on multiple real-life projects related to computer vision, data sciences, Smart IoT systems, and text mining. His research interest aligns with Computer Vision, Artificial Intelligence, Medical Image Processing, Algorithms, and Natural Language Processing.

Moon-Il Joo received a PhD degree in computer engineering from Inje University in 2018. He is currently working as a research professor at the Institute of digital anti-aging Healthcare, Inje University, Korea. His research interests are in software engineering, human-computer-computer interaction, smartphone programming, and component-based development.

Hee-Cheol Kim BSc at the Department of Mathematics, MSc at the Department of Computer Science at SoGang University in Korea, and Ph.D. in Numerical Analysis and Computing Science, at Stockholm University in Sweden. He is a professor and Head of the Department of the Institute. Digital Anti-aging Healthcare, Inje University, S: Korea. His research interests include Machine learning, Text mining, Bioinformatics, Metaverse, Natural Language Processing, and Smart Logistics.

Deep Learning Based Cervical Spine Bones Detection: A case study using YOLO

Muhammad Yaseen*, Maisam Ali*, Sikander Ali*, Ali Hussain*, Ali Athar*, Hee-Cheol Kim*

* Dept. of Digital Anti-Aging Healthcare, Inje University, Gimhae, Republic of Korea

shigriyaseen@gmail.com, maisamali053@gmail.com, sikandershigri77@gmail.com, alihussainrana@gmail.com, ali.athar1401@gmail.com, heeki@inje.ac.kr

Abstract— **Cervical spine bones detection plays a crucial role in various medical applications, such as diagnosis, surgical planning, and treatment assessment. Traditional methods for cervical spine bones detection often rely on manual identification and segmentation, which are time-consuming and prone to errors. In recent years, deep learning approaches have shown great potential in automating the detection process and achieving high accuracy. In this research paper, we propose a deep learning-based approach for detecting cervical spine bones. Our suggested approach employs the YOLOv5 architecture, a cutting-edge object identification system renowned for its effectiveness and precision. The model is trained to recognize and locate bones structures using computed tomography (CT) scans image of the cervical spine as inputs. We conduct extensive evaluations using the trained models on the cervical spine dataset. The mean average precision (mAP) scores achieved by our model are 93% at threshold (mAP_0.5) and 83% at thresholds ranging from (mAP_0.5:0.95), which demonstrate the effectiveness of our approach in accurately detecting and localizing cervical spine bones. Our deep learning-based method for detecting cervical spine bones with high mAP scores presented in this research paper has significant implications for medical applications. With accurate and reliable bones detection, medical professionals can enhance diagnosis, surgical planning, and treatment assessment processes. The achieved mAP scores showcase the performance and potential of our proposed method, contributing to the advancement of bone detection techniques in cervical spine imaging and facilitating collaboration between the medical imaging and deep learning communities.**

Keywords— **Cervical Spine, Bones Detection, Computer Vision, Deep Learning, Computed Tomography (CT), YOLOv5**

I. Introduction

The field of medical imaging has experienced a dramatic revolution in recent years as a result of significant breakthroughs in deep learning and computer vision technology. Among the many applications of these breakthroughs, the recognition and interpretation of anatomical features inside medical images has received a lot of attention [1]. The detection of cervical spine vertebrae is critical in the diagnosis and treatment of spinal diseases and injuries [2]. The cervical spine, which consists of seven vertebrae labelled C1 to C7, is crucial in supporting the skull, protecting the spinal cord, and allowing for a wide variety of head motions [3]. The accurate and efficient identification of these vertebrae is critical for the examination of disorders such as fractures, degenerative disc diseases, and congenital anomalies [4]. Traditionally, radiologists were responsible for manually identifying cervical vertebrae in radiography pictures, a time-consuming and error-prone operation that might be vulnerable to inter-observer variability. To overcome these issues, deep learning-based techniques for automated object detection have emerged as viable alternatives. In this regard, the You Only Look Once (YOLO) family of object detection models, specifically YOLOv5, has attracted attention due to its real-time capabilities and improved accuracy in object localization tasks [5]. Because of its ability to detect many objects at the same time with a single network pass, it is a good option for automating the recognition of cervical spine vertebrae in CT scan images.

This study investigates the use of YOLOv5 in medical imaging, specifically the detection of cervical spine vertebrae in CT scan images. We intend to use deep learning to create a robust and efficient system that can assist radiologists in detecting anomalies or injuries in the cervical spine with high accuracy and in a short amount of time. We aim to contribute to the ongoing efforts to enhance patient care and outcomes in the fields of radiology and spinal health by fusing the cutting-edge technology of deep learning with the wealth of information present in CT scan images.

II. Materials and Methods

A. Dataset Description

In our study, we used the RSNA 2022 Cervical Spine Fracture Detection Challenge dataset [6]. This comprehensive dataset is made up of 2019 CT scans that are specifically focused on the cervical spine area. The availability of segmentations for a portion of the scans is an intriguing feature of the dataset. The provided segmentation labels range from 1 to 7, representing the seven cervical vertebrae from C1 to C7.

B. Data Pre-processing

We used the segmentation data found in the dataset to help us find the cervical spine bone. This segmentation distinguished the different cervical vertebrae, giving each one a specific numerical label that ranged from 1 to 7, as well as a background label that was specified as 0. The next step was to draw bounding boxes. Using the segmentation information, we precisely delineated bounding boxes around each cervical vertebra. These bounding boxes effectively enclosed the areas of interest inside the images. Most importantly, we labeled the slices according to the cervical spine number for a total of 9170 images. Based on the segmentation information, each image was paired with the numerical label of the cervical vertebra to which it referred. This technique endowed each image with a clear and meaningful label, allowing the YOLOv5 model to be trained for precise recognition of cervical spine bone.

C. *YOLOv5 Network*

YOLO (You Only Look Once) has gained fame as a prominent deep neural model for real-time object detection [7], owed to its robust performance in swiftly identifying objects within a real-time environment. The original YOLO version was introduced by Joseph Redmon and his research team in 2016. It distinguishes itself as a single neural network architecture capable of not only detecting object bounding boxes but also predicting the probability of object classes within an image. Unlike earlier object detection tools such as CNN, RCNN, and Faster-RCN [8, 9], YOLO offers a unique advantage in terms of both speed and accuracy, outperforming even mask-RCNN, which excels in precise object detection but lags in processing speed [10]. YOLO is particularly well-suited for scenarios that demand swift and accurate object detection, such as in military applications where speed and precision are critical. It's a single-network solution known for its robustness and accuracy. Over time, YOLO has seen several iterations, including YOLOv2, YOLOv3, v4, and v5, each enhanced with new features and functionalities [11, 12, 13]. YOLOv5 was released in 2020, and it is fast and highly accurate, with various network architectures and data augmentation techniques to enhance performance [14,15]. The YOLO network architecture is made up of three key parts: the backbone, the neck, and the head. The backbone includes modules such as Focus, Convolution with batch normalization and Leaky ReLU (CBL), Mix Conv (Mix convolution), Cross Stage Partial Network (CSP), and Spatial pyramid pooling (SPP). It receives input images with a resolution of 512 x 512 x 3 via the Focus structure, applies slicing operations to minimize image size, and performs convolutional operations with kernels. The backbone outputs are used as input for the next phase of the YOLOV5 design, the neck network. The Neck Network is made up of CBL, CSP, Concat, and Up-sampling. It produces three outputs, which are used as inputs to the head, also known as the detector. This layer makes predictions by displaying the bounding box around the target items, classification of the objects, and probability of the object belonging to a given class.

III. Experiments

This section presents the experimental details of this research study. The constructed model was able to detect the targeted in the images, namely cervical spine vertebrae. The model accurately labelled the objects and categorized the vertebrae by drawing bounding boxes around them.

A. *Training YOLOv5*

The Yolov5 versions were trained in this study, with YOLOv5l being the fastest among them. The dataset used for training consisted of 9170 images, with 80% of the data allocated for model training. A batch size of 16 and an image size of 512 x 512 were chosen. The training utilized a learning rate of 0.01, with a learning rate decay of 0.0005, and the SGD optimizer was employed for network optimization. Training was performed using the train.py script, specifying parameters such as epochs, batch size, and model weights. To facilitate model training, specific folder structures were required. The dataset folders were divided into two, one containing image and the other containing labels in text file format. These label files provided information about object classes, bounding boxes, coordinates, height, and width of the bounding boxes, with one class of objects per line. The experimental parameters are detailed in Table 1 below.

TABLE 1. EXPERIMENTAL PARAMETERS

Parameters	Values
Epochs	200
Image size	512
Learning rate	0.01
Batch size	16
Optimizer	SGD
Momentum	0.937
Weight decay	0.0005

B. *Result and Discussion*

The study used all four versions of YOLOv5, namely YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The findings of a comparative analysis revealed that YOLOv5x performed better than the other models. Values of 0.869, 0.906, 0.933, and 0.839 were obtained for the precision, recall, mAP_0.5, and mAP_0.95 measures, respectively.

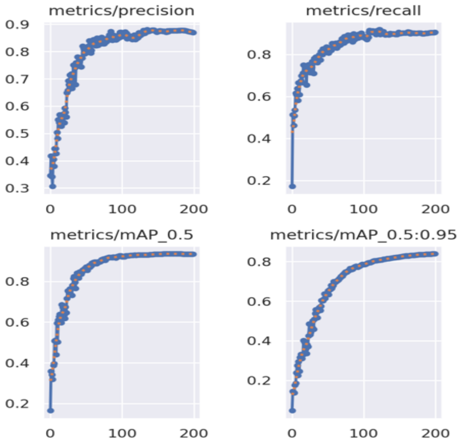


Figure 1. Precision, recall, mAP_0.5, and mAP_0.5:0.95 of YOLOv5x

In Figure 1, the precision value exhibited a gradual increase from 0.346 to reach 0.869, while the recall initially started at 0.171, experienced a sudden rise, and then stabilized with increasing epochs, ultimately reaching 0.906. Additionally, Figure 1 also illustrates mAP @0.5 and mAP @0.5:0.95, which were measured at 0.933 and 0.839, respectively.

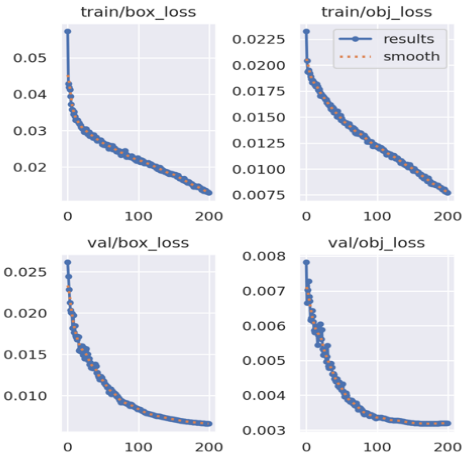


Figure 2. Box loss and object loss for the training and validation, respectively.

In Figure 2, both box loss and object loss are depicted for both training and validation phases. During training, the box loss steadily decreases with increasing epochs, starting at 0.057 at epoch 0 and reaching 0.012 by epoch 200. Similarly, object loss decreases from 0.023 to 0.0075 as the model reaches 200 epochs. In contrast, during validation, the box loss decreases slightly from 0.026 to 0.06 as the model reaches 200 epochs, showing minimal improvement. Concurrently, object loss starts at 0.0028 during validation and gradually decreases with increasing epochs, reaching 0.0027 by epoch 200.

We employed YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x models and conducted training using the cervical spine dataset. The performance of these models was assessed using various evaluation metrics, and the outcomes are presented in Table 2. This table offers a comparative analysis of different YOLO architectures, showcasing the results through the evaluation of parameters such as precision and recall.

TABLE 2. COMPARISON OF YOLO MODELS

Model	Precision	Recall	mAP_0.5	mAP_0.95
YOLOv5s	0.825	0.857	0.892	0.715
YOLOv5m	0.843	0.909	0.921	0.794
YOLOv5l	0.862	0.887	0.919	0.803
YOLOv5x	0.869	0.906	0.933	0.839

Performance metrics including mAP, precision, and recall were computed for all four YOLO architectures, and the experimental findings revealed that YOLOv5x outperformed other YOLO versions. Specifically, YOLOv5xl achieved an mAP_0.5 of 93.3 and an mAP_0.95 of 83.9, highlighting its superior performance.

Figure 3 provides a visual representation of the detection outcomes achieved on the validation dataset, and the remarkable level of accuracy depicted in the results serves as a strong indicator of the exceptional performance exhibited by our model.

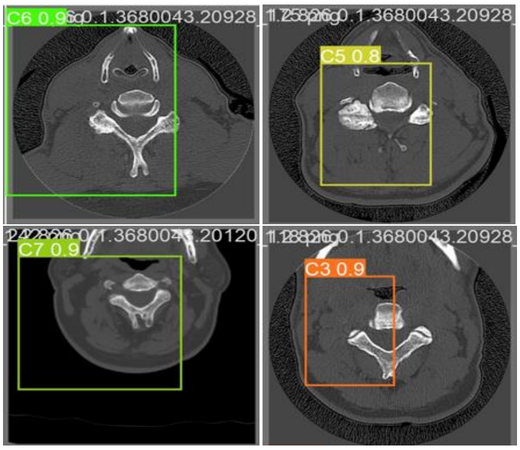


Figure 3. Detection of cervical spine vertebrae

Figure 3 depicts our model's ability to detect cervical bones, specifically C3, C5, C6 and C7. The higher level of accuracy displayed in these detections demonstrates the model's ability to properly recognize these anatomical features within the cervical spine. This level of precision is critical in medical applications where perfect localization of individual vertebrae is required for diagnosis and therapy planning.

Figure 4 displays the confusion matrices of YOLOv5x on our validation dataset. A confusion matrix is a table that is used to evaluate the performance of a classification model. It shows the number of correct and incorrect predictions made by the model compared to the actual outcomes. The matrix is made up of four components: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These confusion matrices provide a comprehensive overview of the YOLOv5 model's performance by showing the alignment between actual classes and predicted classes.

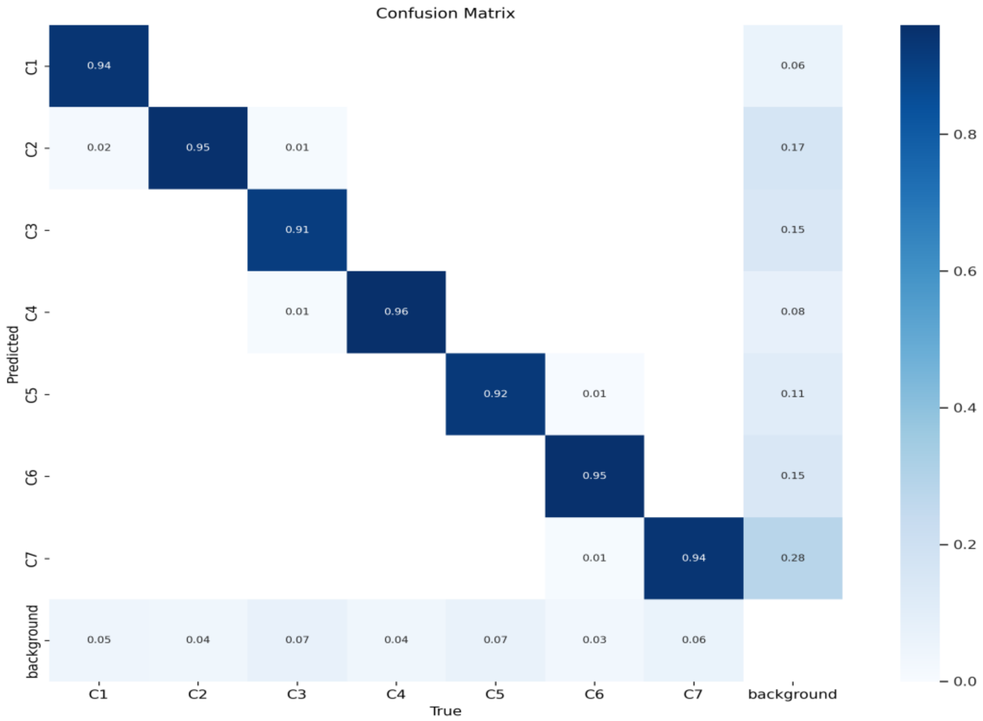


Figure 4. Confusion matrix of YOLOv5x

IV. Conclusions

In this study, using real-time object detection models based on deep learning, we explored the use of YOLO algorithms for the detection of cervical spine structures. The YOLOv5 model, which uses the Pytorch framework and convolutional neural networks as its foundation for feature extraction and object detection, stood out among the other YOLO versions as a highly effective object detection model. We used four distinct YOLOv5 architectures, with YOLOv5x performing particularly well, to identify the cervical spine bones. This model exhibits the capacity to confidently and properly identify vertebrae, which can be a huge help to medical experts when making diagnoses and formulating treatment plans.

Acknowledgment

This research was supported by the MSIT (Ministry of Science ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) in 2022 (2022-0-01091, 1711175863).

References

[1] Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019 Oct 4;7:e7702. doi: 10.7717/peerj.7702. PMID: 31592346; PMCID: PMC6779111.

[2] Zanza C, Tornatore G, Naturale C, Longhitano Y, Saviano A, Piccioni A, Maiese A, Ferrara M, Volonnino G, Bertozzi G, Grassi R, Donati F, Karaboue MAA. Cervical spine injury: clinical and medico-legal overview. Radiol Med. 2023 Jan;128(1):103-112. doi: 10.1007/s11547-022-01578-2. Epub 2023 Jan 31. PMID: 36719553; PMCID: PMC9931800.

[3] Curtis, L., & Ammerman, J. M. (2022, January 4). Cervical Spine Anatomy (Neck). https://www.healthcentral.com/condition/neck-pain/cervical-spine-anatomy-neck .

[4] Panda A, Das CJ, Baruah U. Imaging of vertebral fractures. Indian J Endocrinol Metab. 2014 May;18(3):295-303. doi: 10.4103/2230-8210.131140. Erratum in: Indian J Endocrinol Metab. 2014 Jul;18(4):581. PMID: 24944921; PMCID: PMC4056125.

[5] Wu S, Wang J, Liu L, Chen D, Lu H, Xu C, Hao R, Li Z, Wang Q. Enhanced YOLOv5 Object Detection Algorithm for Accurate Detection of Adult Rhynchophorus ferrugineus . Insects . 2023; 14(8):698. https://doi.org/10.3390/insects14080698.

[6] RSNA 2022 Cervical Spine Fracture Detection | Kaggle. (n.d.). RSNA 2022 Cervical Spine Fracture Detection | Kaggle. https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection/data.

[7] WANG, Xiaolong; SHRIV ASTAVA, Abhinav; GUPTA, Abhinav. A - fast - rcnn: Hard posi tive generation via adversary for o bject detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017 . p. 2606 - 2615.

[8] Ren, Y., Zhu, C., & Xiao, S. (2018). Object detection based on fast/faster RCNN employing fully convolutional architectures. Mathematical Problems in Engineering, 2018, 1-7.

[9] Peng, H., & Chen, S. (2019). BDNN: Binary convolution neural networks for fast object detection. Pattern Recognition Letters, 125, 91-97.

[10] Haralick, R. M., & Shapiro, L. G. (1985). Image segmentation techniques. Computer vision, graphics, and image processing, 29(1), 100-132.

[11] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[12] Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).

[13] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[14] Arifando R, Eto S, Wada C. Improved YOLOv5-Based Lightweight Object Detection Algorithm for People with Visual Impairment to Detect Buses. Applied Sciences. 2023; 13(9):5802.

[15] S. Ali, Abdullah, A. Athar, M. Ali, A. Hussain and H. -C. Kim, "Computer Vision-Based Military Tank Recognition Using Object Detection Technique: An application of the YOLO Framework," 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), Jeddah, Saudi Arabia, 2023, pp. 1-6, doi: 10.1109/ICAISC56366.2023.10085552.

Mr. Muhammad Yaseen received his B.E degree in Electrical Engineering from Hamdard University, Pakistan. He is currently pursuing his master’s degree from Inje University. His research interests are artificial intelligence, machine learning, deep learning, computer vision and medical imaging.

Mr. Maisam Ali received his B.E degree in Electrical and communication Engineering from Hamdard University, Pakistan. He is currently pursuing his master’s degree from Inje University. His research interests are artificial intelligence, machine learning, deep learning, computer vision.

Mr. Sikandar Ali received his B.E. degree in Computer Engineering from Mehran University of Engineering & Technology, Pakistan. He got his MS from the Department of Computer Science from Chung Buk National University, the Republic of Korea. Furthermore, he is now a Ph.D. candidate at Inje University South Korea in the department of Digital Anti-Aging Healthcare. His research interests include artificial intelligence, data science, big data, machine learning, deep learning, reinforcement learning, Computer vision, and medical imaging.

Mr. Ali Hussain received his BSCS. degree in Computer Science from Government College University Faisalabad (GCUF), Pakistan in 2019.Furthermore, he got his master’s from Inje University South Korea in the department of Digital Anti-Aging Healthcare. Currently doing a PhD from the same department. His research interests include artificial intelligence, data science, big data, machine learning, deep learning, Computer vision, reinforcement learning, and medical imaging.

Mr. Ali Athar received his BSSE degree Software Engineering from Government College University Faisalabad (GCUF), Pakistan. He received his MS degree from NUST, Pakistan. He is pursuing his Ph.D. Degree from the Institute of Digital Anti-aging and healthcare at Inje University. His research areas include Text Mining, Machine learning, and Deep learning.

Prof. Hee-Cheol Kim received his BSc at the Department of Mathematics, MSc at the Department of Computer Science in Sogang University in Korea, and Ph.D. at Numerical Analysis and Computing Science, Stockholm University in Sweden in 2001. He is a Professor at the Department of Computer Engineering and Head of the Institute of Digital Anti-aging Healthcare, Inje University in Korea. His research interests include machine learning, deep learning, Computer vision, and medical informatics.

Vision Transformer-based Model for Gastric Cancer Detection and Classification using Weakly Annotated Histopathological Images

Tagne Poupi Theodore Armand*, Subrata Bhattacharjee**, Hyun-Joong Kim*, Ali Hussain*, Sikandar Ali*, Heung-Kook Choi**, and Hee-Cheol Kim*

*Institute of Digital Anti-Aging Healthcare, Inje University, Gimhae 50834, Republic of Korea

**Department of Computer Engineering, u-AHRC, Inje University, Gimhae 50834, Republic of Korea

poupiarmand2@gmail.com, subrata_bhattacharjee@outlook.com, play97509@gmail.com, alihussain.dream@gmail.com, sikandarshigri77@gmail.com, cschk@inje.ac.kr, heeki@inje.ac.kr

Abstract— **Gastric Cancer (GC) is the fifth most diagnosed cancer worldwide. An early diagnosis is a hope for patients suffering from GC. A biopsy is a procedure that helps detect abnormal and suspicious areas to determine whether cancer cells are in the stomach. Tissue samples collected through biopsy are stained using Hematoxylin and Eosin (H&E) and digitalized through scanning to produce a whole slide image (WSI) needed for further analysis. Recently, most prognostics have proven effective using artificial intelligence techniques combined with related computer aid detection systems. This research used a vision transformer to detect and classify gastric cancer from weakly annotated tissue images. After acquiring normal and cancer histopathological samples, we applied the vision transformer (ViT) model for binary classification. We generalized our approach by performing region-based prediction on unannotated tissue samples. The proposed approach will ease diagnosis and support pathologists in decision-making.**

Keywords— **Gastric Cancer, Vision Transformers, Whole Slide Images, Weakly Annotated Images.**

I. Introduction

Computer Aids Detection (CAD) uses various algorithms with different technologies to perform specific tasks such as pattern analysis, anomaly detection, or matches between data. CAD has proven effective and efficient for multiple applications in many industries. In healthcare, CAD using artificial intelligence has demonstrated outstanding capabilities [1]. Machine learning and artificial intelligence algorithms have been applied to detect and classify diseases using customized and pre-trained models [2]. The data collection process is carried out by medical personnel, after which data undergoes necessary preprocessing and must be fed to the model for training using specific algorithms according to the task to be accomplished. For instance, machine learning algorithms have demonstrated some capability in handling disease prediction [3-5].

Furthermore, deep learning methods have expanded the possibilities by offering chances to develop algorithms capable of improving diagnosis at optimal speed [6, 7]. One of the latest deep learning-based algorithms incorporating these features is Vision Transformers (ViTs). ViTs are an innovative approach used to handle image analysis tasks using the transformer architecture. Though the earlier transformer proposed by Vaswani et al. [8] introduced the transformer architecture applicable to natural language processing tasks, vision transformers by Dosovitskiy et al. [9] extended their application to image processing tasks.

In medical imaging, various images (X-ray, CT, MRI, Ultrasound, WSI…) are used for disease diagnoses, prognoses, and treatments. Histopathology Whole Slide Images (WSI) are commonly used for analysis and are often obtained through biopsy. An organ is collected from the patient through biopsy and then stained using Hematoxylin and Eosin (H&E) to prepare the tissue for digitalization through a digital scanner. WSI is being processed using many algorithms, such as machine learning and deep learning algorithms, including CNN and vision transformers. Vision transformers have proven effective for many computer vision tasks, including medical image classification, object detection, and image segmentation. Ikromjanov et al. proposed a vision transformer-based model to detect prostate cancer [10]. The proposed model was built using histopathological images and could detect and grade prostate cancer according to the Gleason grading system using pre-trained ViTs and prostate WSI.

Similarly, Kollias et al. [11] proposed a Covid-19 diagnosis tool developed with ViTs. This model was able to detect region-based lesions on CT scan images. Zeid et al. [12] developed a vision transformer and CNN-based model known as compact convolutional transformer to achieve the classification of colorectal cancer using histopathology images. ViTs have also been influential in the segmentation tasks. An efficient Coronary artery segmentation was done by Ning et al. using vision transformers [13]. Their model consisted of three main modules: one for feature extraction, another for modeling the features to high-level texture features, and a last module for context aggregation. The experimental result demonstrated good performance with a more than 75% dice coefficient. Table 1 below shows more applications of vision transformers in medical imaging tasks.

TABLE 1: APPLICATION OF VISION TRANSFORMERS IN MEDICAL IMAGING

SN	Description	Archi.	Ref.
1	Classification of retail disease using fundus images and ViTs	Base	[14]
2	Multi-scale Hybrid Vision Transformer for Learning Gastric Cancer Histology	Hybrid	[15]
3	Applying ViTs for Breast ultrasound image classification	Base	[16]
4	MIL transformer-based model for classification of histological Whole Slide images	Hybrid	[17]
5	Multiclass Colorectal Cancer Histology Images Classification Using Vision Transformers	Base	[12]

In this research, we use the base vision transformer with weakly annotated areas of WSI to detect and classify gastric cancer. The main contribution of this work is to use weakly annotated regions of whole slide images and ViTs to efficiently classify gastric cancer histopathological images in order to support physician diagnostics and decision support systems.

II. DATASET

The histopathological samples used in this experiment were sourced from the GasHisSDB dataset [26], accessible at https://gitee.com/neuhwm/GasHisSDB. This dataset comprises images with diverse imaging characteristics and is publicly available for research. The histopathological samples were curated by four pathologists from Longhua Hospital, Shanghai University of Traditional Chinese Medicine, resulting in 560 cancerous images and 140 normal images, each with a resolution of 2048×2048 pixels. The dataset was prepared through H&E staining at a magnification of 20X using Nikon (Japan) and Olympus (Japan) Microscopes. These images were extracted from original microscopic biopsy images related to gastric lesions, specifically highlighting cancerous regions. The dataset has been well-annotated by medical experts. The example samples of the dataset are shown in Figure 1.

III. METHODS

In this research, we used 2048x2048 weakly annotated region extracted from a whole slide image of gastric cancer. The weak annotation provides region-level labels of patients with cancer or normal, used for disease detection and classification using the vision transformer model. To achieve our goal, we had to prepare the data and feed it into the vision transformer to train the model that will serve as a region-level classifier of the WSI patches, which will help to achieve the complete classification of any given gastric cancer WSI into the cancer and normal classes.

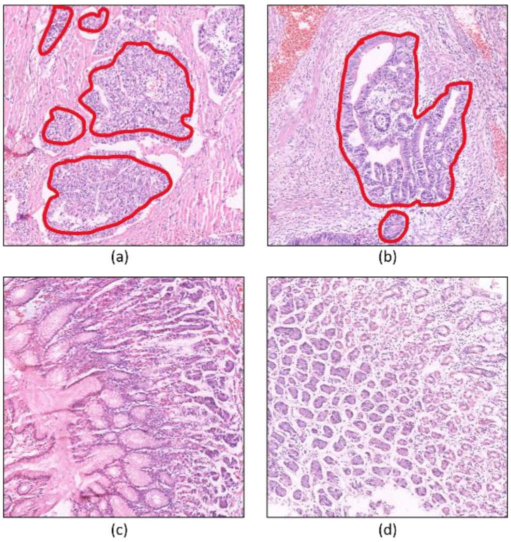


Figure 1. Sample images of the dataset. (a, b) Cancer samples with annotated regions. (c, d) Normal samples.

A. Data Preprocessing

For each region of interest (ROI) annotated by the pathologist, we conducted manual weak annotation for weakly supervised learning. Tissue samples of size 2048×2048 were generated from these weakly annotated ROIs. The patching process enhances scalability and facilitates efficient analysis by allowing precise examination of specific areas. Utilizing these patch images reduces computational requirements, resulting in more optimal learning, inferencing, and likelihood estimation times than directly manipulating whole slide images (WSIs). The preprocessing involves a series of operations on the patch images to prepare them for analysis. This includes removing the background, as illustrated in Figure 2, to gain insights into the image information.

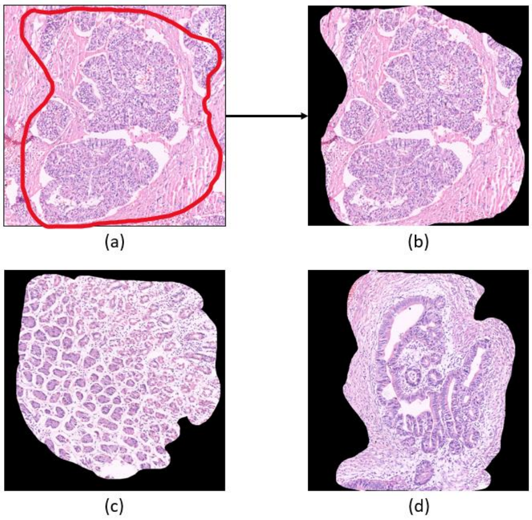


Figure 2. Preprocessed tissue samples. (a) Weakly annotated tissue sample. (b) Extracted region of interest from (a) by removing the background. (c) Example sample of normal tissue. (d) Example samples of cancer tissue.

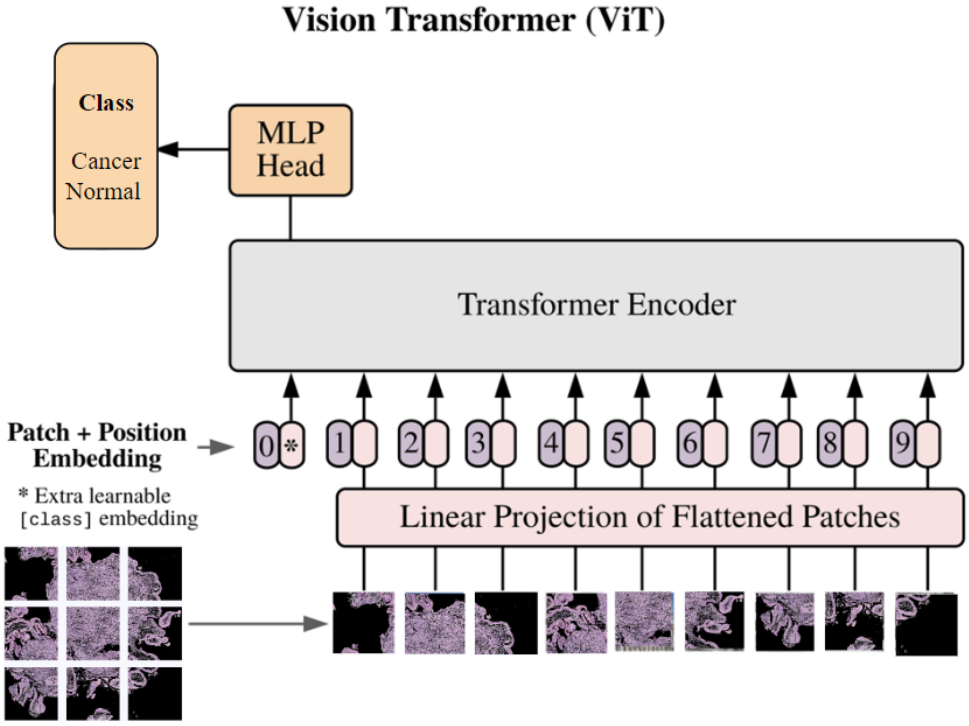

Vision Transformer (ViT)
Class
Cancer
Normal
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches

Figure 3. Architecture of Vision transformer.

B. *Vision transformer*

Vision transformers are deep learning models designed for computer vision-related tasks (detection, classification, and segmentation). The success of the transformers model in natural language processing inspired the vision transformer presented by Dosovitskiy et al. [9]. Traditionally, computer vision tasks were mostly achieved using CNN; nowadays, the vision transformers are a new approach recently gaining much attention from industry and academia. ViTs use a self-attention mechanism that allows the computer to understand the relationship between various parts of an image. It assesses the importance of the patch score and focuses on the most relevant information helpful for the model. The patch size defined in the original paper was set to 16×16 and can be changed. In this study, we considered a 16×16 patch size, generating 16,3884 patches for a single input image of size 2048×2048. The input images are tokenized and flattened; in this process, the 2-D patch of 16×16 is converted into a 1-D array of 256 tokens that will be input into the linear projection layer. The linear projection layer will transform the 1-D vector into lower-dimensional vectors while preserving relevant information. For any 1-D vector x_1, the weight W and a bias b obtained during training produce an output expressed as $Wx_1 + b$ representing the lower dimensionality transform vector. The lower dimensionality vectors are the reduction in the number of features used to describe an object. This process benefits our algorithm by reducing noise, computational power, and memory requirements, thus increasing speed. Position embedding takes the lower dimensional vector as input and provides the position information to the transformer encoder. Unlike the original transformer, the ViTs don't have a decoder part. The encoder transformer's first layer is the self-attention layer (Multi-Head Attention Layer), which captures dependencies between patches and enables the model to consider the global context. The output of each patch passes through a feed-forward network that captures the complex non-linear relationship between the patches. The next stage, at the final layer, is the classification layer, which matches the output of the transformer layer (Multilayer Perceptron head) and the desired output (cancer/normal classification in our case).

IV. Results and Discussion

In this paper, we focused on the various hospital cohorts for training and testing using the deep learning framework, the ViT model, to classify region-level whole slide images of cancer and normal. The ViT model used in the study provided some sound results in classifying region-level histopathologic images containing aggressive and nonaggressive tumor cells. The experiment was also carried out using CNN-based algorithms, notably ResNet50, DenseNet121, and EfficientNetB1, which were chosen for their effectiveness in medical image classification tasks.

The ViT-16 model prompted a test accuracy of 85.9%, outperforming all other models. The overall performance is of all the models are shown in Table 2. The performance metrics used for model evaluation are accuracy, precision, recall, and f1-score, computed as shown in equations 1-4. Moreover, we split the dataset into training and validation with an 80:20 ratio while reserving 40 independent weakly annotated slides for testing purposes.

TABLE 2: OVERALL CLASSIFICATION PERFORMANCE OF THE VIT MODEL

Class	Accuracy	Precision	Recall	F1-score
ResNet50	70.59	0.68	0.63	0.64
DenseNet121	74.51	0.75	0.73	0.74
EfficientNetB1	78.71	0.76	0.81	0.76
ViT	**85.9**	**0.77**	**0.85**	**0.78**

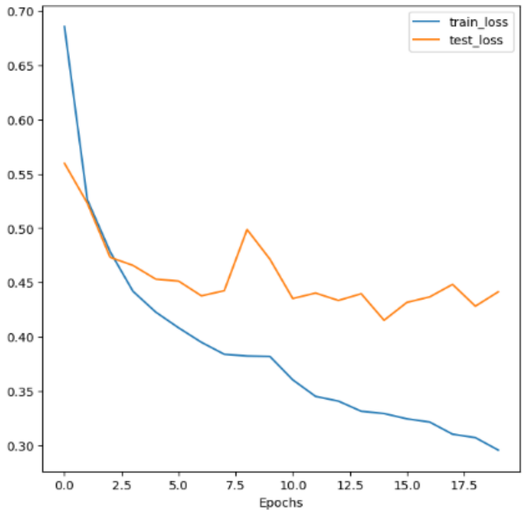
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

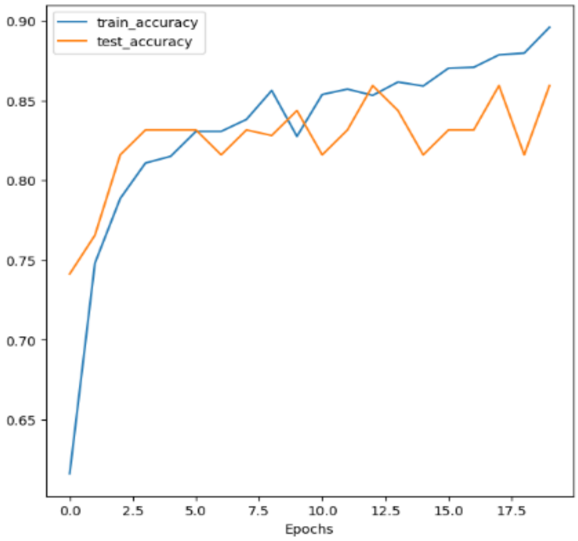
$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

$$F1-score \; = 2 * \frac{Precision*Recall}{Precision + Recall} \qquad (4)$$

The results of our ViT model show that it can achieve a competitive performance and be effectively used in classifying cancer and normal gastric cancer cases. From the pathological perspective, cancer cells are graded based on their level of differentiation. Well-differentiated cells closely resemble normal, healthy cells, while poorly differentiated cells look very different from normal cells and tend to be more aggressive. Poorly differentiated adenocarcinoma is associated with a higher grade and is often more aggressive in its growth and spread. The proposed model aligns with the pathological analysis, and some observations can be made on the output image. Our model was trained with the most suitable hyperparameters with a few epochs set to $n = 20$. We evaluated the performance of our model with a variety of metrics. Figure 4 illustrates the model learning graphs of accuracy and loss that give insights into the training process and performance. The model's efficiency can also be observed from the confusion matrix shown in Figure 5. From the confusion matrix, we can see that there are a few mispredictions, and it is because of the similarity samples of cancer and normal; as a result, the model did not perform well for some specific images.



(a)



(b)

Figure 4. Learning curves for the ViT model. (a) Loss curve; (b) Accuracy curve.

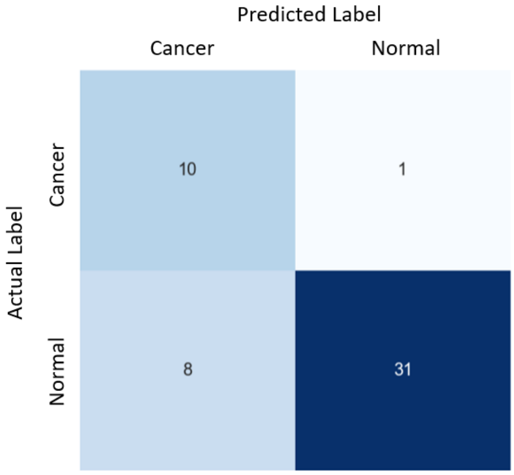


Figure 5. Confusion matrix of the ViT-based model

The model we used in the work showed astounding results for the cancer class by providing an overall accuracy and recall of 0.85 and 0.85, respectively. The ViT-based classification models are very much necessary for classifying region-level WSI images which can assist pathologists in the identification of tumor and non-tumor cells in H&E histopathological samples. There are many existing papers on weakly supervised approaches where they performed patch-level multiple-instance learning (MIL) for binary classification. However, we presented a state-of-the-art method for weakly supervised learning by feeding the global features to the model extracted from region-level weakly annotated gastric cancer samples. This approach was carried out to bypass the information loss that occurs in other patch-based MIL approaches.

V. Conclusions

In conclusion, this paper proposes a method to detect and classify gastric cancer using ViTs and weakly annotated histopathological images. Unlike the CNN approaches that use low-level to high-level feature extractor mechanisms for classification with effects on computational and statistical efficiency, the vision transformer uses the attention mechanism to overcome most CNN challenges while reducing dependencies powered by its parallel processing architecture. The obtained results showed the potential to assist the pathologist in decision-making. Region-level segmentation can be further used to compare our approach with fully supervised learning-based methods.

Acknowledgment

This research was supported by a grant from Korea Health Technology R&D Project through the Korea Health Industry Development Institute validation (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant No: HI21C0977)

References

[1] Shiraishi, Junji, et al. "Computer-aided diagnosis and artificial intelligence in clinical imaging." *Seminars in nuclear medicine*. Vol. 41. No. 6. WB Saunders, 2011.

[2] Ahsan, Md Manjurul, Shahana Akter Luna, and Zahed Siddique. "Machine-learning-based disease diagnosis: A comprehensive review." *Healthcare*. Vol. 10. No. 3. MDPI, 2022.

[3] Ansari, A.Q.; Gupta, N.K. Automated diagnosis of coronary heart disease using neuro-fuzzy integrated system. In Proceedings of the 2011 World Congress on Information and Communication Technologies, Mumbai, India, 11–14 December 2011; pp. 1379–1384.

[4] Levey, A.S.; Coresh, J. Chronic kidney disease. Lancet 2012, 379, 165–180.

[5] Vidushi, A.R.; Shrivastava, A.K. Diagnosis of Alzheimer disease using machine learning approaches. Int. J. Adv. Sci. Technol. 2019, 29, 7062–7073.

[6] Serag, Ahmed, et al. "Translational AI and deep learning in diagnostic pathology." *Frontiers in medicine* 6 (2019): 185.

[7] Litjens, Geert, et al. "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis." *Scientific reports* 6.1 (2016): 26286.

[8] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[9] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[10] Ikromjanov, Kobiljon, et al. "Whole slide image analysis and detection of prostate cancer using vision transformers." *2022 international conference on artificial intelligence in information and communication (ICAIIC)*. IEEE, 2022.

[11] Kollias, Dimitrios, et al. "Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[12] Zeid, Magdy Abd-Elghany, Khaled El-Bahnasy, and S. E. Abo-Youssef. "Multiclass colorectal cancer histology images classification using vision transformers." *2021 tenth international conference on intelligent computing and information systems (ICICIS)*. IEEE, 2021.

[13] Ning, Yang, et al. "Cac-emvt: Efficient coronary artery calcium segmentation with multi-scale vision transformers." *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021.

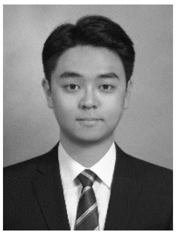
[14] Yu, Shuang, et al. "Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification." *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*. Springer International Publishing, 2021.

[15] Oh, Yujin, et al. "Multi-Scale Hybrid Vision Transformer for Learning Gastric Histology: AI-Based Decision Support System for Gastric Cancer Treatment." *IEEE Journal of Biomedical and Health Informatics* (2023).Gheflati, Behnaz, and Hassan Rivaz. "Vision transformers for classification of breast ultrasound images." *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022.

[16] Shao, Zhuchen, et al. "Transmil: Transformer based correlated multiple instance learning for whole slide image classification." *Advances in neural information processing systems* 34 (2021): 2136-2147.

Tagne Poupi Theodore Armand was born in Cameroon and received an MSc in Information Systems and Networking at ICT University USA, Cameroon Campus, in 2021. He is a Ph.D. research scholar at the Institute of Digital Anti-aging and healthcare at Inje University. His research interest field includes image processing, focusing on medical image analysis, deep learning, machine learning, metaverse, and digital technology.

Subrata Bhattacharjee received his B.S. in information technology (IT) from the University of Derby, UK 2016. He is pursuing his M.S. leading Ph.D. from the Graduate School of Computer Engineering, Inje University, Korea. He is a Research Scholar and Teaching Assistant at the Medical Image Technology Laboratory (MITL) at Inje University. His research interests include multimodal medical imaging, tissue and cell image analysis, digital pathology, image reconstruction, cell and gland segmentation, machine learning, and deep learning.



Hyun Joong Kim was born in the Republic of Korea in 1994. He received his Bachelor's Degree in Nursing from Yonsei University in 2019. He is currently pursuing a Master's Degree majoring in artificial intelligence in healthcare at the Institute of Digital Anti-aging and healthcare, Inje University. His research activity is concentrated on Image Processing, Deep Learning, and Machine Learning.

Ali Hussain received his BSCS degree in computer science from Government College University Faisalabad (GCUF), Pakistan, in 2019. Furthermore, he got his master's from InjeUniversity South Korea in the Department of Digital Anti-Aging Healthcare. Currently doing PhD from the same department. His research interests include artificial intelligence, data science, Big data, machine learning, deep learning, Computer vision, reinforcement learning, and medical imaging.

Sikandar Ali received his B.E. degree in Computer Engineering from Mehran University of Engineering & Technology, Pakistan. He got his MS from the Department of Computer Science from Chungbuk National University, the Republic of Korea. Furthermore, he is now a Ph.D. candidate at Inje University South Korea in the Department of Digital Anti-Aging Healthcare. His research interests include artificial intelligence, data science, big data, machine learning, deep learning, reinforcement learning, Computer vision, and medical imaging.

Heung Kook Choi received his Bachelor of Engineering and Master of Engineering from Linköping University, Sweden, in 1988 and 1990, respectively. In 1996, he pursued his Ph.D. at the Centre for Image Analysis in Computerized Image Analysis from Uppsala University, Sweden. He has been a Professor at Inje University from 1997 to 2020. He is an Emeritus Professor of Computer Engineering at Inje University and operates the Medical Image Technology Laboratory (MITL). He is the author of eight books, with more than 400 articles and more than 40 inventions. His research interests include computer graphics, multimedia, image processing, and analysis.

Hee-Cheol Kim received his BSc at the Department of Mathematics, MSc at the Department of Computer Science at SoGang University in Korea, and Ph.D. in Numerical Analysis and Computing Science at Stockholm Stockholm University in Sweden in 2001. He is a professor at the Department of Computer Engineering and Head of the Institute of Digital Anti-aging Healthcare Inje University in Korea. His research interests include machine learning, deep learning, Computer vision, and medical informatics.

Overview of the potentials of multiple instance learning in cancer diagnosis: Applications, challenges, and future directions

Tagne Poupi Theodore Armand*, Subrata Bhattacharjee**, and Hee-Cheol Kim*
*Institute of Digital Anti-Aging Healthcare, Inje University, Gimhae 50834, Republic of Korea
**Department of Computer Engineering, u-AHRC, Inje University, Gimhae 50834, Republic of Korea
poupiarmand2@gmail.com, subrata_bhattacharjee@outlook.com, heeki@inje.ac.kr

Abstract— **The outcome of cancer patients mostly depends on the diagnosis process and the treatment strategies. Computer-aided diagnosis (CAD) methods have demonstrated the potential to handle accurate diagnostics using artificial intelligence techniques such as machine learning and deep learning. The nature of the data used in training the AI-based model determined the paradigm, often classified as supervised and unsupervised learning for scenarios with labeled and unlabeled data, respectively. Due to the cost of time and resources, most datasets are nowadays partially labeled and used for training. The weakly supervised learning approach enables the AI models to be trained with incompletely labeled, noisy, or imbalanced data. In recent years, multiple instance learning (MIL) has emerged as a promising weakly supervised learning approach in many fields, including cancer diagnosis. Unlike traditional supervised learning methods, MIL allows the classification of groups of instances, known as bags, where only the bag's label is available. This comprehensive review aims to provide an in-depth analysis of the applications of MIL in cancer diagnostic tasks, highlighting its advantages, challenges, and future directions. By examining these advantages, challenges, and future trends, the review aims to contribute to advancing MIL as a powerful tool in improving cancer diagnostic accuracy and patient outcomes.**

Keywords— **Bag, Cancer diagnosis, Instance, Multiple Instance Learning, Weakly supervised learning**

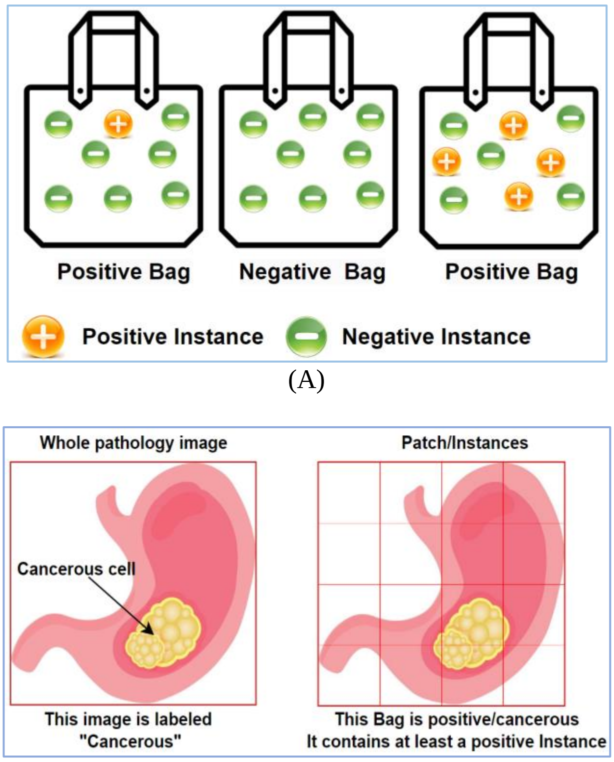
I. Introduction

Cancer remains a global health challenge, with significant implications for patient outcomes and healthcare systems. According to a World Cancer Research Funds International report, 18.1 million cancer cases were estimated in 2020, with 1,806,590 new cases in the United States. Cancers are named for the area where the cancerous cells originated (breast, liver, colon, etc.), even if they spread to other body parts (metastasis). The most common cancers include breast cancer, lung and bronchus cancer, prostate cancer, colon and rectum cancer, melanoma of the skin, bladder cancer, non-Hodgkin lymphoma, kidney and renal pelvis cancer, endometrial cancer, leukemia, pancreatic cancer, thyroid cancer, and liver cancer [1]. Despite advances in cancer research, the mortality rate of cancer patients is still alarming, rendering it a severe threat to human life [2]. Accurate and timely cancer diagnosis is crucial in determining appropriate treatment strategies and improving patient survival rates [3]. However, cancer diagnosis presents numerous challenges due to the complexity and heterogeneity of tumors and the limitations of traditional diagnostic methods. In recent years, advanced machine-learning techniques have emerged as powerful tools to address these challenges and enhance cancer diagnosis [4]. These techniques improve early detection of the condition, tumor classification, prognosis and survival predictions, risk assessment, and clinical trial matching, leading to a better patient outcome, thus reducing healthcare costs and advancing our understanding of the disease. Still, cancer diagnostic challenges arise from several factors, including the diversity of cancer types, the variability in disease manifestations, and the limitations of existing diagnostic approaches.

Traditional diagnostic methods, such as histopathological analysis, imaging techniques, and biomarker assessments, rely on subjective interpretations and may have limitations in terms of sensitivity, specificity, and reproducibility [5]. Additionally, while using Computer-aided diagnosis (CAD) and machine learning methods, cancer datasets often exhibit characteristics such as class imbalance, noisy labels, and complex interactions between features, further complicating accurate diagnosis [6]. Overcoming these challenges enables a growing need for advanced machine-learning techniques in cancer diagnosis. Advanced machine learning algorithms can analyze large volumes of complex data and extract meaningful features and patterns to make accurate predictions. By leveraging computational power and advanced algorithms, machine-learning approaches can improve the sensitivity and specificity of cancer diagnosis, enhance risk stratification, and aid in treatment decision-making [7].

Multiple instance learning (MIL) is a machine learning paradigm that has gained attention for its applicability in many areas, including image classification, drug activity prediction, text classification, and object recognition in computer vision, among others [8]. MIL is a weakly supervised learning approach that deals with specific types of problems and is helpful in cases where the supervised methods are unsuitable because of the lack of colossal labels or annotations in the dataset. In MIL, training data is organized into bags with instances. An instance is a small data point, e.g., a patch image.

An instance can be either positive (belonging to a specific target class) or negative (not belonging to the target class). A bag is a collection of instances. A bag is said to be positive if it contains at least one positive instance and negative in the opposite case. Figure 1 illustrates the concept of bags and instances.



(A)

(B)

Figure 1. Illustration of the concept of bags. A bag is labeled positive if one or more instances are positive.

Unlike traditional supervised learning, where each instance is labeled individually, MIL operates on groups of instances known as bags. A whole slide image can be represented in histopathological imaging as a bag of instances and its patches as instances using MIL [8]. Similarly, a bag could represent a region of interest in an image, a patient's medical record, or a collection of molecular features [9]. The fundamental key concept of MIL is the above-mentioned: bags are labeled as positive if at least one instance within the bag is positive, and the bag is labeled negative if all instances are negative. This allows MIL to handle situations where the exact localization or identification of cancerous regions within a bag may be challenging or uncertain. The MIL algorithms are therefore designed and implemented to learn a model that can accurately classify bags of instances based on the tumor's presence/absence, considering the ambiguity and flexibility within the bags [10]. In this review, we highlight the benefits of MIL in cancer diagnosis with applicable use cases. We further discuss challenges associated with MIL and propose strategies to overcome them. The review also suggests future research directions for cancer diagnosis using MIL.

II. Fundamental of Multiple Instance Learning

Multiple Instance Learning (MIL) is gaining interest from industry and academia because it fits a variety of problems and allows us to leverage weakly labeled data. MIL finds applications in computer vision, natural language processing, drug activity prediction, bioinformatics, CAD, and document and text classification [8]. It appears that almost every field can use the MIL approach for problem-solving since the bag and instance representation can be reproduced in real-life scenarios. Multiple Instance Learning (MIL) is a form of weakly supervised learning where training instances are arranged in sets called bags, and a label is provided for the entire bag. MIL organizes its data into bags containing multiple instances; this can be replicated in the medical domain, where a bag can be considered as medical report and, an instance, a clinical test result. MIL uses bag-level labels, which can be either positive or negative. A positive bag contains at least one instance of the target class (positive instance), while a negative bag does not contain any instances from the target class (negative instance). Each individual instance is described by a set of covariates (or features). The instance label is ambiguous in practice as it may not be directly observed or clearly identified. This ambiguity is the primary challenge of MIL since there is no evidence of individual precise labels of instances (positive or negative). Therefore, MIL algorithms are responsible for handling this doubt. MIL uses the relationship between bags and instances through a learning process to predict the bag-level labels. The learning process is weakly supervised due to ambiguous instance labels and aims to predict whether a bag is negative or positive. MIL uses various classifiers on bags to predict the bag-level label; the prediction output depends on the aggregation method.

MIL algorithms are usually divided into two main parts: An instance-level classifier and a bag-level classifier. An instance-level classifier provides information about the individual instances within each bag. The instance-level classifiers help to access the features of instances within each bag, and their training results in differentiating between positive and negative instances without directly using bag-level classification. Instance-level classifiers can assist in understanding the characteristics of instances associated with positive or negative bags and provide additional insights into the underlying patterns [9]. Instance-level classifiers can be very effective in certain situations, especially when the data is highly variable or complex. However, they can also be sensitive to noise and outliers, which can affect the accuracy of the classification results. Implementing instance-level classification uses various algorithms such as SVM, logistic regression, or k-nearest neighbors. For instance, the k-nearest neighbor algorithm finds the k-nearest instances to a given example and then assigns the class label that is most common among those instances. Another example is the support vector machine, which finds the hyperplane that separates the different data classes. On the other hand, A bag-level classifier is responsible for making predictions at the bag level based on the collective information of the instances. The output of the bag-level classifier is typically a probability or confidence score indicating the likelihood of a bag being positive or negative. MIL uses majority voting, average probability, max probability, and clustering techniques to achieve bag-level classification. In majority voting, the class of the bag is that of the majority of its instances. For all instances in a bag, the average probability technique computes the average probability of belonging to a class. It sets it as a threshold λ used to classify the bag as of class c if $P(\text{inst} \in c) \geqslant \lambda$. The max probability uses the highest likelihood of an instance belonging to a class for bag-level classification. Clustering methods group instances per cluster according to their similarities. Various machine learning algorithms can be employed as bag-level classifiers, including support vector machines (SVM), decision trees, random forests, and neural networks [9].

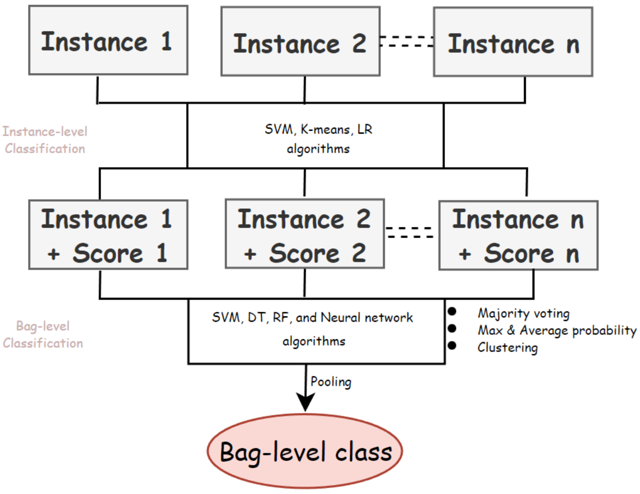


Figure 2. Overview of MIL algorithm

MIL algorithms can be classified into various types depending on their application areas. Regarding cancer diagnosis, some standard MIL algorithms include CNN-based, Deep-MIL, and attention-based MIL, among others, and each has its strengths and weaknesses. However, the advantages of MIL are numerous, as it appears to be a promising solution in the field of cancer diagnosis and prognosis. Due to the combination of approaches and methods, the MIL hybrid algorithms are countless reasons why we will present a few applied to cancer diagnosis tasks.

- *Multiple Instance Learning via Embedded Instance Selection (MILES)*: The MILES algorithm selects representative instances from each bag and transforms the problem into a standard supervised learning task. MILES has been applied in breast cancer histopathology image analysis for cancer detection and grading [11].
- *Deep Multiple Instance Learning (DMIL)*: DMIL combines deep learning architectures with MIL to handle complex, high-dimensional data. DMIL has been used for breast cancer subtype classification, leveraging molecular data to predict breast cancer subtypes [12]. Moreover, Pal et al. and Chang et al. used DMIL for abnormal cell detection in cervical histopathology images [13] and the prediction of chemotherapy response in non-small cell lung cancer using pretreatment CT images [14].
- *Multiple Instance Learning Neural Network (MILNN-MILCNN)*: MILNN utilizes a multiple instance neural network and CNN to classify bags. These algorithms are applicable in a variety of cancer diagnosis tasks. Wei et al. [15] proposed a cluster-based MILNN called the CBMIL model for binary classification (benign or malignant) of lung and breast cancers using WSI. Li et al. also developed a Dual-Stream MILNN for WSI classification; the approach learns from instances and bags while providing critical instances followed by an attention score representing the similarity of the critical instance and other instances [16].
- *Attention-based Multiple Instance Learning*: This approach extends the classical MIL approach using the attention mechanism [17]. Lu et al. [18] proposed a clustering-constrained attention multiple instance learning (CLAM) that uses attention-based learning and MIL to identify subregions of high diagnostics value to classify WSI effectively. CLAM approach is applicable to various cancer types and provides an interpretable weakly supervised deep-learning method for data-efficient WSI processing and learning that only requires slide-level labels.

These MIL algorithms, among others, demonstrate the applicability of MIL in cancer diagnostic tasks, leveraging bag-level labels to effectively identify cancerous regions, subtype classification, and risk prediction. The choice of algorithm depends on the specific cancer diagnostic task and the characteristics of the available data. However, it should be noted that the few types and examples of MIL algorithms presented are a non-exhaustive list because medical imaging and cancer diagnosis are rapidly evolving, and new research papers and methodologies are continuously being published.

III. Applications of Multiple Instance Learning in Cancer Diagnosis

The attention MIL gets can be justified by its ability to fit in various problems and allow it to leverage weakly labeled data. The challenges encountered in getting labeled data are alarming and raise the need for methods dealing with partially labeled data, such as MIL. In cancer diagnosis, pathologists are expected to provide labeled data that can be used for model training. The labeling process is time-consuming and cost-effective; pathologists can successfully achieve partial annotation of the medical images. MIL algorithms use bag and instances representation to handle the classification of partially annotated data. Here, we discuss the applications of MIL in breast cancer diagnosis, lung cancer diagnosis, and prostate cancer diagnosis, highlighting some studies and research that have utilized MIL techniques.

A. Breast Cancer Diagnosis using MIL.

MIL has been applied to breast cancer diagnosis tasks, including subtype classification, mammogram analysis, and lesion detection. In a study by Sudharshan et al., MIL was employed for breast cancer subtype classification using histopathological images [19]. The MIL-based approach captured bag-level representations of breast tissue samples, enabling accurate classification into different molecular subtypes such as luminal A, luminal B, HER2-enriched, and basal-like subtypes. MIL algorithms have also been utilized for mammogram analysis in breast cancer detection. Cheplygina et al. investigated MIL methods for detecting and classifying breast lesions in mammograms [20]. The MIL framework allowed bag-level labels to identify regions of interest and accurately classify mammograms as positive or negative for cancerous lesions.

Furthermore, Wang et al. [21] developed a prediction model for HER-2-positive breast cancer prognosis patients using WSI and MIL. The patient’s histopathological whole slide was considered a bag and was split into thousands of patches (instances) and clustered using the k-means algorithm. The clustered instances were aggregated into bag feature-level representation through graph attention networks to predict the prognostics of the patients. This research showed MIL's potential effectiveness and efficiency in analyzing unlabeled gigapixel WSI to predict Her2-positive breast cancer prognosis. Applications of MIL for breast cancer are diverse and extend beyond ordinary diagnosis. Piumi et al. [22] used MIL on histopathological images for survival prediction in triple-negative breast cancer. A WSI is cropped before the feature extraction process. A pre-trained encoder is used to extract the features before clustering. Each cluster is composed of a group of instances that is fed into a prediction model. The proposed model can output the risk score, a useful biomarker in triple-negative breast cancer prediction.

B. Lung Cancer Diagnosis using MIL.

MIL techniques have been applied to lung cancer diagnosis tasks, focusing on nodule detection and risk prediction.

Dou et al. proposed an MIL-based approach for false positive reduction in lung nodule detection from computed tomography (CT) scans [23]. By considering the entire CT scan as a bag, the MIL model effectively detected and classified lung nodules, aiding radiologists in identifying potentially cancerous regions. Furthermore, MIL-based models have been developed for lung cancer risk prediction by integrating diverse clinical and imaging data. The MIL framework captured bag-level relationships between imaging and clinical variables, providing accurate patient risk estimation. Junhua et al. [24] proposed a lung cancer diagnosis model using deep attention-based MIL. The proposed approach uses the radiomics feature to feed the deep-attention MIL module, which is equipped with an attention mechanism that provides higher interpretability by estimating each instance’s importance in the final diagnosis set. A total of 103 features were extracted per nodule to generate a bag of features that were used for model training. The results of this study demonstrate that the proposed approach will serve as a reliable indicator of the importance of each nodule in the diagnosis process. The module will support medical personnel and patients in interpreting medical results better and handling the disease.

Similarly, Fadre et al. [25] used MIL lung pathophysiological CT scans to detect nodules, fibrosis, and emphysema. Each CT scan was a bag, and the various sections of the scan were considered as instances to adapt to the MIL problem structure. The Radiomic Bag Generator and Hounsfield Units Bag Generator were used for the implementation, and the obtained bags were labeled positive and negative according to the presence of an instance of the target class. A Kernel Density Estimation (KDE) was used to select the most positive instance in the bag, and a sampling method was applied before classification. The model could classify the presence or absence of tumor cells, making it useful for pathologists to handle the disease characterization.

C. Prostate Cancer Diagnosis using MIL.

MIL techniques have been employed in prostate cancer diagnosis, particularly in histopathology image analysis and detection in multiparametric MRI. MIL can potentially use an artificial immune recognition system based on classifying histopathological images into different Gleason scores to facilitate prostate cancer grading. MIL approach allowed the analysis of entire tissue slides as bags, capturing relevant patterns and providing valuable information for cancer diagnosis. MIL-based methods have been developed for prostate cancer detection and localization in multiparametric MRI. Litjens et al. proposed a MIL framework for automatic prostate cancer detection in multiparametric MRI, achieving improved sensitivity and specificity compared to traditional supervised learning methods [26]. The MIL model effectively captured bag-level interactions and relationships among different MRI sequences, enabling precise identification of cancerous regions. Li et al. [27] proposed a multi-resolution MIL model for Gleason grade group classification using MIL. Their model uses end-to-end training with slide-level labels (bags labels) and an aggregated bag-level feature vector clustered and trained to predict cancer grade using slide-level attention distribution. Unlike most classification algorithms focusing on patch segmentation, the proposed MIL-based approach uses multi-resolution instance learning algorithms that can detect suspicious regions accurately for a fine-grained grade. In addition, Golara et al. [28] used ultrasound images from systematic prostate biopsy to develop a deep network for cancer detection using MIL. The MIL enabled learning from ultrasound image regions associated with some statistically labeled data. This approach showed promising results and could be used as a generic extension to other similar cases.

D. Other Cancer Diagnosis using MIL

In practice, providing an exhaustive list of the applications of MIL techniques in diagnosing various cancer types is impossible. In this section, we underline some studies that have employed the MIL approach to achieve diagnosis or related tasks in other cancer types. The baseline remains the same; in the case of partially annotated images, MIL is one of the best and most famous approaches for machine learning-based solutions.

TABLE 1: OTHER CANCER TYPES DIAGNOSED BY MIL

SN	Cancer type	Description	Ref.
1	Bladder cancer	Using hierarchical deep multiple-instance learning with double-stage attention mechanism for high precision of gene mutation prediction.	[29]
2	Lymph node metastasis in various cancers	Two-stage MIL for Lymph Node Metastasis (LNM) Classification. The first stage consists of a double Max-Min MIL to select suspected top-K positive instances. The second stage consists of a transformer-based MIL aggregator combined to achieve classification in LNM.	[30]
3	Ovarian tumors	MRI-Based multiple instance convolutional neural networks for increased accuracy in the differentiation of borderline and malignant epithelial ovarian tumors	[31]
4	Gastritis	Attention MIL and network-based MIL are combined to diagnose chronic gastritis. WSI label is based on the gastritis pathology report, and the MIL is used for accurate diagnosis. The results were compared to the diagnosis results of pathologists and gave total satisfaction.	[32]
5	Colorectal cancer	Detect and grade lesions in colorectal cancer with higher sensitivity using multiple instance learning and feature aggregation methods.	[33]

IV. ADVANTAGES OF MIL IN CANCER DIAGNOSIS

MIL offers several advantages, and the primary one lies in its ability to handle weakly labeled or ambiguous data in cancer diagnostic tasks.

- *Reduced reliance on precise instance-level labels*: Obtaining accurate instance-level labels can be challenging in cancer datasets, especially when dealing with complex imaging modalities or heterogeneous tumor regions. MIL allows for the utilization of bag-level labels, reducing the dependence on precise instance annotations and accommodating the uncertainty inherent in cancer diagnosis [11]. This is particularly beneficial when only bag-level labels, like the case may be when dealing with whole-slide histopathology images or medical records.
- *Accommodation of inter-instance variability*: In cancer datasets, instances within a bag may exhibit variability in terms of size, shape, texture, or intensity. MIL models can effectively handle this variability by considering the collective information of instances within a bag, enabling the identification of common patterns or features associated with the presence or absence of cancer [34].
- *Robustness to noisy or mislabeled instances*: MIL is inherently more robust to noisy or mislabeled instances compared to traditional supervised learning. Even if some instances within a bag are mislabeled or ambiguous, the bag-level label provides a more reliable indication of cancer’s overall presence or absence. This robustness allows MIL models to handle imperfect or incomplete labeling in cancer datasets, improving their generalization performance [35].

V. CHALLENGES ASSOCIATED WITH MIL

Despite its advantages, MIL also poses some challenges in cancer diagnostic tasks:

- *Model interpretability*: MIL models often operate at the bag level, making it more challenging to interpret the specific instances or features responsible for the bag-level predictions. Interpreting the decision-making process becomes complex when the exact localization or identification of cancerous regions within a bag is uncertain. However, various interpretability techniques, such as saliency maps or attention mechanisms, can be employed to gain insights into the contribution of instances or features to the bag-level predictions [36].
- *Bag-level feature representation*: One of the critical aspects of MIL is the representation of bags as feature vectors. Designing effective bag-level feature representations that capture the relevant information from instances is crucial for the success of MIL models. This requires careful selection and extraction of informative features that discriminate between positive and negative bags. Incorporating domain knowledge or leveraging advanced feature extraction techniques, such as deep learning architectures or transfer learning, can improve the representation of bags and enhance the performance of MIL models [34].

To overcome the limitations and improve the performance of MIL-based cancer diagnostic models, several strategies can be employed:

- *Instance selection or weighting*: Selecting informative instances within bags or assigning instance weights can enhance the discrimination between positive and negative bags. Techniques such as instance selection based on confidence scores or importance weighting based on instance-level predictions can improve the model's performance by emphasizing more informative instances [37].
- *Ensembling and boosting*: Ensembling multiple MIL models or applying boosting techniques can enhance the overall performance and robustness of the models. Combining the predictions from multiple models or boosting the performance of weak MIL models through iterative training can improve the accuracy and generalization capabilities of the models [38].

- *Integration with other techniques*: Combining MIL with other machine learning techniques, such as active learning, transfer learning, or domain adaptation, can further improve the performance of cancer diagnostic models. Active learning can be used to select informative instances for labeling, reducing the labeling effort required for training MIL models. Transfer learning and domain adaptation methods can leverage knowledge from related datasets or pre-trained models to enhance the generalization capabilities of MIL models in different cancer diagnostic tasks or datasets [39].

VI. Future Directions and Emerging Trends

- *Integration of MIL with deep learning for enhanced cancer diagnostic accuracy*: Integrating multiple instance learning (MIL) with deep learning techniques holds great potential for advancing cancer diagnostic accuracy. Deep MIL models, which combine the representation learning capabilities of deep neural networks with the bag-level learning framework of MIL, can effectively capture complex patterns and relationships within bags of instances. By leveraging the hierarchical representations learned by deep models, MIL can potentially improve the discrimination between positive and negative bags in cancer diagnosis tasks. Future research efforts should focus on developing deep MIL architectures designed explicitly for cancer diagnostic applications, exploring convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms within MIL frameworks.
- *Exploration of MIL in other cancer types and clinical domains*: While the application of MIL in breast, lung, and prostate cancer diagnosis has shown promising results, there is a need to explore the potential of MIL in other cancer types and clinical domains. Each cancer type presents unique challenges and characteristics that can benefit from the MIL framework. For instance, using dermoscopy images, MIL can be applied to skin cancer diagnosis to classify melanoma, non-melanoma, and benign lesions. MIL can also be employed in gastrointestinal cancer diagnosis to detect and classify polyps or tumors in endoscopic videos or images. Exploring MIL in these and other cancer types will expand its applicability and impact in diverse clinical scenarios.
- *Incorporating domain knowledge and expert annotations into MIL frameworks*: To further improve the performance and interpretability of MIL-based cancer diagnostic models, incorporating domain knowledge and expert annotations into MIL frameworks is crucial. Domain knowledge can guide the selection of informative features, the design of appropriate bag-level representations, and the definition of task-specific constraints. Expert annotations, such as region-of-interest annotations or weak instance-level labels, can provide additional supervision during model training. Combining MIL with domain knowledge and expert annotations can help overcome model interpretability challenges and enhance the model's ability to capture clinically relevant patterns. Future research should focus on developing MIL frameworks that seamlessly integrate domain knowledge and expert annotations to improve cancer diagnostic accuracy.

VII. Conclusions

The potential impact of MIL in improving cancer diagnosis and patient outcomes is substantial. By leveraging bag-level labels and capturing complex patterns within bags, MIL models can aid in early detection, accurate classification, risk prediction, and personalized treatment planning. Developing reliable MIL-based diagnostic tools can enhance clinical decision-making, reduce inter-observer variability, and improve patient outcomes. This review has discussed the challenges associated with MIL, such as model interpretability and bag-level feature representation. Strategies to overcome these limitations have been explored, including instance selection or weighting, ensembling, and integration with other techniques. These strategies can improve the performance of MIL-based cancer diagnostic models and enhance their interpretability and generalization capabilities. Integrating MIL with deep learning is a promising direction to improve cancer diagnostic accuracy. Exploring MIL in other cancer types and clinical domains will expand its applicability and impact in diverse clinical scenarios. Additionally, incorporating domain knowledge and expert annotations into MIL frameworks can further improve the performance and interpretability of cancer diagnostic models. With further advancements and research, MIL-based approaches are promising to improve cancer diagnostic accuracy, ultimately leading to better patient outcomes.

Acknowledgment

This research was supported by a grant from Korea Health Technology R&D Project through the Korea Health Industry Development Institute validation (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant No: HI21C0977).

References

[1] https://www.cancer.gov/types/common-cancers accessed on 16/10/2023

[2] https://www.who.int/news-room/fact-sheets/detail/cancer accessed on 16/10/2023

[3] Bray, Freddie, et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA: a cancer journal for clinicians 68.6 (2018): 394-424.

[4] Iqbal, Muhammad Javed, et al. "Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future." Cancer cell international 21.1 (2021): 1-11.

[5] Gurcan, Metin N., et al. "Histopathological image analysis: A review." IEEE reviews in biomedical engineering 2 (2009): 147-171.

[6] Yanase, Juri, and Evangelos Triantaphyllou. "The seven key challenges for the future of computer-aided diagnosis in medicine." International journal of medical informatics 129 (2019): 413-422.

[7] Anshad, PY Muhammed, and S. S. Kumar. "Recent methods for the detection of tumor using computer aided diagnosis—A review." 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). IEEE, 2014.

[8] Fatima, Samman, Sikandar Ali, and Hee-Cheol Kim. "A Comprehensive Review on Multiple Instance Learning." Electronics 12.20 (2023): 4323.

[9] Dietterich, Thomas G., Richard H. Lathrop, and Tomás Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles." Artificial intelligence 89.1-2 (1997): 31-71.

[10] Maron, Oded, and Tomás Lozano-Pérez. "A framework for multiple-instance learning." Advances in neural information processing systems 10 (1997).

[11] Chen, Yixin, Jinbo Bi, and James Ze Wang. "MILES: Multiple-instance learning via embedded instance selection." IEEE transactions on pattern analysis and machine intelligence 28.12 (2006): 1931-1947.

[12] Das, Kausik, et al. "Detection of breast cancer from whole slide histopathological images using deep multiple instance CNN." IEEE Access 8 (2020): 213502-213511.

[13] Pal, Anabik, et al. "Deep multiple-instance learning for abnormal cell detection in cervical histopathology images." Computers in Biology and Medicine 138 (2021): 104890.

[14] Chang, Runsheng, et al. "Deep multiple instance learning for predicting chemotherapy response in non-small cell lung cancer using pretreatment CT images." Scientific Reports 12.1 (2022): 19829.

[15] Wu, Wei, et al. "Clustering-Based Multi-instance Learning Network for Whole Slide Image Classification." International Workshop on Computational Mathematics Modeling in Cancer Analysis. Cham: Springer Nature Switzerland, 2022.

[16] Li, Bin, Yin Li, and Kevin W. Eliceiri. "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[17] Ilse, Maximilian, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning." International conference on machine learning. PMLR, 2018.

[18] Lu, Ming Y., et al. "Data-efficient and weakly supervised computational pathology on whole-slide images." Nature biomedical engineering 5.6 (2021): 555-570.

[19] Sudharshan, P. J., et al. "Multiple instance learning for histopathological breast cancer image classification." Expert Systems with Applications 117 (2019): 103-111.

[20] Cheplygina, Veronika, Marleen de Bruijne, and Josien PW Pluim. "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis." Medical image analysis 54 (2019): 280-296.

[21] Wang, Y., Zhang, L., Li, Y., Wu, F., Cao, S., & Ye, F. (2023). Predicting the prognosis of HER2-positive breast cancer patients by fusing pathological whole slide images and clinical features using multiple instance learning. Mathematical Biosciences and Engineering, 20(6), 11196-11211.

[22] Sandarenu P, Millar EKA, Song Y, Browne L, Beretov J, Lynch J, Graham PH, Jonnagaddala J, Hawkins N, Huang J, Meijering E. Survival prediction in triple negative breast cancer using multiple instance learning of histopathological images. Sci Rep. 2022 Aug 25;12(1):14527. doi: 10.1038/s41598-022-18647-1. PMID: 36008541; PMCID: PMC9411153.

[23] Dou, Qi, et al. "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection." IEEE Transactions on Biomedical Engineering 64.7 (2016): 1558-1567.

[24] Chen J, Zeng H, Zhang C, Shi Z, Dekker A, Wee L, Bermejo I. Lung cancer diagnosis using deep attention-based multiple instance learning and radiomics. Med Phys. 2022 May;49(5):3134-3143. doi: 10.1002/mp.15539. Epub 2022 Mar 3. PMID: 35187667; PMCID: PMC9310706.

[25] Frade J, Pereira T, Morgado J, Silva F, Freitas C, Mendes J, Negrão E, de Lima BF, Silva MCD, Madureira AJ, Ramos I, Costa JL, Hespanhol V, Cunha A, Oliveira HP. Multiple instance learning for lung pathophysiological findings detection using CT scans. Med Biol Eng Comput. 2022 Jun;60(6):1569-1584. doi: 10.1007/s11517-022-02526-y. Epub 2022 Apr 6. PMID: 35386027.

[26] Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., & Huisman, H. (2014). Computer-aided detection of prostate cancer in MRI. IEEE transactions on medical imaging, 33(5), 1083-1092.

[27] Li, J., Li, W., Sisk, A., Ye, H., Wallace, W. D., Speier, W., & Arnold, C. W. (2021). A multi-resolution model for histopathology image classification and localization with multiple instance learning. Computers in biology and medicine, 131, 104253.

[28] Javadi, G., Samadi, S., Bayat, S., Pesteie, M., Jafari, M. H., Sojoudi, S., ... & Abolmaesumi, P. (2020). Multiple instance learning combined with label invariant synthetic data for guiding systematic prostate biopsy: a feasibility study. International Journal of Computer Assisted Radiology and Surgery, 15, 1023-1031.

[29] Yan, R., Shen, Y., Zhang, X., Xu, P., Wang, J., Li, J., ... & Zhou, S. K. (2023). Histopathological bladder cancer gene mutation prediction with hierarchical deep multiple-instance learning. Medical Image Analysis, 87, 102824.

[30] Y. Chen et al., "dMIL-Transformer: Multiple Instance Learning Via Integrating Morphological and Spatial Information for Lymph Node Metastasis Classification," in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 9, pp. 4433-4443, Sept. 2023, doi: 10.1109/JBHI.2023.3285275.

[31] Jian, J., Li, Y. A., Xia, W., He, Z., Zhang, R., Li, H., ... & Qiang, J. (2022). MRI‐Based Multiple Instance Convolutional Neural Network for Increased Accuracy in the Differentiation of Borderline and Malignant Epithelial Ovarian Tumors. Journal of Magnetic Resonance Imaging, 56(1), 173-181.

[32] Huang, D., et al. "A novel attention fusion network-based multiple instance learning framework to automate diagnosis of chronic gastritis with multiple indicators." Zhonghua Bing li xue za zhi= Chinese Journal of Pathology 50.10 (2021): 1116-1121.

[33] Neto, P. C., Oliveira, S. P., Montezuma, D., Fraga, J., Monteiro, A., Ribeiro, L., ... & Cardoso, J. S. (2022). iMIL4PATH: A semi-supervised interpretable approach for colorectal whole-slide images. Cancers, 14(10), 2489.

[34] Carbonneau, Marc-André, et al. "Multiple instance learning: A survey of problem characteristics and applications." Pattern Recognition 77 (2018): 329-353.

[35] Natarajan, Nagarajan, et al. "Learning with noisy labels." Advances in neural information processing systems 26 (2013).

[36] Tolomei, Gabriele, et al. "Interpretable predictions of tree-based ensembles via actionable feature tweaking." Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017.

[37] Carbonneau, Marc-André, et al. "Multiple instance learning: A survey of problem characteristics and applications." Pattern Recognition 77 (2018): 329-353.

[38] Zhou, Zhi-Hua, Yu-Yin Sun, and Yu-Feng Li. "Multi-instance learning by treating instances as non-iid samples." Proceedings of the 26th annual international conference on machine learning. 2009.

[39] Melendez, Jaime, et al. "On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis." Ieee transactions on medical imaging 35.4 (2015): 1013-1024.

Tagne Poupi Theodore Armand was born in Cameroon, received Msc in information System and networking at ICT University USA, Cameroon Campus in 2021. Currently, he is a Ph.D. research scholar at the Institute of Digital Anti-aging and healthcare at Inje University. His research interest field includes image processing with a focus on medical image analysis, Deep Learning, Machine Learning and Metaverse.

Subrata Bhattacharjee is pursuing his M.S. leading Ph.D. from the Graduate School of Computer Engineering, Inje University, Korea. He is a Research Scholar and Teaching Assistant at the Medical Image Technology Laboratory (MITL) at Inje University. His research interests include multimodal medical imaging, tissue and cell image analysis, digital pathology, image reconstruction, cell and gland segmentation, machine learning, and deep learning. He has some SCIE paper publications and has participated in International and Domestic conferences.

Hee-Cheol Kim received his BSc at the Department of Mathematics, MSc at the Department of Computer Science at SoGang University in Korea, and Ph.D. in Numerical Analysis and Computing Science at Stockholm Stockholm University in Sweden in 2001. He is a professor at the Department of Computer Engineering and Head of the Institute of Digital Anti-aging Healthcare Inje University in Korea. His research interests include machine learning, deep learning, Computer vision, and medical informatics.

A Study on Real-time Evaluation of Uncertainty of PM-10 Concentration Determined by Tele-measuring Instrument

Jeeho KIM [*] [**] [***], Jin-Chun WOO [**], Young SUNWOO [*]

[*] Department of Environmental Engineering, Konkuk University, Korea
[**] KOSTEC(Korea Standard Technology), Inc, Korea
[***] KNJ Engineering, Inc, Korea
jeeho.kim@knj-eng.co.kr, Fax: +82 31 459 7321, Tel: +82 31 451 7082

Abstract— **We developed a real-time particulate matter(PM-10) and its uncertainty monitoring instrument assisted by beta-ray absorption method and continuous tele-measuring system. According to the GUM as an authentic reference, 4 steps of operational procedures for uncertainty evaluation were included to general type of instrument; at first, establishment of a model equation of PM concentration, at second, calculation of each standard uncertainty, at third, calculation of combined standard uncertainty, at fourth, calculation of expanded uncertainty at 95% confidence level.**

The developed instrument was tested at Gwanpyeong-dong, Da ejeon during Nov. 2023. Through this field application, expanded uncertainties at 720 data points of PM concentration were obtained. At the low level of PM concentration, 22.2 μg/m^3, the expanded uncertainty value was 2.0 μg/m^3 at the 95 % confidence level, while at the high level of PM concentration, 118.6 μg/m^3, it was 9.8 μg/m^3. The expanded uncertainty values versus PM concentration were well fitted with a second-order regression equation, $y = 0.00007\ x^2 + 0.073\ x + 0.36$ (μg/m^3). In comparison with each uncertainty variance, it was revealed that the most important uncertainty source is the uncertainty of linearity quantified by equivalence evaluation between the PM concentration values obtained by beta-ray absorption method and direct weighing method. And the contribution rate of this uncertainty was 32% of total uncertainty.

It has revealed that the developed instrument has successfully managed to calculate and display real-time measurement uncertainty with various uncertainty source budgeting data.

Keywords—**Real-time uncertainty evaluation. Measurement uncertainty, Particulate matter, PM-10, Air quality monitoring instrument, Tele-measuring system**

I. Introduction

In Korea, the beta-ray absorption method is adopted as the fundamental measurement technique[1] for monitoring ambient particulate matter(PM-10). National Ambient air quality Monitoring Information System(NAMIS), as a continuous ambient air quality monitoring system, operates to measure air pollutants including PM-10 and collects measured air quality data across the country. NAMIS also collects data in real-time utilizing the internet network. The collected data are used as foundational information about air quality and contribute to the national and local air quality improvement policies[2].

Maintaining measurement accuracy requires regular calibration and quality control processes for the monitoring instrument. For this processes, essential tasks include periodic zero/span calibration for PM-10 concentration and the calibration of all sensors measuring sample temperature, pressure and flow rate. During the continuous measurement operations of ambient particulate of PM-10, additional uncertainty factors may arise in comparison to typical single measurement operation[3]. Since the long-term operation leads to changes of reproducibility and linearity of calibration, their quality control process is one of the crucial works for this continuous measurement of PM-10 concentration.

Recently many laboratories adopt “Guide to the Expression of Uncertainty in Measurement(GUM)”[4] for the quality assurance of measurement result[4-13]. Therefore, it is necessary to follow the procedures and show measurement uncertainty in continuous monitoring instrument. In this study, we developed an instrument for simultaneous real-time monitoring of ambient particulate matter(PM) concentration and uncertainty. This instrument was assisted by the beta-ray absorption method and a tele-measuring system. Applying this instrument, PM concentration and its uncertainty values were collected in Gwanpyeong-dong, Daejeon during Nov. 2023. With the measurement results and corresponding uncertainty values, we discussed the measurement quality and the feasibility of the developed instrument.

II. Experiments

A. *Instrument and modification of software*

In this study, the continuous ambient particulate monitoring instrument, K-501A(KOSTEC Inc.) was used and operational software was upgraded to enable onsite evaluation of measurement uncertainty. The basic specifications of K-501A are outlined in Table 1[14].

For the modification of conventional type of instrument, 4 steps of operational procedures of uncertainty included; at first, establishment of a model equation of PM concentration, at second, calculation of each standard uncertainty, at third, calculation of combined standard uncertainty, and at fourth, expanded uncertainty of 95% confidence level[3-12, 15]. Uncertainty factors treated in the processes included repeatability of beta-ray intensity ratio, uncertainty of temperature, pressure and flow measurements, uncertainty of attenuation coefficient, uncertainty of sampling area, linearity and background uncertainty, and drift of standard film.

TABLE 1. K-501A SPECIFICATIONS

Items	Specifications
Measurement method	β-ray attenuation
Measurement range	0-1,000 μg/m³
Minimum detection limit	<4 μg/m³ (2σ)
Measurement cycle	1h
Radiation Source	14C, 60 μCi(<.22×106 Bg) source
Filter tape	Glass Fiber Roll Filter
Air flow rate	16.7 L/min
Flow system	Mass flow controller
Operating temperature	0~50 ℃

B. Approach for the Evaluation of Uncertainty

Generally, basic approach for calculating measurement uncertainty is based on GUM[4-9]. In this study, the operational software in K-501A was designed and included with procedure of GUM. So, the final frame of calculating and displaying measurement uncertainty is such as following Figure 1.

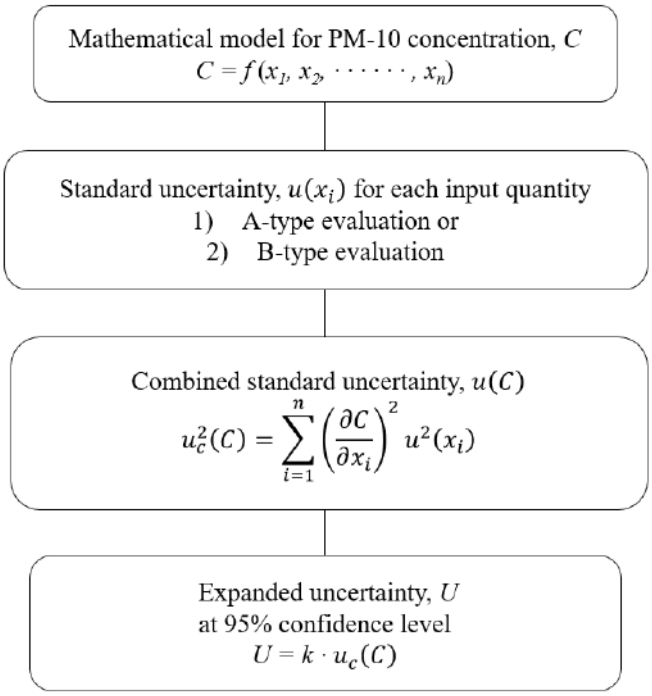

Mathematical model for PM-10 concentration, C
$C = f(x_1, x_2, \cdots\cdots, x_n)$
Standard uncertainty, $u(x_i)$ for each input quantity
1) A-type evaluation or
2) B-type evaluation
Combined standard uncertainty, $u(C)$
$u_c^2(C) = \sum_{i=1}^{n} \left(\frac{\partial C}{\partial x_i}\right)^2 u^2(x_i)$
Expanded uncertainty, U
at 95% confidence level
$U = k \cdot u_c(C)$

Figure 1. The procedure for accessing real-time measurement uncertainty.

C. Establishing Mathematical Model of Measurand

For the determination of PM-10 concentration, beta-ray absorption method was incorporated with Beer-Lambert relationship[16, 17]. The equation of measurand with beta ray attenuation before and after sampling particulate from ambient air is such as following Eq. 1.

$$C = \frac{10^6 \cdot A \cdot ln\left(\frac{I_0}{I}\right)}{\mu \cdot Q \cdot \Delta_t \cdot \left(\frac{P}{P_{std}}\right) \cdot \left(\frac{T_{std}}{T}\right)} \qquad \text{Eq. 1.}$$

where, C, measured PM-10 concentration(μg/m^3), A, area of the sample filter(cm^2), I_0/I, beta-ray intensity ratio before and after sampling, μ, attenuation coefficient(cm^2/mg), Q, sampling flow rate(L/min), $\triangle t$, sampling time(min), P_{std}, pressure at standard conditions(mmHg), P, sampling pressure(mmHg), T_{std}, temperature at standard conditions(K), T, sampling temperature(K).

In the continuous operations of ambient particulate of PM-10, additional uncertainty factors may arise. These type of uncertainty factors include the span drift quantified by calibration using standard film and equivalence evaluation[18] results between PM-10 concentration obtained by K-501A analyzer and by direct weighing method. Final mathematical model of measurand, which is suitable to evaluate uncertainty, should be modified as following Eq. 2.

$$C = \frac{10^6 \cdot A \cdot ln\left(\frac{I_0}{I}\right)}{\mu \cdot Q \cdot \Delta_t \cdot \left(\frac{P}{P_{std}}\right) \cdot \left(\frac{T_{std}}{T}\right)} \cdot f_{drift} \cdot f_{lin} + \delta_{bg} \qquad \text{Eq. 2,}$$

where f_{drift} is span drift correction factor using standard film, f_{lin} is linearity correction factor of slope and δ_{bg} is intercept correction factor.

D. Standard uncertainty and degrees of freedom

For the input quantities in Eq. 2, the standard uncertainty(SU), and degrees of freedom(DF) should be determined[4] according to the GUM guidelines. Generally, The standard uncertainty can be calculated through Type A or Type B evaluation[4, 12]. In this experiment, all the standard uncertainty values are evaluated by Type B evaluation. For the Type B evaluation, past measurement data, prior quality control experimental results and certificates of various references are essentially needed.

1) SU and DF of sampling area

After determining the diameter of sampling hole using digital micrometer, the area of sampling hole can be quantified by the following Eq. 3.

$$A = \pi \left(\frac{D}{2}\right)^2 \qquad \text{Eq. 3.}$$

Assuming that the probability density distribution is rectangular in the determination of diameter, then the standard uncertainty of diameter is calculated as shown in Eq. 4[4, 10, 11].

$$u(D) = \frac{a}{\sqrt{3}} \qquad \text{Eq. 4.}$$

where, a, resolution of digital micrometer. By the uncertainty propagation method, the standard uncertainty of area is as following Eq. 5.

$$u(A) = \frac{\pi D}{2} \cdot u(D) \qquad \text{Eq. 5.}$$

The DF in the Type B evaluation of uncertainty can be obtained as shown in Eq. 6[4].

$$\nu(x_i) = \frac{1}{2}\left(\frac{100}{R}\right)^2 \qquad \text{Eq. 6,}$$

where, R = 100 - uncertainty of uncertainty evaluation with the unit of percentage.

2) SU and DF of Attenuation Intensity Ratio

In the measurement process of PM-10, the standard deviation with 3 consecutive measurement results were continuously calculated. And a pooled standard deviation was calculated using smaller values than 30 % from bottom of most recent one-week data. The pooled standard deviation as shown in Eq. 7 was used as the standard uncertainty of attenuation intensity ratio.

$$s_p = \sqrt{\frac{\sum_{i=1}^{n} s_i^2 \cdot \nu_i}{\sum_{i=1}^{n} \nu_i}} \qquad \text{Eq. 7.}$$

And degrees of freedom(ν) of pooled standard uncertainty of attenuation intensity ratio is as shown in Eq. 8.

$$\nu = \sum_{i=1}^{n} \nu_i \qquad \text{Eq. 8.}$$

3) SU and DF of Attenuation Coefficient

To determine the standard uncertainty of μ, it is necessary to measure the weight(W_{film}) and area(A_{film}) of a standard film, and the beta-ray intensity ratio(I_0/I) before and after determining the standard film. With determined data, attenuation coefficient can be derived as following Eq. 9[16].

$$\mu = \frac{ln\left(\frac{I_0}{I}\right)}{\left(\frac{W_{film}}{A_{film}}\right)} \qquad \text{Eq. 9.}$$

To quantify the standard uncertainty of μ, it is prerequisite to evaluate the standard uncertainty of each input quantity. For the uncertainty values of area and beta-ray intensity ratio, same procedures were adopted as described previous sections. And the standard uncertainty of weight of standard film can be calculated using the following Eq. 10 with quality control limit(a) data.

$$u(\,W_{film}) = \frac{a}{\sqrt{3}} \qquad \text{Eq. 10.}$$

Finally, the standard uncertainty of μ was combined using uncertainty propagation rule as shown in Eq. 11.

$$\left(\frac{u(\mu)}{\mu}\right)^2 = \left(\frac{u(I_0/I)}{(I_0/I) \cdot ln(I_0/I)}\right)^2 + \left(\frac{u(\,W_{film})}{W_{film}}\right)^2 + \left(\frac{u(\,A_{film}\,)}{A_{film}}\right)^2 \qquad \text{Eq. 11.}$$

4) SU and DF of Sampling Time

In the case of sampling time(Δt), the standard uncertainty is estimated by Eq. 4 with resolution(a) data of the timer used in the software of K-501A. Since the determination of lapsed time involves two times of determination, the standard uncertainty of Δt can be calculated as shown in Eq. 12

$$u(\Delta t) = \frac{\sqrt{2}\,a}{\sqrt{3}} = \mathrm{a}\sqrt{\frac{2}{3}} \qquad \text{Eq. 12.}$$

The degrees of freedom can be quantified as the same way as shown in Eq. 6.

5) SU and DF of Temperature, Pressure and Flow Rate

For the evaluation of standard uncertainty of determined values from many sensors in the PM monitor, calibration data with certificates of references are needed. Therefore, the standard uncertainty and degrees of freedom of sample flow rate(Q), temperature(T) and pressure(P) can be extracted from prior calibration data with the certificates provided.

6) SU and DF of Linearity and Background

For the linearity and background control, it is essential to compare measurement results periodically between beta-ray attenuation and direct weighing methods. This equivalence evaluation should be conducted once in every 2 years in Korea as a regulation[2]. To quantify these two standard uncertainty values, most recent equivalence evaluation results were assisted.

We conducted this equivalence evaluation with two different measurement systems for the comparisons of PM-10 concentration. In this study, one measurement system was composed with K-501A in which an air-sampling adaptor was included. And, another system as a reference was composed with the E-SEQ-FRM(Metone Inc.) manual sampler and Microbalance(Satorius Inc.). This evaluation was conducted continuously for more than 2 months on experimental field site[2]. The 24-hour average data measured by two methods were directly compared[18, 19].

In order to quantify standard uncertainty values, at first, a linear regression line($y= b_1x+b_0$) is calculated with obtained comparison data. And the uncertainty variance matrix(U) of the slope(b_1) and intercept(b_0) of the regression line was derived as shown in Eq. 13[4].

$$U = \begin{pmatrix} u^2(b_0) & u(b_0,b_1) \\ u(b_0,b_1) & u^2(b_1) \end{pmatrix} \qquad \text{Eq. 13,}$$

$$= s^2 \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}^{-1} .$$

where, $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - b_0 + b_1 x_i)^2}{n-2}}$.

Since the standard uncertainty of linearity factor(f_{lin}), includes both the bias of slope and uncertainty of slope, the standard uncertainty of linearity factor can be derived as shown Eq. 14.

$$u^2(f_{lin}) = \left(u(b_1)\right)^2 + \left(\frac{b_1 - 1}{\sqrt{3}}\right)^2 \qquad \text{Eq. 14.}$$

With same idea on the treatment of linearity, the standard uncertainty of δ_{bg} can be derived as shown Eq. 15.

$$u^2(\delta_{bg}) = \left(u(b_0)\right)^2 + \left(\frac{b_0}{\sqrt{3}}\right)^2 \qquad \text{Eq. 15.}$$

7) SU and DF of Drift

For the instrumental drift control, periodic calibration and quality control with standard film are essential. So, the standard uncertainty of drift can be calculated using the $u(\,m_{film})$ and quality control limit(a) data.

E. Combined Standard Uncertainty and DF of Measurand

Basically, the combined standard uncertainty, $u_c(C)$ of PM-10 concentration as a measuand can be calculated by the uncertainty propagation rule shown in Figure 1[4-9]. In this study, at first, equation of combined standard uncertainty can be derived using uncertainty propagation rule and partial derivatives of input quantities. Next, both side of the derived equation can be divided by PM-10 concentration(C) and arranged to appropriate form of equation as shown in following Eq. 16.

$$\left(\frac{u(C)}{C}\right)^2 = \left(\frac{u(A)}{A}\right)^2 + \left(\frac{u(Q)}{Q}\right)^2 + \left(\frac{u(\Delta_t)}{\Delta_t}\right)^2 + \left(\frac{u(\mu)}{\mu}\right)^2 \qquad \text{Eq. 16.}$$
$$+\left(\frac{u(I_0/I)}{(I_0/I)\cdot ln(I_0/I)}\right)^2 + \left(\frac{u(T)}{T}\right)^2 + \left(\frac{u(P)}{P}\right)^2 \qquad .$$
$$+\left(\frac{u(f_{drift})}{f_{drift}}\right)^2 + \left(\frac{u(f_{lin})}{f_{lin}}\right)^2 + \left(\frac{u(\delta_{bg})}{C}\right)^2 \qquad .$$

The effective degrees of freedom(ν_{eff}), which are the degrees of freedom of combined standard uncertainty, can be obtained with the standard uncertainty and DF of each input quantity, and the partial derivatives, as shown in Equation 17.

$$\nu_{eff} = \frac{u_c^4(C)}{\left(\frac{u_i^4(A)}{\infty} + \frac{u_i^4(Q)}{\infty} + \frac{u_i^4(\Delta_t)}{\infty} + \frac{u_i^4(\mu)}{\infty} + \frac{u_i^4(I_0/I)}{\nu_{I_0/I}} + \frac{u_i^4(T)}{\infty} + \frac{u_i^4(P)}{\infty} + \frac{u_i^4(f_{drift})}{\infty} + \frac{u_i^4(f_{lin})}{\infty} + \frac{u_i^4(\delta_{bg})}{\infty}\right)} \qquad \text{Eq. 17,}$$

where, $u_i(x_i) = \frac{\partial f}{\partial x_i} \cdot u(x_i)$,
if the model equation is $y = f(x_1,\ x_2, \cdots\cdots\ \ x_n)$.

F. Expanded Uncertainty as Measurand

Basically, the expanded uncertainty(U) is statistical interval of expected measurand value and calculated by multiplying the combined standard uncertainty to coverage factor(k). Since the coverage factor is determined depending on the probability density distribution of measured quantity, t-value from t-distribution with 95 % confidence level can be used in this study. The expanded uncertainty was calculated as shown in Eq. 18[4].

$$U = k \cdot u_c(C) \qquad \text{Eq. 18.}$$

As a result of accessing uncertainty, the PM-10 concentration in ambient air can be expressed in a simple form, as shown in Equation 19.

$$C\ \pm U \qquad \text{Eq. 19.}$$

This indicates that the estimated value of PM-10 concentration is located from C - U to C + U at 95 % confidence level.

G. Uncertainty Budget

We calculated how much each uncertainty factor contributes to the uncertainty of the final measurement value. For the comparison with the same unit of variance, contribution rate was derived as shown Eq. 18.

$$\left(\frac{\frac{\partial f}{\partial x_i} \cdot u_i(x_i)}{u_c(C)}\right)^2 \times 100(\%) \qquad \text{Eq. 20}$$

With the results, we compared the contributions of uncertainty components and verified important sources of uncertainty [4].

H. Installation of the Software of Uncertainty

In this study, particulate monitoring instrument, K-501A was used and operational software was modified by the developed procedure of uncertainty evaluation and expression for PM-10 concentration. The main control menu of K-501A was originally composed of 4 category, such as Setting-up measurement condition, Diagnostics, Calibration and Quality Control. To display real-time calculated uncertainty, we modified and added software of uncertainty as shown in the Figure 2.

With modified operational software, one can periodically exchange all the newly prepared standard uncertainty values and obtain real-time PM-10 concentration and its uncertainty.

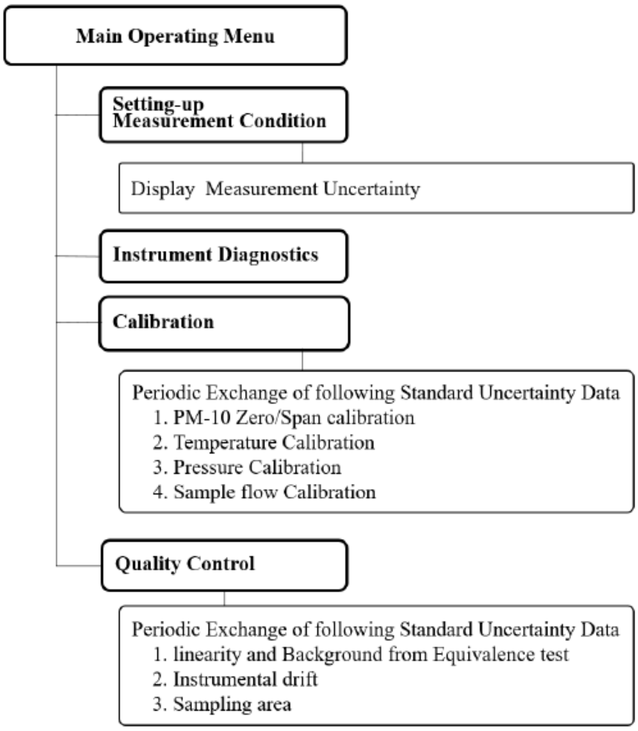

Main Operating Menu
Setting-up Measurement Condition
Display Measurement Uncertainty
Instrument Diagnostics
Calibration
Periodic Exchange of following Standard Uncertainty Data
1. PM-10 Zero/Span calibration
2. Temperature Calibration
3. Pressure Calibration
4. Sample flow Calibration
Quality Control
Periodic Exchange of following Standard Uncertainty Data
1. linearity and Background from Equivalence test
2. Instrumental drift
3. Sampling area

Figure 2. The operating menu for accessing uncertainty.

I. Field Application of Developed Software of Uncertainty

With the modified K-501A, the PM concentration and its uncertainty values were collected for one month at Gwanpyeong-dong, Daejeon during Nov. 2023. Through this field application, PM concentration and their expanded uncertainties of 95% confidence level were obtained and their various characteristics were evaluated.

III. Results

We have obtained many PM-10 concentration and uncertainty values, which are representative results produced by the procedures explained at the previous sections. Here, as an example, we present detailed results only on a PM-10 concentration value, 22.15 μg/m^3 obtained once at the field application of measurement and onsite evaluation of uncertainty. For the selected value, all input quantities and associated uncertainty values obtained by the procedures explained at previous sections are such as following Table 1.

TABLE 2. The standard uncertainties and degrees of freedom for each input quantity and measurand

Input/Output Quantity	Input/Output Quantity Value	Standard uncertainty, $u(x_i)$	Unit	DF
A	1.130	0.000022	cm^2	∞
I_0/I	1.005	0.000091	-	13
μ	0.310	0.0054	cm^2/mg	∞
$\triangle t$	50.00	0.000053	min	∞
Q	16.67	0.036	L/min	∞
P	764	0.45	mmHg	∞
T	5.8	0.10	K	∞
f_{drift}	1	0.017	-	∞
f_{lin}	1	0.026	-	∞
δ_{bg}	0	0.30	-	∞
C	22.15	1.00	μg/m^3	∞

The standard uncertainty values of linearity and background shown in Table 2 were evaluated with the results of the most recently performed equivalence evaluation using the K-501A. The scatter plot for the paired comparison data obtained by the equivalence evaluation is shown in Figure 3.

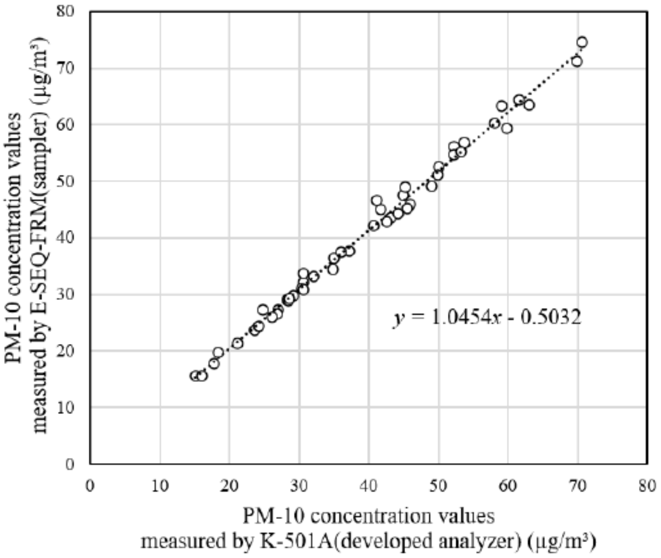
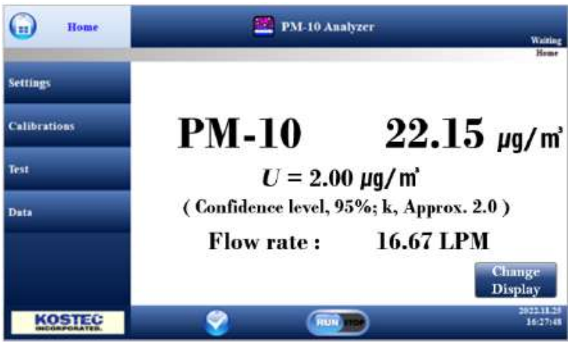


Figure 3. The scatter plot for the equivalence evaluation results.

The slope of regression line was 1.045 and the intercept of that was -0.503 μg/m^3. For the standard uncertainty of linearity and background, both the bias and regression uncertainty were combined by Eq. 14 and Eq. 15.

The combined standard uncertainty was obtained using Eq. 16 and the value was 1.00 μg/m^3. And the effective degrees of freedom were obtained using Eq. 17 and the value was 2×10^{17}(approximately ∞). Since the probability density function was *t*-distribution with approximately ∞ of degrees of freedom, the probability density function of measurand could be assumed as a Gaussian distribution.

The expanded uncertainty as an expected statistical interval of measurand was calculated by Eq. 18. And its value was 2.0 μg/m³ at the confidence level of 95 %, with k = 2.0. By the procedure explained at previous sections, each result was quantified by the software in K-501A and was displayed at the front display panel as shown in Figure 4.



PM-10 22.15 μg/㎥

U = 2.00 μg/㎥

(Confidence level, 95%; k, Approx. 2.0)

Flow rate : 16.67 LPM

Figure 4. The display of PM-10 concentration measurement results

We compared the contribution rate of each uncertainty source as described with Eq. 20 and found that highest contributed component is uncertainty of linearity(contribution rate, 32 %) found at consistency test. And the contribution rate of the uncertainty of flow rate, temperature and pressure were not so high(contribution rate, blow 1.0 %). The uncertainty contributions rate for each input quantity was derived and plotted in Figure 5.

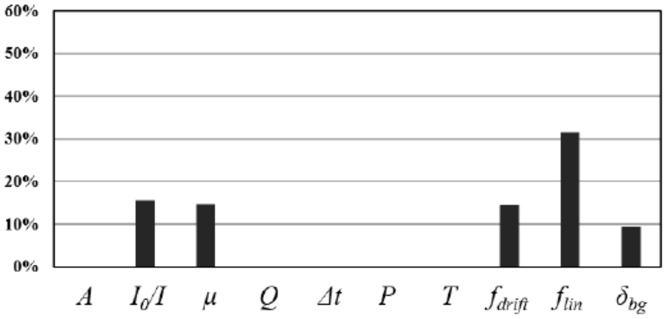


Figure 5. The uncertainty contributions for each input quantity.

During the field application period in November 2023, the measurement uncertainties for hourly average data were collected for PM-10 concentration value, and all the data of uncertainties were depicted versus determined PM-10 concentration as shown in Figure 6. As the PM concentration increased, the expanded uncertainty also increased. For a high concentration level of PM-10, 118.55 μg/m³, the expanded uncertainty was 9.8 μg/m³ at 95 % confidence level. For a low concentration level of PM-10, 22.15 μg/m³, the expanded uncertainty was 2.0 μg/m³ at 95 % confidence level.

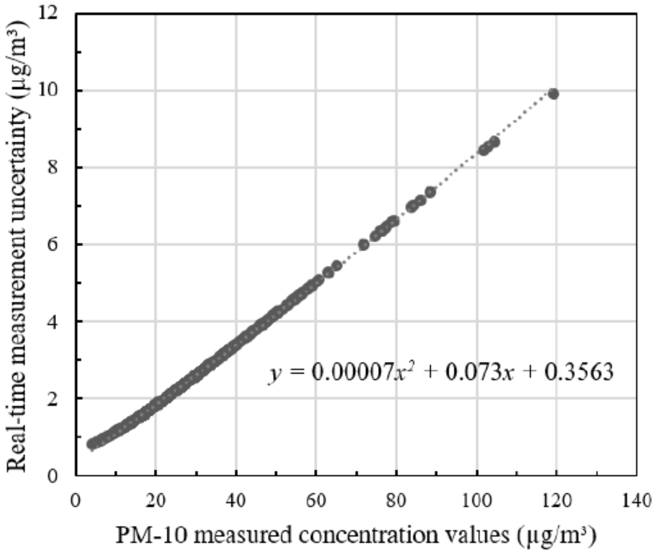


Figure 6. The real-time measurement uncertainty based on PM-10 measured concentration value.

Based on the results, the regression equation of the expanded uncertainty(y) versus PM-10 concentration(x) was formulated as shown in Equation 21. The data was well fitted with a second-order regression line.

$$y = 0.00007x^2 + 0.073x + 0.36 \text{ (μg/m}^3\text{)} \qquad \text{Eq. 21}$$

IV. Conclusion

We developed a real-time particulate matter(PM-10) and its uncertainty monitoring instrument assisted by beta-ray absorption method and tele-measuring system. According to the GUM, operational procedure of uncertainty in the instrument included 4 steps; at first, establishment of a model equation of PM concentration, at second, calculation of each standard uncertainty, at third, calculation of combined standard uncertainty, and at fourth, expanded uncertainty of 95% confidence level. Uncertainty factors in the process included repeatability of beta-ray intensity ratio, uncertainty of temperature, pressure and flow measurements, uncertainty of attenuation coefficient, uncertainty of sampling area, equivalence evaluation with direct weighting measurement value after air sampling, and sensitivity drift of standard film.

We tested the developed instrument and the PM concentration and its uncertainty values were collected at Gwanpyeong-dong, Daejeon during Nov. 2023. Through this field application, expanded uncertainties at 720 data points of PM concentration were obtained. At the low level of PM concentration, 22.2 μg/m³, the expanded uncertainty value was 2.0 μg/m³ at the 95 % confidence level, while at the high level of PM concentration, 118.6 μg/m³, it was 9.8 μg/m³. The expanded uncertainty values were also well fitted with a second-order regression equation, $y = 0.00007\, x^2 + 0.073\, x + 0.36$ (μg/m³). In comparison with each uncertainty variance, it was revealed that the most important uncertainty source is the uncertainty quantified by equivalence evaluation between the PM concentration values obtained by beta-ray absorption method and direct weighing method. And the contribution rate of this uncertainty was 32% of total uncertainty.

we have studied real-time uncertainty evaluation of PM-10 measurement in ambient air assisted by tele-measuring instrument. As a conclusion, it has revealed that the developed instrument has successfully managed to calculate and display real-time measurement uncertainty with various uncertainty source budgeting data.

Acknowledgment

All authors extend their sincere gratitude to Korea Standard Technology(KOSTEC), Republic of Korea, for their generous support throughout this research. We appreciate CEO Lee Jae-yong's unwavering support, which has been instrumental in bringing this research to its conclusion. Lastly, we would like to express our thanks to all the researchers for their encouragement and support.

References

[1] Airkorea, www.airkorea.or.kr

[2] Guidelines for the Installation and Operation of Ambient Air Quality Monitoring Networks, Ministry of Environment(MOE) and National Institute of Environmental Research(NIER), 2022.

[3] G. Buonanno, M. Dell 'Isola, L. Stabile, A. Viola, "Critical aspects of the uncertainty budget in the gravimetric PM measurements", *Measurement*, Volume 44, 139-147, 2011

[4] Uncertainty of measurement – Part 3: Guid to the expression of uncertainty in measurement(GUM:1995), ISO/IEC Guide 98-3:2008

[5] Measurement Uncertainty Analysis Principles and Methods, NASA Measurement Quality Assurance Handbook − ANNEX 2, NASA-HDBK-8739.19-3, 2010

[6] International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM 3rd edition),"ISO/IEC Guide 99:2007

[7] Expression of the Uncertainty of Measurement in Calibration, EA-4/, http://www.european-ccreditation.org/publication

[8] Uncertainty of measurement - Part 1: Introduction to the expression of uncertainty in measurement, ISO/IEC Guide 98-1:2009

[9] Extension to any number of output quantities, ISO/IEC Guide 98-3:2008/Suppl. 2:2011

[10] C.F Dietrich, Uncertainty, calibration and probability, second edition, Adam-Hilger, Bristol, 1991

[11] Expression of the Uncertainty of Measurement in Calibration, EA-4/, http://www.european-ccreditation.org/publication

[12] Propagation of distributions using a Monte Carlo method, ISO/IEC Guide 98-3:2008/Suppl. 1:2008

[13] J. G. Watson, R. J. Tropp, S. D. Kohl, X. Wang, J. C. Chow, "Filter processing and gravimetric analysis for suspended particulate matter samples", *Aerosol Sci. Eng.*, 1, 93–105, 2017

[14] Korea Standard Technology(KOSTEC), INC *Beta Ray PM Monitor K-501A*, http://www.ikostec.co.kr/product/product_view.htm?product_category=&idx=2142

[15] J. Pokhariyal, A. Mandal, S. G. Aggarwal, "Uncertainty Estimation in PM10 Mass Measurements", *MAPAN* 34, 129–133, 2019

[16] Air Pollution Source Testing Standards - Automatic Measurement Method for Fine Particulate Matter(PM-10) in Ambient Air Using Beta Attenuation, National Institute of Environmental Research(NIER), 2021.

[17] J. C. Chow, "Measurement methods to determine compliance with ambient air quality standards for suspended particles". *J. Air Waste Manage. Assoc.*, 45, 320–382, 1995

[18] H. Hauck, A. Berner, B. Gomiscek, S. Stopper, H. Puxbaum, M. Kundi and O. Preining, "On the equivalence of gravimetric PM data with TEOM and beta-attenuation measurements", *J. Aerosol Sci.*, 35, 1135–1149, 2004

[19] Notice on the Type Approval and Performance Testing of Environmental Measurement Instruments, National Institute of Environmental Research(NIER), 2017.

Jeeho Kim, his B.S degree in Mechanical Engineering from Konkuk university, in 2019 and M.S. degree in Environmental Engineering from Konkuk university in 2021.
He joined the Korea Standard Technology(KOSTEC) where he is a researcher in development of air quality monitor and study of measurement uncertainty. He works as an engineer at KNJ Engineering, INC.

Jin-Chun Woo, his B.S degree in Chemistry from Sogang university, Seoul, in 1978 and M.S. degree in Inorganic Chemistry from Sogang university, Seoul, in 1980. And his Ph.D. degree in Material Science and Engineering from Nagoya University, Nagoya, in 1996.
He joined the Korea Research Institute of Standard and Science(KRISS) in 1982 as a researcher of measurement standards on the analytical chemistry. He joined the Korea Standard Technology(KOSTEC) in 2020 as a technical director. His research interests include measurement uncertainty, quality assurance of measurement results, instrumentation of gas analysis, standard procedure of gas analysis.

Young Sunwoo, his B.S degree in Chemical Engineering from Yonsei university, Seoul, in 1984 and M.S. degree in Material Engineering from University of Iowa, Iowa, in 1987. And his Ph.D. degree in Environmental Engineering from University of Iowa, Iowa, in 1993.
He completed his post-doctoral researcher at Princeton University, New Jersey and is a professor of environmental engineering at Konkuk University, Seoul, since 1994. He is the secretary general of the Internationl Union of Air Pollution Prevention and Environmental Protection Associations(IUAPPA), a member of the editorial board of the international journal "Atmosphere" and the specialized committee of the United National Environment Programme(UNEP) Asia-Pacific Clean Air.

Table of Contents
(Journal)

Volume 12 Issue 1 January 2023

Page: 1475 - 1482

Title: An Adaptive User Scheduling Algorithm for 6G Massive MIMO Systems

Author : Prof. Robert Akl, Prof. Robin Chataut

Institute : Fitchburg State University

Country : USA

Volume 12 Issue 2 March 2023

Page: 1483 - 1493

Title: A Blockchain based Security Assessment Framework

Author : Mr. NANDURI SATYANARAYANA

Institute : CDAC

Country : India

Volume 12 Issue 3 May 2023

Page: 1494 - 1506

Title: A Horizontal Federated Learning Approach to IoT Malware Traffic Detection: An Empirical Evaluation with N-BaIoT Dataset

Author : Mr. Phuc Hao Do, Dr. Tran Duc Le, Prof. Vladimir Vishnevsky, Prof. Aleksandr Berezkin, Prof. Ruslan Kirichek

Institute : sut.ru

Country : Viet Nam

Volume 12 Issue 4 July 2023

Page: 1507 - 1513

Title: A Deep learning Framework for Cultural Heritage Damage Detection for Preservation; Based on the case of Heunginjimun and Yeongnamnu in South Korea

Author : Dr. Sang-Yun Lee, Mr. Daekyeom Lee

Institute : ETRI

Country : Korea(South)

Volume 12 Issue 5 September 2023

Page : 1514 - 1520

Title: A Study on Connectivity Evaluation Among Peer Groups in Pure P2P Networks

Author : Dr. Yutaka Naito, Dr. Takumi Uemura, Dr. Takashige Hoshiai

Institute : Sojo University

Country : Japan

Volume 12 Issue 6 November 2023

Page : 1521 - 1527

Title: Quick Blocking Operation of IDS/SDN Cooperative Firewall Systems by Reducing Communication Overhead

Author : Mr. Akihiro Takai, Mr. Yusei Katsura, Prof. Nariyoshi Yamai, Dr. Rei Nakagawa, Dr. Vasaka Visoottiviseth,

Institute : Tokyo University of Agriculture and Technology

Country : Japan

Page : 1528 - 1540

Title: Automated Vulnerability Assessment Approach for Web API that Considers Requests and Responses

Author : Mr. Yuki Ishida, Prof. Masaki Hanada, Prof. Atsushi Waseda, Dr. Moo Wan Kim

Institute : Tokyo University of Information Sciences

Country : Japan

An Adaptive User Scheduling Algorithm for 6G Massive MIMO Systems

Robin Chataut[*] and Robert Akl[**]
[*]School of Computing and Engineering, Quinnipiac University, USA
[**]Department of Computer Science, University of North Texas, USA
robin.chataut@quinnipiac.edu, robert.akl@unt.edu

***Abstract*—Massive MIMO (Multiple-Input Multiple-Output) is a promising wireless access technology that has emerged as a solution to the ever-increasing demand for network capacity. Massive MIMO is expected to play a crucial role in the deployment of 5G and upcoming 6G networks, enabling the realization of their full potential capacity. Despite the numerous benefits, user scheduling during downlink communication in Massive MIMO systems is a challenging task due to the large number of antenna terminals. In this paper, we propose a novel scheduling algorithm aimed at improving the area throughput, sum capacity, error performance, and ensuring fairness among all users. The proposed algorithm uses the average channel rate as the scheduling criteria, which is calculated from the channel state information obtained from the users during uplink transmission. To evaluate the performance of our proposed algorithm, we conducted simulations using Matlab. Our results demonstrate that our proposed channel rate-based scheduling algorithm is superior to conventional scheduling algorithms in terms of sumrate, throughput, and bit error performance while also ensuring fairness among all users. The proposed algorithm can address the challenge of user scheduling in Massive MIMO systems and contribute to the efficient deployment of 5G and 6G networks. The ability to improve system capacity, area throughput, and provide fairness in communication is of great importance in meeting the high demands of future wireless networks. Our approach could pave the way for further research in improving the performance of Massive MIMO systems, thereby advancing the potential of 5G and 6G networks.**

***Index Terms*—5G, 6G, Massive MIMO, User Scheduling**

I. Introduction

THe demand for high data rates has skyrocketed due to the growing usage of mobile devices and the emergence of new applications that require high-speed data transfer. This has led to the development of next-generation wireless systems such as 5G, beyond 5G, and 6G networks, which are expected to provide high data rates, low latency, and better quality of service. Multiple-input Multiple-output (MIMO) technology has been a key factor in the development of previous generation wireless networks such as 3G and 4G, and it is expected to continue playing a critical role in future wireless systems. MIMO technology utilizes multiple antennas at both the transmitter and receiver to create multiple signal paths, which can be used to improve the robustness of the link against fading and interference [1]–[6].

One of the major benefits of MIMO technology is diversity gain, which refers to the improvement in signal quality due to the use of multiple signal paths. By leveraging the spatial dimension, MIMO technology can create multiple independent signal paths between the transmitter and receiver, which can help mitigate the effect of fading on the signal strength. This is particularly useful in environments with a high degree of multipath propagation, such as urban areas. Another key benefit of MIMO technology is multiplexing gain, which refers to the increase in data rate due to the use of multiple signal paths. By sending independent data streams on each signal path, MIMO technology can effectively increase the bandwidth of the link, resulting in higher data rates. This is particularly useful in applications that require high-speed data transfer, such as video streaming, online gaming, and virtual reality.

To cater to more users with better quality of service, the MIMO technique called massive MIMO plays a critical role. Massive MIMO employs hundreds of antennas at the base station, serving multiple users simultaneously. Massive MIMO is a wireless access technology operating below 6GHz, which plays a crucial role in current 5G and upcoming 6G networks by offering high spectral and energy efficiency with

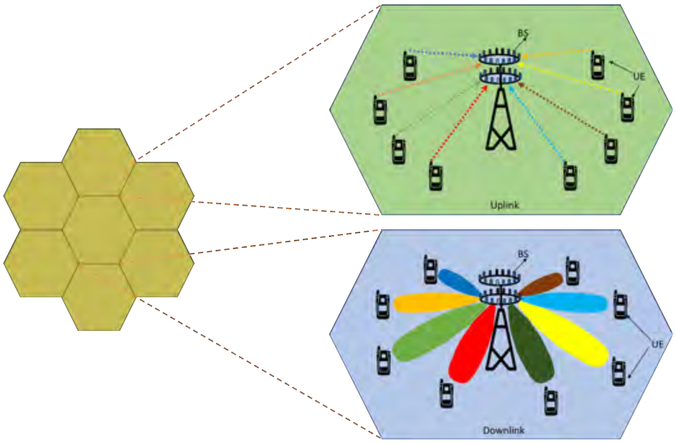


Fig. 1. Massive MIMO uplink and downlink system.

Manuscript submitted Dec 08, 2022. This is a follow-up of the invited journal to the accepted and presented paper entitled ”An Adaptive User Scheduling Algorithm for 6G Massive MIMO Systems” at the 25th International Conference on Advanced Communication Technology (ICACT2023).

Robin Chataut is with the School of Computing and Engineering, Quinnipiac University, Hamden, CT, USA, 06518; (corresponding author, e-mail: Robin.Chataut@quinnipiac.edu;).

Robert Akl is with the Department of Computer Science, University of North Texas, Denton, Texas, USA 76201; e-mail: rakl@unt.edu

low latency [9]- [16]. It uses hundreds of antennas at the base station to serve tens of users simultaneously, providing high multiplexing and diversity gains while mitigating fading effects. Massive MIMO is essential to support the increasing demand for high data rates driven by applications such as blockchain, cyber-security, Smart Vehicles, the Internet of Things, augmented reality, virtual reality, and extended reality. The technique uses beamforming to direct signals towards users during the downlink, and the narrower beams resulting from more antennas improve spatial focus. Figure 1 shows a typical massive MIMO system where uplink pilot signals are transmitted by users towards the base station during uplink communication, and the downlink communication uses beamforming to direct signals towards the users. As the number of antennas increases, the beams become narrower, resulting in better spatial focus on users [17], [18].

However, with hundreds of antenna terminals, user scheduling during downlink communication is one of the major challenges in massive MIMO system deployment. A suitable user scheduling method during the downlink is necessary to enhance the throughput of massive MIMO systems when the number of active users is greater than the number of base station antenna terminals. Scheduling users with better channel conditions can improve the total area throughput. However, maintaining an adequate fairness level is equally important to ensure timely scheduling for users with weaker channel conditions. Considerable research has been conducted to develop optimal user scheduling algorithms. Greedy algorithms have been discussed in [19]- [21], which provide better fairness performance but fail to achieve optimal throughput. Traditional algorithms, Round Robin (RR), and Proportional Fair (PF) are better in terms of fairness but do not achieve optimal fairness. Linear methods like Zero Forcing (ZF) and Minimum Mean Square Error (MMSE) have been explored in [22]- [23]. The authors in [24]–[26], [28]–[31] have investigated user scheduling methods for downlink MIMO systems, but optimal performance in terms of both throughput and fairness has not been achieved. In this paper, we propose an adaptive user scheduling algorithm based on channel rate to provide the user with optimal throughput and ensure fairness among all the users.

A. *Contribution of the Paper*

1) The user scheduling issue during the downlink massive MIMO system is investigated
2) An adaptive user scheduling scheme based on instantaneous channel rate is proposed
3) The sum rate, per-user throughput, and error performance of the proposed algorithm are accessed and compared with traditional scheduling algorithms
4) We evaluated the fairness index of the proposed algorithm. We have used Jain's fairness index to compute the fairness index.
5) The results obtained from the Matlab simulations show that the proposed algorithm is fair and performs better than the traditional user scheduling algorithm in terms of sumrate, per-user throughput, and error performance.

B. *Outline*

The remainder of the paper is structured as follows: Section II defines the downlink system model for massive MIMO with M antennas and N users. The proposed adaptive algorithm is described in III. The simulation steps, required parameters, and algorithm analysis are presented in IV. Finally, V concludes the paper by encapsulating the major concepts of the paper.

C. *Notations*

In this paper, there are specific notations and terminologies used to represent various mathematical concepts. Column vectors are denoted by lower-case letters, while matrices are denoted by upper-case letters. The inverse of a matrix is denoted by $(.)^{-1}$, and the transpose is represented by $(.)'$. The hermitian transpose is denoted by $(.)^H$. The circular symmetric complex Gaussian distribution with zero mean and co-variance V is represented by $\mathcal{CN}(0, V)$. The space of M-element complex vectors is denoted by $\mathbb{C}^M$, where M is a positive integer. The $M \times M$ identity matrix is represented by I_M, which is a square matrix with ones on the diagonal and zeros elsewhere.

II. SYSTEM MODEL

In massive MIMO, the BS is equipped with numerous antennas, typically numbering in the hundreds or thousands. Downlink is the data transmission from the base station to the user equipment, such as mobile phones or laptops. The fundamental concept behind massive MIMO is to exploit the large number of antennas at the BS to concurrently communicate with multiple users utilizing the same time-frequency resource. This is accomplished by spatially combining signals from the BS's antennas to create unique signal combinations for each user. Downlink massive MIMO systems capitalize on the numerous antennas at the BS to enhance the quality and capacity of the wireless link to the users. Beamforming methods are employed by the BS to direct the transmitted signal towards each user, increasing the signal-to-noise ratio (SNR) and reducing interference from other users.

A massive MIMO downlink system is considered with M base station antenna terminals and N users. In the course of the downlink communication, the base station will send an independent and autonomous signal to each active user. If U users are waiting for their turn to be scheduled, the base station selects S users ($S <= U$) according to the scheduling algorithm. The base station will apply a precoder before sending the downlink signal towards the user. The primary objective of precoding is to optimize the wireless channel between the base station and the users by modifying the phase and amplitude of the transmitted signal. The purpose of this modification is to reduce interference and enhance the quality of the signal received by the user. Precoding is a process that involves the application of a matrix operation to the data signals transmitted from the base station antennas. This operation is intended to decrease the interference between users and improve the signal-to-noise ratio (SNR) at the receiver's end. Several precoding algorithms are used in Massive MIMO systems, including zero-forcing (ZF) precoding and minimum

Copyright © 2024 Global IT Research Institute (GiRI)

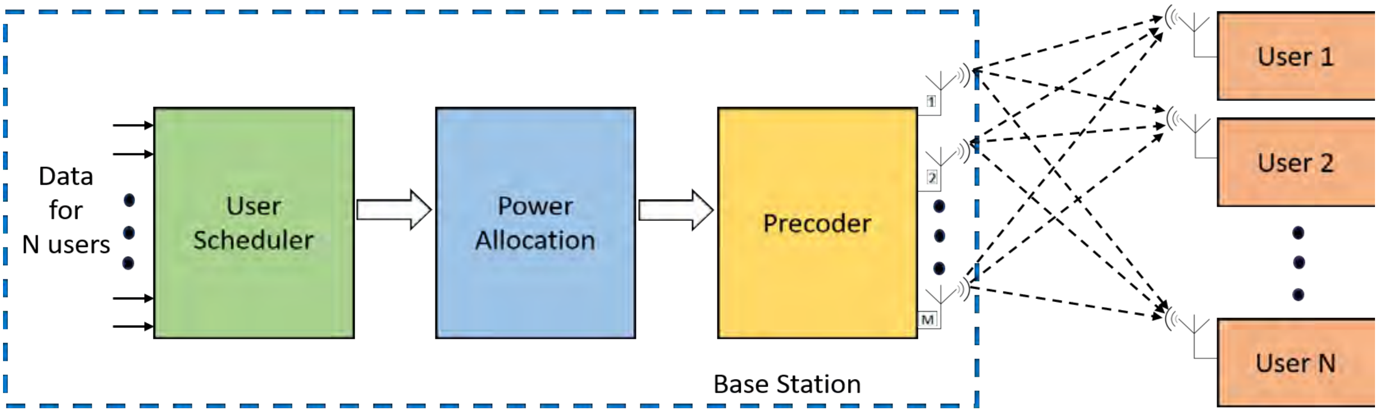


Fig. 2. System Model with M base station antenna serving N users.

mean squared error (MMSE) precoding. The implementation of precoding during user scheduling can lead to increased data rates, better spectral efficiency, and improved overall system performance in terms of signal quality and interference reduction. Therefore, precoding is a critical component in the design and optimization of Massive MIMO systems. The signal received by user i can be represented as:

$$y_i = Hx_i + n_i \tag{1}$$

Where,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ . \\ . \\ . \\ x_i \end{bmatrix} and \quad H_i = \begin{bmatrix} h^i_{11} & h^i_{12} & . & h^i_{1N} \\ h^i_{21} & h^i_{22} & . & h^i_{2N} \\ . & . & . & . \\ . & . & . & . \\ h^i_{N1} & h^i_{N2} & . & h^i_{NM} \end{bmatrix}$$

y_i is the signal received by the i_{th} user, and x_i is the signal sent towards the user from the base station i. $H \in \mathbb{C}^{N\times M}$is the channel vector between the user terminals and the base station antenna terminals, where elements of H are independent and identically distributed. n_i is the added white Gaussian noise at the i_{th}.

We do the precoding before scheduling the user to minimize multi-user interference. We get the matrix for precoding by stacking the beamforming vectors and user signals.

$$y_i = HWx_i + n_i \tag{2}$$

Where,

$$W = \begin{bmatrix} p_1 & p_2 & . & . & . & p_j \end{bmatrix}$$

$W \in \mathbb{C}$ is the precoding matrix, which contains a set of precoders. p_j is the vector used for precoding the j_{th} user. For our simulations, we have applied two simple linear precoders, ZF and MMSE [34]:

$$W_{ZF} = H^H(HH^H)^-1 \tag{3}$$

$$W_{MMSE} = H^H(HH^H + \sigma^2 I)^-1 \tag{4}$$

We compute the sumrate by considering the uniform power allocation among each user as [32]:

$$Sumrate = \sum_{i=1}^{N} \log_2\left(1 + \frac{|b_i h_i|^2}{1 + \sum_{j=1,j\neq i}^{N} |b_i h_j|^2}\right) \tag{5}$$

where, b_k is the k_{th} row of precoding matrix B and h_k is the k_{th} row of the channel matrix H.

III. PROPOSED ALGORITHM FOR DOWNLINK USER SCHEDULING

The proposed algorithm is summarized in 1. We initialize the active users set U, including N active users. The set of selected users is S, which is null initially as non of the users are scheduled. Then we calculate the instantaneous channel rate for each user:

$$C_j = log_2\left(1 + \sqrt{\sum_{j=1}^{N} |h_j|^2}\right) \tag{6}$$

The mean channel rate is computed based on the active users waiting to be scheduled. The calculated mean channel rate will also be the selection criteria for the proposed algorithm.

$$\bar{C} = \frac{\sum C_j}{N} \tag{7}$$

The user with an instantaneous channel rate closest to the mean channel rate is selected first. Once the selected user is scheduled, we update the set containing the remaining active and selected users.

$$\pi(j) = argmin|||A_j| \tag{8}$$

$$S = S\, U\, \pi(j) \tag{9}$$

$$U = U - S \tag{10}$$

The process of user selection is repeated until all the active users are scheduled. Then, the mean channel rate is re-evaluated for the next set of active users.

$$U \neq \{\phi\} \tag{11}$$

Algorithm 1 Proposed Algorithm for Massive MIMO Downlink Scheduling

```
Initialization:
1. $U = \{1, 2, 3, 4, ....N\}$
2. $S = \{\phi\}$
3. $j = 0$
Channel Rate Calculation:
4. $C_j = log_2\left(1 + \sqrt{\sum_{j=1}^{N} |h_j|^2}\right)$
5. $\bar{C} = \frac{\sum C_j}{N}$
Selection Criteria:
6. $A_j = |C_j - \bar{C}|$
Algorithm iteration:
do
    7. $\pi(j) = argmin|||A_j|$
    8. $S = S\,U\,\pi(j)$
    9. $U = U - S$
    10. $i = i + 1$
While $U \neq \{\phi\}$
```

IV. Simulation Results and Analysis

In this section, we analyze the results obtained from the Matlab simulations. For simulations, we set up a massive MIMO base station with many antenna terminals (16 to 512). We assume that all the antenna terminals are communicating with 128 single active users simultaneously. We have considered various antenna configurations with different modulation techniques (QPSK, 16QAM, 16QAM) for conducting the simulations. The system's bandwidth is set to 20 MHz, whereas a carrier frequency of 2.5 GHz is used. A perfect channel state information (CSI) is assumed between the user and the base station, and the Rayleigh fading channel model is used for simulations. We have compared our proposed algorithm with traditional schedulers like Proportional Fair (PF) and Round Robin (RR) algorithms for analysis. In addition, we have used ZF and MMSE precoding to reduce the effect of multi-user interference and to simplify the processing required at the receiver. The simulation parameters used are shown in I.

TABLE I
Simulation Parameters

Parameter	Value
Base Station Antenna Terminal	16 to 512
Number of Users	128
Carrier Frequency	2.5 GHz
Bandwidth	20 MHz
Coherence Internal	200 Symbols
Channel Model	Uncorrelated Rayleigh Fading
Signal Variance	2
SNR	0 dB - 25dB
Modulation	QPSK, 16QAM, 64QAM

Fig. 3 depicts the error performance of the proposed algorithm with 16 users, 16 base station antenna terminals, 16QAM modulation, and MMSE precoding. The proposed algorithm exhibits better BER performance than the traditional algorithm across the entire range of user SNR in the simulation. For instance, at a BER of 10^{-2}, the proposed algorithm achieves a 6dB gain over the RR algorithm and a 4dB gain over the PF algorithm. Similarly, conducting the same experiment with 16 users, 16 base station antenna terminals, and MMSE precoding, but with 64QAM modulation, results in degraded error performance for all algorithms, as shown in Fig. 4. At BER 10^{-1}, the proposed algorithm achieves almost 3 dB gain over the PF algorithm and 4dB gain over the RR algorithm. Nonetheless, the per-user throughput increases for all algorithms with higher modulation order. Additionally, the simulation with comparable parameters and QPSK modulation, depicted in Fig. 5, results in improved error performance for all algorithms. This improvement stems from QPSK being less susceptible to interference and noise in comparison to higher modulation orders such as 16QAM and 64QAM used in our experiments.

Fig.6 illustrates the simulation outcomes when 16 users,

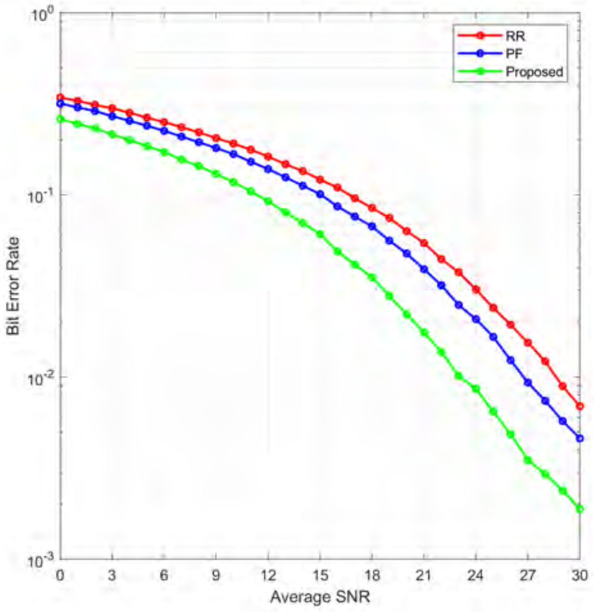


Fig. 3. BER vs. SNR performance with 16 users, 16 base station antennas, 16QAM modulation, and MMSE precoding.

Copyright © 2024 Global IT Research Institute (GiRI)

16 base station antenna terminals, and 16QAM modulation are employed, but with Zero Forcing (ZF) precoding. The performance trend follows the same pattern as the previous experiment; however, the overall performance of all algorithms has decreased. Specifically, at a BER of 10^{-1}, the proposed algorithm outperforms the RR algorithm by 5 dB and PF by 3.5 dB, underscoring the superiority of the proposed algorithm's BER performance compared to conventional algorithms. In Fig.7, QPSK modulation is employed, and all algorithms demonstrate enhanced error performance. This improvement can be attributed to the lower susceptibility of lower modulation orders, such as QPSK, to noise and interference. In contrast, higher modulation orders, such as 16QAM, are more susceptible to noise and interference, resulting in degraded error performance.

Fig.8 illustrates the analysis of the sumrate performance of the proposed algorithm. This simulation involved 16 base

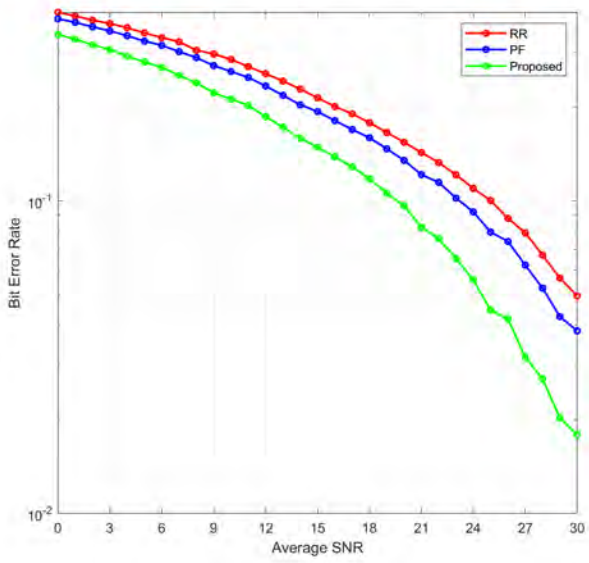


Fig. 4. BER vs. SNR performance with 16 users, 16 base station antennas, 64QAM modulation, and MMSE precoding.

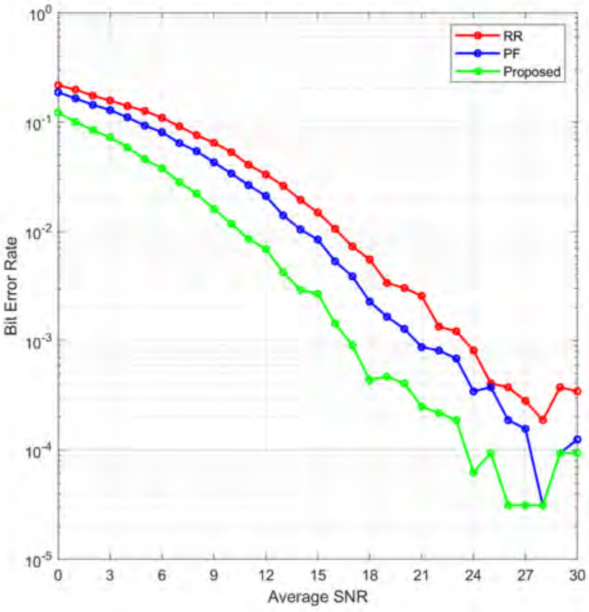


Fig. 5. BER vs. SNR performance with 16 users, 16 base station antennas, QPSK modulation, and MMSE precoding.

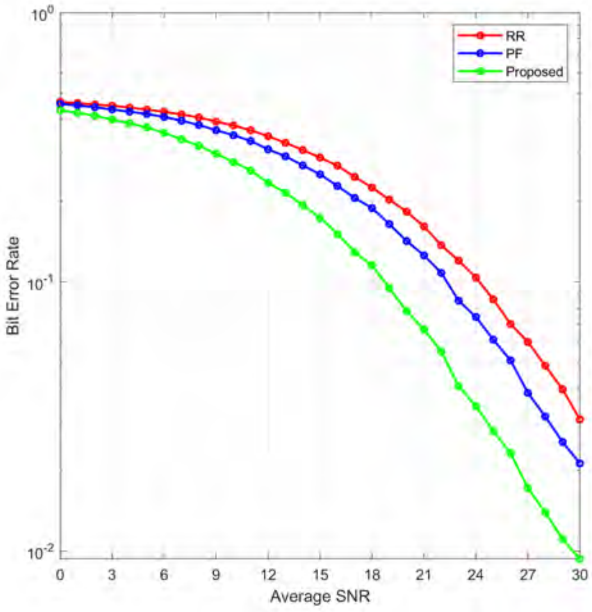


Fig. 6. BER vs. SNR performance with 16 users, 16 base station antennas, 16QAM modulation, and ZF precoding.

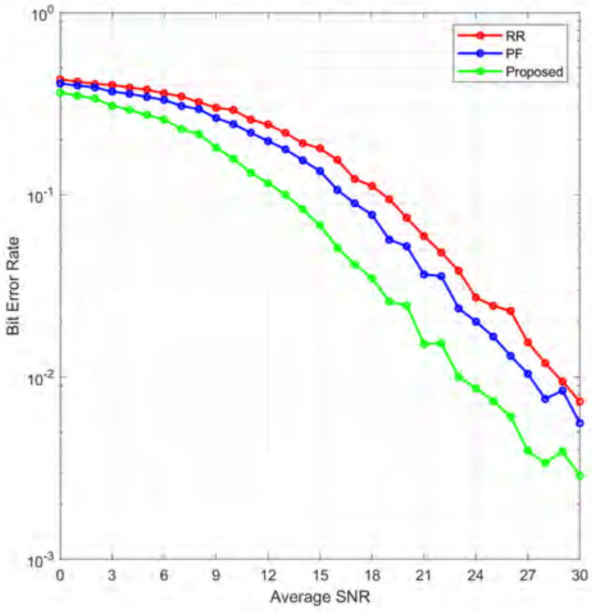


Fig. 7. BER vs. SNR performance with 16 users, 16 base station antennas, QPSK modulation, and ZF precoding.

Copyright © 2024 Global IT Research Institute (GiRI)

station antenna terminals communicating with 16 users using 16QAM modulation and MMSE precoding. The simulation results indicate that the proposed algorithm outperforms the traditional algorithms. For instance, at an SNR of 21dB, the proposed algorithm achieves a sum rate of 60 bits/s/Hz, whereas the PF algorithm attains a sum rate of 43 bits/s/Hz, and the RR algorithm exhibits the poorest performance, with a sum rate of 38 bits/s/Hz. The high sum rate is primarily attributed to the increased number of antenna terminals. Nevertheless, as the number of active users in a cell grows, the sum rate will eventually reach a saturation point.

We conducted a similar experiment using ZF precoding, and the sumrate performance was comparable to that of MMSE precoding, as demonstrated in Fig.9. With ZF precoding at an SNR of 21dB, the proposed algorithm attained a sumrate of 60 bits/s/Hz, while the PF and RR algorithms recorded a sum rate of 42 bits/s/Hz and 37 bits/s/Hz, respectively.

We then considered the performance of our proposed algorithm with several modulation techniques. This simulation was administered with 16 base station antenna terminals communicating with 16 users using 16QAM modulation and MMSE precoding. As shown Fig.10, QPSK exhibited the best error performance across a range of SNRs, while 64QAM displayed the best performance due to its ability to transmit more data per symbol. However, higher modulation orders are more susceptible to noise and interference, leading to higher error rates. Therefore, the optimal modulation order depends on the application and the user's requirements. Furthermore, we performed a simulation with ZF precoding using 16 base station antenna terminals communicating with 16 users via 16QAM modulation, as shown in Fig.11. We observed that the performance was nearly identical to that of MMSE precoding.

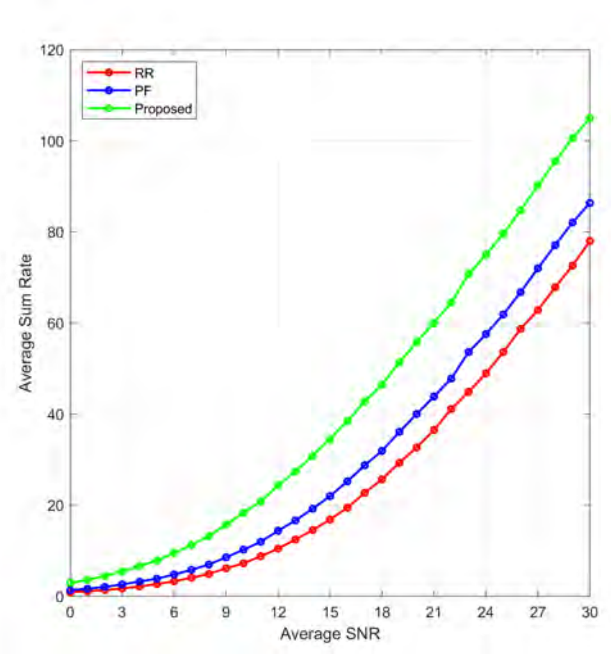


Fig. 8. Sumrate vs. BER performance with 16 users, 16 base station antennas, 16QAM modulation, and MMSE precoding.

We evaluated the proposed algorithm's average throughput per user performance. This simulation was administered with 16 base station antenna terminals communicating with 16 users using 16QAM modulation and MMSE precoding. As shown in Fig. 12, the average per-user throughput for the proposed algorithm was best among the compared algorithms. Our algorithm achieved a per-user throughput of 3.14 Mbps, whereas, for RR and PF algorithms, it was found to be 2.33 Mbps and 2.53 Mbps, respectively.

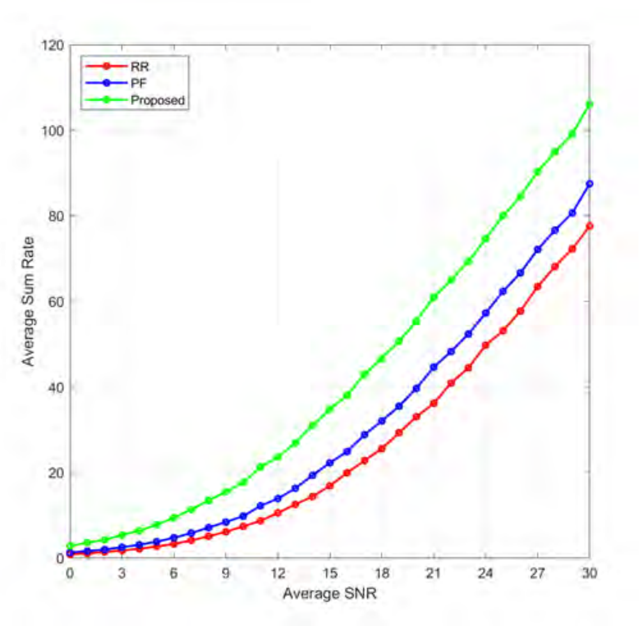


Fig. 9. Sumrate vs. BER performance with 16 users, 16 base station antennas, 16QAM modulation, and ZF precoding.

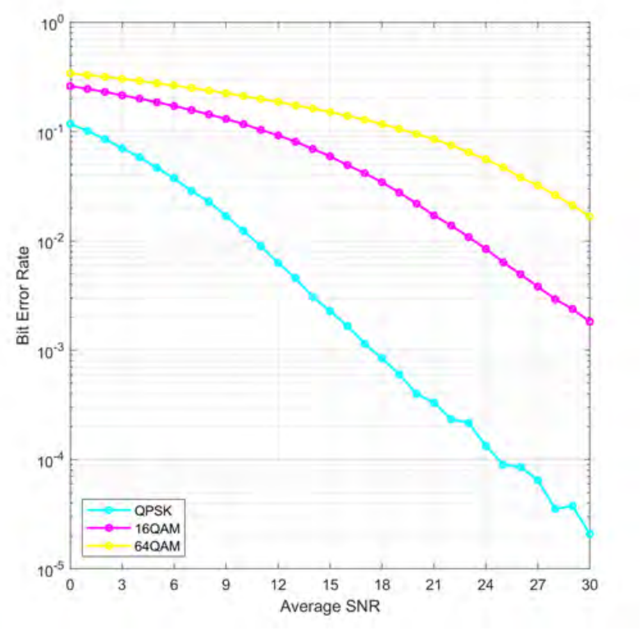


Fig. 10. BER performance of the proposed algorithm with several modulation schemes with 16 users, 16 base station antennas, and MMSE precoding

Copyright © 2024 Global IT Research Institute (GiRI)

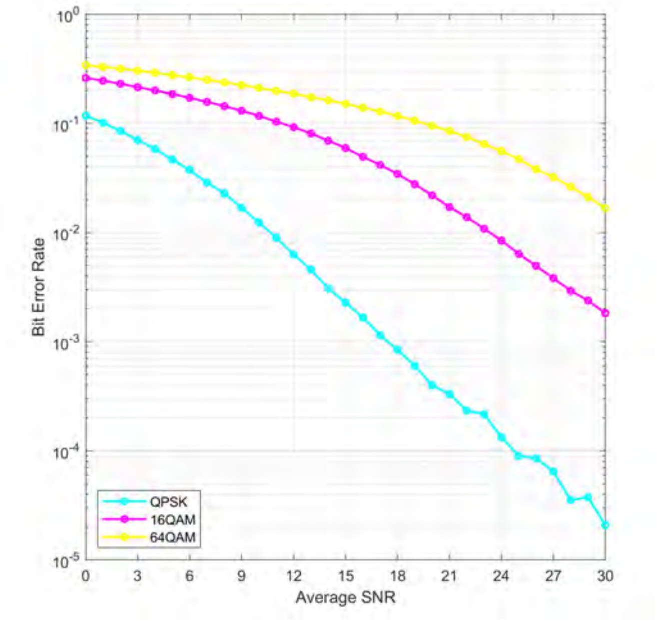


Fig. 11. BER performance of the proposed algorithm with several modulation schemes with 16 users, 16 base station antennas, and ZF precoding

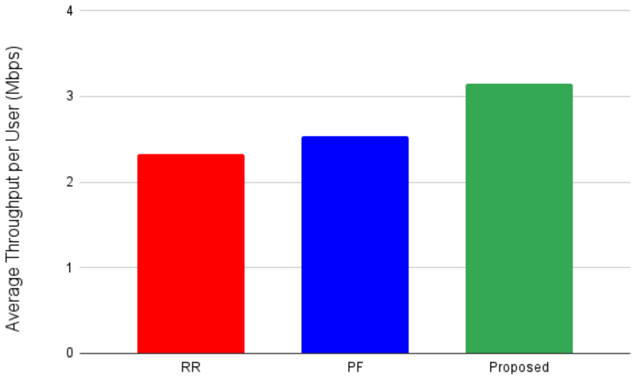


Fig. 12. Average throughput per user with 16 users, 16 base station antennas, 16QAM modulation, and MMSE precoding.

We use Jain's fairness index to evaluate the performance of the proposed algorithm. Jain's fairness index is a widely-used metric for assessing fairness in the distribution of limited resources, especially in the context of networking and telecommunications. This metric is particularly useful when multiple users or applications are competing for a finite amount of resources like CPU time or wireless bandwidth. It offers an objective way of quantifying the degree of fairness in resource allocation and comparing different allocation schemes. The fairness of resource allocation is critical in networking to prevent congestion, service degradation, or even network failure. Jain's fairness index allows for the comparison of resource allocation schemes by measuring how equitably resources are distributed among users or applications. An index value close to 1 suggests that resources are being allocated fairly to all users or applications, whereas a value closer to 0 indicates an unfair distribution, with some users or applications receiving more than their fair share of resources.

We measured Jain's fairness index for all the algorithms [33].

$$\mathcal{F}(X) = \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{\sum_{i=1}^{N} x_i^2} \tag{12}$$

Where $\mathcal{F}$ is the fairness index whose values are between 0 and 11, and x_j is throughput for ith user. As shown in II, simulation results show that the fairness provided by the proposed algorithm is similar to that of the traditional algorithms.

TABLE II
FAIRNESS INDEX COMPARISON

Scheduling Algorithm	Fairness Index
Round Robin	0.973
Proportional Fair	0.983
Proposed	0.999

V. CONCLUSION

In conclusion, this paper addressed the issue of user scheduling during downlink signaling in a massive MIMO system. The proposed algorithm takes into account the instantaneous channel rate, which enables it to adaptively schedule users based on their current channel conditions. This results in a significant improvement in the sum rate and per-user throughput, as well as providing better error performance and fairness among all users. The simulation results also showed that the performance of the proposed algorithm varied with different modulation techniques. Specifically, 64QAM provided the best data rate, while QPSK provided the best error rate. This indicates that the choice of modulation technique can significantly affect the performance of the user scheduling algorithm, and it is essential to choose an appropriate modulation technique that suits the requirements of the system.

Furthermore, the fairness of the proposed algorithm was assessed using Jain's fairness index, which is a commonly used metric for measuring fairness in communication systems. The fairness index of 0.99 obtained from the proposed algorithm indicates that the algorithm ensures fairness among all users, which is an essential requirement for any scheduling algorithm. The proposed adaptive user scheduling algorithm based on instantaneous channel rate is a suitable candidate for downlink user scheduling in a massive MIMO system with a large number of antennas. The algorithm provides improved performance in terms of sum rate, per-user throughput, error performance, and fairness, and can be adapted to different modulation techniques to suit the requirements of the system.

REFERENCES

[1] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, ”Toward 6G Networks: Use Cases and Technologies,” *IEEE Commun. Mag.*, vol. 58, pp. 55-61, 2020.

[2] A. Kurve, ”Multi-user MIMO systems: The future in the making,” *IEEE Potentials*, vol. 28, pp. 37-42, 2009.

Copyright © 2024 Global IT Research Institute (GiRI)

[3] G. J. Foschini and M. J. Gans, ”On Limits of Wireless Communications in a Fading Environment When Using Multiple Antennas,” *Wireless Pers. Commun.*, vol. 6, pp. 311-335, 1998.

[4] Q. H. Spencer, C. B. Peel, A. L. Swindlehurst, and M. Haardt, ”An introduction to the multi-user MIMO downlink,” *IEEE Commun. Mag.*, vol. 42, pp. 60-67, 2004.

[5] A. Paulraj and T. Kailath, ”Increasing Capacity in Wireless Broadcast Systems Using Distributed Transmission/Directional Reception (DTDR),” U.S. Patent 5,345,599, Sep. 6, 1994.

[6] D. Nojima, L. Lanante, Y. Nagao, M. Kurosaki, and H. Ochi, ”Performance evaluation for multi-user MIMO IEEE 802.11ac wireless LAN system,” in *Proceedings of the 2012 14th International Conference on Advanced Communication Technology (ICACT)*, PyeongChang, Korea, Feb. 19-22, 2012, pp. 804–808.

[7] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*. New York, USA: Cambridge University Press, 2008.

[8] IEEE Draft Standard Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Amendment 4: Enhancements for Higher Throughput. P802.11n D3.00, Sept. 2007.

[9] 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (EUTRA); Multiplexing and channel coding (Release 9). 3GPP Organizational Partners TS 36.212 Rev. 8.3.0, May 2008.

[10] J. Hoydis, K. Hosseini, S. Ten Brink, and M. Debbah, ”Making smart use of excess antennas: Massive MIMO, small cells, and TDD,” *Bell Labs Technical Journal*, vol. 18, no. 2, pp. 5-21, Sep. 2013.

[11] R. Chataut and R. Akl, ”Optimal pilot reuse factor based on user environments in 5G Massive MIMO,” *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, 2018, pp. 845-851.

[12] R. Chataut, R. Akl and U. K. Dey, ”Least Square Regressor Selection-Based Detection for Uplink 5G Massive MIMO Systems,” *2019 IEEE 20th Wireless and Microwave Technology Conference (WAMICON)*, Cocoa Beach, FL, USA, 2019, pp. 1-6.

[13] T. L. Marzetta, ”Noncooperative cellular wireless with unlimited numbers of base station antennas,” *Wireless Communications, IEEE Transactions on*, vol. 9, no. 11, pp. 3590-3600, 2010.

[14] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, ”Massive MIMO for Next Generation Wireless Systems,” *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186-195, Feb. 2014.

[15] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, ”Scaling up MIMO: Opportunities and Challenges with Very Large Arrays,” *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40-60, Jan. 2013.

[16] T. L. Marzetta, ”Massive MIMO: An Introduction,” in *Bell Labs Technical Journal*, vol. 20, pp. 11-22, 2015.

[17] R. Chataut and R. Akl, ”Massive MIMO Systems for 5G and beyond Networks—Overview, Recent Trends, Challenges, and Future Research Direction,” *Sensors*, vol. 20, no. 10, p. 2753, 2020.

[18] R. Chataut, R. Akl, U. K. Dey, and M. Robaei, ”SSOR Preconditioned Gauss-Seidel Detection and Its Hardware Architecture for 5G and beyond Massive MIMO Networks,” *Electronics*, vol. 10, no. 5, p. 578, 2021.

[19] G. Dimic and N. D. Sidiropoulos, ”On downlink beamforming with greedy user selection: performance analysis and a simple new algorithm,” *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3857-3868, Oct. 2005.

[20] J. Wang, D. J. Love, and M. D. Zoltowski, ”User selection with zero-forcing beamforming achieves the asymptotically optimal sum rate,” *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3713-3726, Aug. 2008.

[21] M. Kobayashi and G. Caire, ”Joint beamforming and scheduling for a multi-antenna downlink with imperfect transmitter channel knowledge,” *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1468-1477, Sep. 2007.

[22] K. Lyu, ”Capacity of multi-user MIMO systems with MMSE and ZF precoding,” in *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, San Francisco, CA, 2016, pp. 1083-1084.

[23] D. L. Colon, F. H. Gregorio, and J. Cousseau, ”Linear precoding in multi-user massive MIMO systems with imperfect channel state information,” in *2015 XVI Workshop on Information Processing and Control (RPIC)*, Cordoba, 2015, pp. 1-6.

[24] M. Sharif and B. Hassibi, ”On the capacity of MIMO broadcast channels with partial side information,” *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 506–522, Feb. 2005.

[25] T. Al-Naffouri, M. Sharif and B. Hassibi, ”How much does transmit correlation affect the sum-rate scaling of MIMO Gaussian broadcast channels?,” *IEEE Trans. Commun.*, vol. 57, no. 2, pp. 562–572, Feb. 2009.

[26] T. Yoo and A. Goldsmith, ”On the optimality of multi-antenna broadcast scheduling using zero-forcing beamforming,” *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

[27] H. Yang, ”User Scheduling in Massive MIMO,” in *Proceedings of the 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Kalamata, Greece, 25–28 June 2018, pp. 1–5.

[28] S. Huang, H. Yin, J. Wu, and V.C.M. Leung, ”User selection for multi-user MIMO downlink with zero-forcing beamforming,” *IEEE Trans. Veh. Technol.*, vol. 62, pp. 3084–3097, Sept. 2013.

[29] Y. Cai, J. Yu, Y. Xu, and M. Cai, ”A comparison of packet scheduling algorithms for OFDMA systems,” in *Proceedings of the 2008 2nd International Conference on Signal Processing and Communication Systems*, Gold Coast, QLD, Australia, 15–17 Dec. 2008, pp. 1–5.

[30] M.F. Hamdi, R.A. Saeed, and A. Abbas, ”Downlink scheduling in 5G massive MIMO,” *J. Eng. Appl. Sci.*, vol. 13, pp. 1376–1381, Apr. 2018.

[31] K. Djouani, G. Maina, M. Mzyece, and G. Muriithi, ”A low complexity greedy scheduler for multiuser MIMO downlink,” in *Proceedings of the Southern African Telecommunications Networks and Applications Conference (SATNAC)*, Port Edward, South Africa, 5–8 Sept. 2010.

[32] R. Chataut and R. Akl, ”Channel Gain Based User Scheduling for 5G Massive MIMO Systems,” in *2019 IEEE 16th International Conference on Smart Cities: Improving Quality of Life Using ICT and IoT, and AI (HONET-ICT)*, Charlotte, NC, USA, 2019, pp. 049-053.

[33] R. Jain, D. Chiu, and W. Hawe, ”A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Systems, Digital Equipment Corporation,” *Technical Report DEC-TR-301*, Tech. Rep., 1984.

[34] R. Akl, ”An Efficient and Fair Scheduling for Downlink 5G Massive MIMO Systems,” in *11th IEEE Texas Symposium on Wireless and Microwave Circuits and Systems (TSWMCS 2020)*, May 2020.

Robin Chataut is an assistant professor in the School of Computing and Engineering at Quinnipiac University, Hamden, USA. He obtained his undergraduate degree in Electronics and Communication Engineering from Pulchowk Campus, Tribhuvan University, Nepal, in 2014 and his Ph.D. in Computer Science and Engineering from the University of North Texas, Texas, USA, in 2020. Prior to completing his Ph.D., he worked as a senior software developer.

His research interests are in the areas of wireless communication and networks, 5G, 6G, and beyond networks, vehicular communication, smart cities, Internet of Things, wireless sensor networks, and network security. He has designed, implemented, and optimized several algorithms and hardware architectures for precoding, detection, user scheduling, channel estimation, and pilot contamination mitigation for massive MIMO systems for 5G and beyond networks. He has authored and co-authored several research articles. He is an active reviewer in several international scientific journals and conferences.

Robert Akl received his B.S. in Computer Science and B.S. in Electrical Engineering in 1994, his M.S. in Electrical Engineering in 1996, and his D.Sc. in Electrical Engineering in 2000, all from Washington University in Saint Louis. He is currently a Tenured Associate Professor at the University of North Texas and a Senior Member of IEEE. He has designed, implemented, and optimized both hardware and software aspects of several wireless communication systems for cellular, Wi-Fi, and sensor networks.

Dr. Akl has broad expertise in wireless communication, Bluetooth, Cellular, Wi-Fi, VoIP, telephony, computer architecture, and computer networks. He has been awarded many research grants by leading companies in the industry and the National Science Foundation. He has developed and taught over 100 courses in his field. Dr. Akl has received several awards and commendation for his work, including the 2008 IEEE Professionalism Award and was the winner of the 2010 Tech Titan of the Future Award.

Copyright © 2024 Global IT Research Institute (GiRI)

A Blockchain-based Security Assessment Framework

N. Satyanarayana
eSecurity De*partment, Centre for Development of Advanced, Computing* Hyderabad, India
nanduris@cdac.in

***Abstract*–Using Blockchain Technology for Security Assessment results in effective monitoring capabilities especially when data analytics components are inbuilt in such a system. At present days, we can see the availability of many Security Information and Event Management (SIEM) tools that follow a client-server model for capturing data from different resources and performing data analysis on the server side. However, such tools serve the purpose of a single institute and depend on the trust level in a multi-institute or multi-center-project kind of environment where they can be used. Another limitation could be that if the server is attacked, the whole exercise would be futile. The lack of trust and concerns about data integrity in such an environment makes performing root cause analysis of security risks difficult. Blockchain technology ensures a tamper-proof, time-stamped, and decentralized storage repository that helps in maintaining data integrity even in complex and untrusted multi-institute or multi-center-project environments while assuring data provenance. This article presents a unified and comprehensive security assessment framework that produces a compliance report along with threat perception level by monitoring and assessing resources across multi-institute or multi-center-project in different geographical locations while supporting data privacy by leveraging Blockchain Technology capabilities.**

Keywords - Blockchain, Security Assurance Policy, Continuous Monitoring

I. Introduction

Ccontinuous monitoring of critical digital resources, processes, networks, and resource utilization patterns, and reporting the incident about the observed violations is an essential part of any organization's security framework. The success of an audit process depends on data integrity while running security assessment tools such as SIEM tools. SIEM tools produce reports instantaneously when they run in a machine. For a detailed analysis of security assessment history of events and related data need to be captured in a manner such that the data integrity is maintained forever. Hence, we need a technology that maintains data provenance in a tamper-proof and time-stamped manner so that the security framework is assured of data integrity at any time.

Moreover, such a provision will help SIEM tools to produce more effective reports when data analytics components are integrated for fine-grained analysis. Apart from the above, there should be a mechanism using which one can see to what extent the underlying security policy is conformant and its current severity level to indicate a perceived threat. If unique state replication of data provenance is ensured at the premises of the service provider and other stakeholders then compliance to the organization's security policy framework can be provided as a Software Service with continuous monitoring capabilities in a decentralized manner.

This kind of arrangement ensures transparency and trust in the organization's security policy assessment process among the stakeholders. Blockchain Technology ensures unique state replication of data in the underlying network based on consensus.

In a multi-institute or multi-center-project scenario where critical resources are spread across different geographical locations, we can utilize a permissioned Blockchain-based security assessment mechanism for resource monitoring and security analysis. The usage of permissioned Blockchain eliminates the concerns about the integrity of security-relevant data that is available at different geographical locations as it maintains data in a tamper-proof, authentic, and shares information through a secured communication channel. On top of that the permissioned Blockchain also supports data privacy by sharing security-relevant data between intended members/stakeholders only.

Blockchain Technology is a distributed ledger technology that ensures unique state replication across the participating nodes in a tamper-proof and time-stamped manner. The data once stored cannot be modified or deleted. This is because data will be stored in the form of a block whose header contains a hash of the previous block and so on. The same process of appending a new block to the existing chain of blocks in each of the participating nodes will be ensured by the consensus algorithms in the Blockchain network. Hence, to modify data in the Blockchain network, an attacker not only has to recompute the hash of the corresponding block but also the hash of the next blocks up to the end block. And this entire effort has to be replicated in each and every participating node of the Blockchain network. Considering the effort required to modify the data in the Blockchain network we can be assured of the security and integrity of data that is stored in the Blockchain network. There are tools that provide Application Programming Interface to collect various metrics from the cloud providers and leave it to the interested party to deduce the inference, information about violations, and corresponding mitigation plan in a particular machine/VM etc. There is also research work done on using

Manuscript received Dec 09, 2022. This work was supported by the Centre for Development of Advanced Computing, Hyderabad through a funded project sponsored by the Ministry of Electronics and Information Technology, Government of India. The manuscript is submitted in the follow-up of the invited journal to the accepted and presented paper entitled "A Blockchain-based Security Information and Event Monitoring Framework" of the 25th International Conference on Advanced Communication Technology (2023).

N. Satyanarayana is a Joint Director at the Centre for Development of Advanced Computing, Hyderabad, India. (e-mail: nanduris@cdac.in).

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

Blockchain technology to maintain the provenance of data objects corresponding to the cloud platform. To the best of our knowledge, a unified approach where evidence collection in a tamper-proof and time-stamped manner and maintenance of the same based on consensus among its stakeholders in a distributed storage system on one side and presenting the standards conformance along with perceived severity level in the whole system on another side is not present. We addressed this problem in our paper.

Major contributions of this paper include,

- Maintenance of data provenance of security log information in a tamper-proof and time stamped using Hyperledger Fabric Blockchain Platform.
- Design of a unified security assurance platform considering best practices for preventing security threats and a corresponding verification mechanism with information about perceived severity threat level.
- Provision of Security Assessment as a Service with unique state replication across multiple nodes from which stakeholders can ascertain information about current security policy conformation.
- Describes an architecture wherein multi-centre or multi-centre-project stakeholders can perform data analysis in an independent manner while maintaining data in a decentralized environment.
- Briefs about how to circumvent issue of dealing with storage space availability in this distributed and decentralized system.

In section 2, we shall present the background study, our approach detailing the cyber security policy framework model and implementation details in section 3, results in section 4, and conclusion in section 5.

II. Related Study

Hyperledger Fabric (HLF) is a Linux Foundation's Blockchain initiative. The transaction flow (data) of HLF, ledger maintenance, and the purpose of the channel can be seen in [3][20][21]. Researchers claimed that up to 2500 tps could be achieved by them when HLF is used as a Blockchain platform [19]. In HLF, a channel is nothing but a logical subnetwork that binds its members together so that any information can be exchanged among the members themselves and others will not be able to see the data exchange. We can logically group auditors, service providers, and customers of a business entity as members of one or more logical organizations and can make them subscribe to a particular channel. In HLF such logical organizations become part of a consortium. We can have multiple consortiums configured in HLF so that no two consortium members can share data between themselves because of a channel as described above.

Major security vulnerabilities that were exploited in Cloud services were reported in Top Security Threats in Cloud Computing [4] by Cloud Security Alliance (CSA). From the information provided in [4] it can be inferred that 24x7 monitoring of security baseline controls such as password quality verification, continuous monitoring of behavioral anomalies w.r.t users, processes etc., and network policies etc., would provide a protective or vigilant cloud environment. The CSA Cloud Controls Matrix (CCM) is a cyber-security control framework for cloud computing, composed of 133 control objectives that are structured in 16 domains. It can be used as a guide to determine which security controls should be implemented by which actor for the systematic assessment of a cloud implementation. The controls in the CCM are mapped against industry-accepted security standards, regulations, and control frameworks including but not limited to: ISO 27001/27002/27017/27018, NIST SP 800-53, AICPA TSC, ENISA Information Assurance Framework, German BSI C5, PCI DSS, ISACA COBIT, NERC CIP, and many others [5].

Cloudwatch[6] from Amazon collects monitoring and operational data in the form of logs, metrics, and events providing the end user with a unified view of AWS resources, applications, and services that run on AWS and on-premise servers. This is a service provided by Amazon itself for monitoring the health information of its instances and other relevant data to the end-users. Lynis [7] is a security tool for systems running Linux, macOS, or Unix-based operating systems. It performs an extensive health scan of systems to support system hardening and compliance testing. CIS-CAT (CIS-Configuration Assessment Tool) [8] compares the configuration of target systems to the security configuration settings recommended in machine-readable content, provided the content conforms to Security Content Automation Protocol (SCAP). Both the tools (Lynis and CIS-CAT) generate security-relevant information of critical resources in the same physical asset where the tool is deployed and depends on the running status of certain services such as "auditd" in Linux-based systems for the collection of information. These tools either leave the job of security assessment inference to the end-users through an API framework or they can be used in single nodes.

Several researchers have identified the need for addressing cloud security and explored using Blockchain in Cloud Platforms for security-related aspects [1][2][15][16][17]. Blockchain-based data provenance architecture for cloud environments has been proposed by [9][14][18]. In their approach, the focus was mainly on providing the ability to audit or privacy protection of cloud data objects or related operations of a cloud platform. The emphasis was more on using the data provenance capability of the underlying blockchain network for security analysis.

III. Our Approach – Model, Policy Framework and Implementation

A. Security Assurance Model

From the background study, we can observe that several efforts have been made in Cloud security by different organizations or standard bodies, primarily CSA, ISO, and individual cloud platform vendors. However, it is confusing which standard to follow from a plethora of standards by different organizations or else which platform to choose from the available platforms in the market as each one has its own metrics resulting in vendor lock-in. Each effort is in its own direction with a common goal of ensuring cloud security assurance. We need a holistic view of the security assurance framework. To fill this gap, we developed a security assurance policy framework based on the model defined in Fig. 1. and implemented monitoring of 14 controls as given in Table I below.

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

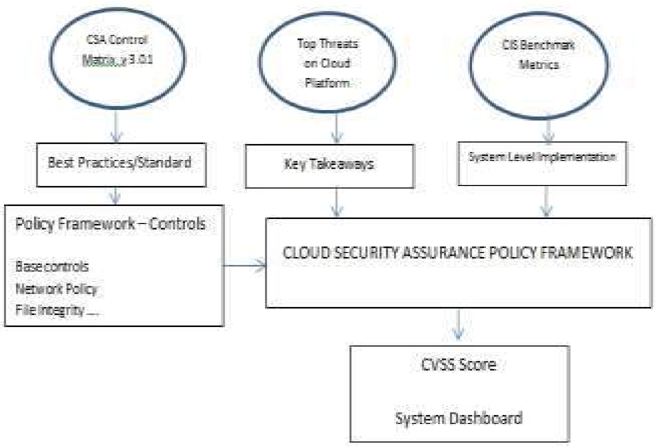

CSA Control Matrix, v 3.01
Top Threats on Cloud Platform
CIS Benchmark Metrics
Best Practices/Standard
Key Takeaways
System Level Implementation
Policy Framework – Controls
Base controls
Network Policy
File Integrity ....
CLOUD SECURITY ASSURANCE POLICY FRAMEWORK
CVSS Score
System Dashboard

Fig. 1. Security Assurance Policy Framework Methodology

The list is not limited to these controls itself and can be extended in the future. The controls are all relevant to the key takeaways that were studied based on CSA Top Cloud Threats and corresponding preventive mechanisms. These controls have been implemented for continuous monitoring using the Centre for Internet Security benchmark metrics [10]. These controls have been mapped to different control domains of CSA Control Matrix v 3.0.1 in order to present the compliance w.r.t Cloud Control Matrix best practices. In Table I below, column 1 represents the description of the control being monitored, and column 2 represents the Cloud Control Matrix control domain to which the control name is mapped. When monitoring of a control name is implemented based on CIS benchmark metrics, its security threat perception can be graded in accordance with the Common Vulnerability Scoring System (CVSS) score. The combination of the control name, it's mapping to the control domain, and its CVSS score thus form the security assessment policy framework from which the overall compliance of all critical resources of the business entity can be ascertained. The CVSS score could be one of None, Low, Medium, High, or Critical. The scores would be calculated in pursuance to [12]. The rule engine to calculate the CVSS score for each policy rule can be obtained based on the below-shown pseudo code.

TABLE I
UNIFIED CLOUD SECURITY ASSURANCE POLICY FRAMEWORK

Control Name	CCM Control Domain
File Integrity	AIS-04 Data Security/Integrity
Apache Loaded Modules	IVS-07 (OS Hardening and Base Controls)
User Login Attempts	IVS-07 (OS Hardening and Base Controls)
Password Policy	IVS-07 (OS Hardening and Base Controls)
Secure Boot Setup	IVS-07 (OS Hardening and Base Controls)
Process Monitoring	IVS-07 (OS Hardening and Base Controls)
Network Policy (IP Tables)	IVS-06 (Network Security) IAM-01 (Audit Tool Access) IVS-07 (OS Hardening and Base Controls)
Cron Service	IVS-07 (OS Hardening and Base Controls)
Syslog Configuration	IVS-07 (OS Hardening and Base Controls), IAM-01 (Audit Tool Access)
Secure Service User Accounts	IVS-07 (OS Hardening and Base Controls), IAM – 01 (Audit Tool Access)
Apache Logs Configuration	IVS-07 (OS Hardening and Base Controls), IAM – 01 (Audit Tool Access)
Apache User Configuration	IVS-07 (OS Hardening and Base Controls), , IAM – 01 (Audit Tool Access)
TCP Wrapper	IAM – 01 (Audit Tool Access), IVS-06 (Network Security)
SSH Service Configuration	IVS-07 (OS Hardening and Base Controls), IAM – 01 (Audit Tool Access)

Implementation of each control name may have a dependence on the occurrence of more than one event associated with it. For example, Process monitoring control compliance would be verified based on various events such as (a) Whether a process is system-oriented or network-oriented, (b) if it is network oriented whether access restrictions are present in the network policy or not (c) whether the log information corresponding to the process has appropriate access permissions or not (d) if it is network oriented whether the process owner has "nologin" shell or not. For each event, the CVSS score would be computed and the highest severity level among all the events that correspond to a control name would become that control's severity level.

Pseudo code for determining CVSS score

```
foreach (observed_error) {
   if (user_account has loginshell) {
            if (network_outer_allowed) {
               if (public_ip) {
                  av = network;
               } else {av = adjacent'}
            } else {av = local}
   } else {av = physical;}

   if (user_account == nologin_shell) {
           ac = high; scope_change = false; pr = high;
      confidentiality = none;
      availability = none;
   integrity = none; }
     elsif (passwd_policy == weak &&
                          user_account == login_shell) {
      ac = low; and scope_change = true; pr = low;
      confidentiality = high;
      availability = high;
      integrity = high;
   }elsif (passwd_policy == strong && user_account ==
login_shell) {
   ac = high; and scope_change = false; pr = high;
         confidentiality = low;
      availability = low;
      integrity = low;   }
    }
```

This model gives three important dimensions from which the security assessment of the business entity can be ascertained. Security Threat Perception of a critical resource being monitored. (2) Strength/Weakness of the critical resource in a particular domain in a particular control domain in accordance with the CCM. (3) Overall compliance to the security assessment policy framework of the business entity considering all the critical resources.

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

B. Implementation of Security Assurance Framework – System Architecture

In order to realize the effectiveness of the policy framework the policy has to be implemented as a software module. Fig. 2. depicts the system architecture which implements the proposed security policy framework using Blockchain Technology.

We used Hyperledger Fabric (HLF) [11] as a blockchain platform. In each resource that is being monitored, there will be an agent program running hereinafter called 'Agent'. The agent program reads the log contents from the respective resource corresponding to each observable control as defined in Table 1, encapsulates the same in a JSON object format, and transfers the same to the Blockchain platform. During this process each resource that is being monitored encrypts the JSON object using its private key and the same will be sent to the HLF Blockchain node. The array containing error ids represents events where policy violations have been detected. A hash value of all the fields in the JSON object is computed and is also made part of the JSON object. This is useful for data validation upon receipt by the Blockchain node. The JSON object format to be stored in the Blockchain is as follows.

```
{
   "hashvalue":" ab232lalkn23,nlk….",
   "data": {
     "host":"xx.xx.xx.xx",
         "timestamp":"02-02-2021",
           "valid":"false" //in case of detection of   policy
                                violation
           "result":[errid1,   errid2,   errid3….]   //list   of
           observed violations
      "metricid":"3232" //policy control id
          "condition": {
          Loginshell: [portno, exists or not exists, uname]
              public_ip: yes or no //error can occur from
              outer network
          }
   }
}
```

Blockchain node upon receiving the JSON object would decrypt the received JSON object using the resource public key and recalculates the hash of all the fields in the JSON object and verify its integrity. Modified JSON objects would be rejected and will not be stored in the Blockchain node. This verification mechanism is implemented as a smart contract so that the same validation rules will be applied to each and every node of the Blockchain network.

C. Role of Blockchain Nodes

The advantages of using Blockchain in this architecture are (a) The security policy-relevant information collected from each resource gets replicated in all the Blockchain nodes in a tamper-proof and time-stamped manner based on the underlying consensus mechanism (b) Reduces log processing load on the resource being monitored by delegating the data processing to the Blockchain platform. (c) Due to consensus-based unique state representation of data across all Blockchain nodes, all stakeholders viz. service providers, consumers, and auditors of a business entity would see the same data provenance or its analysis report at any given time without loss of generality. This unique state representation of data in the HLF blockchain network is supported by the Raft consensus algorithm [13] which is considered a robust crash-tolerant consensus algorithm. These Blockchain nodes each can be kept at different geographical locations or in a single data center as per the business entity's need. Since this is a permissioned Blockchain only authorized users can access/retrieve data from the Blockchain nodes. A more detailed use case scenario is explained in the next section.

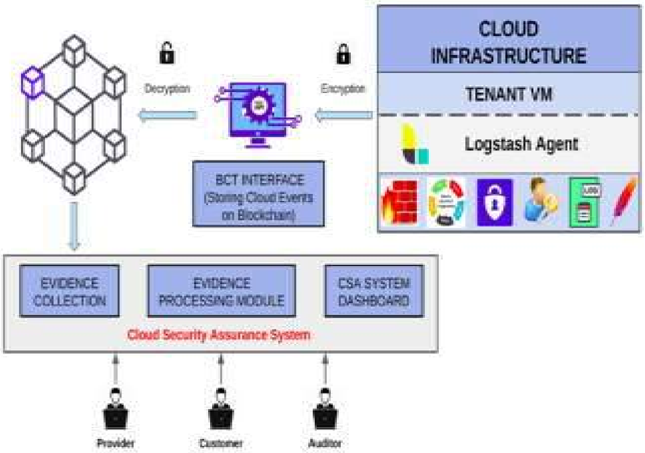


Fig. 2. The architecture of the Security Assurance Framework powered by Blockchain

In this system architecture, we defined three logical organizations namely Customer, Service Provider, and Information Security Auditor. Each logical organization can be assumed as a representation of real-time entities of respective stakeholders. The advantage of using three different logical organizations is that the nodes which are part of these organizations will maintain a unique state of log data as well as users of respective logical organizations only can initiate transactions (insert/query) on Blockchain network. Since logical organizations have been defined based on the functionalities of respective stakeholders who are concerned about the security aspects of the business entity, and all stakeholders are ensured of unique state replication with tamper-proof capability the security assessment of the business entity can be regarded as a fool-proof, robust and trustworthy system. Users from different geo-graphical locations can access such a system as Software-as-a-Service. Having been informed about the role of the Blockchain network and its necessity, we now focus on how the system reports its findings. For this purpose, we have incorporated two different metrics (a) Compliance rate – Which informs to what extent the service provider is compliant with the pre-defined security assurance policy as per table 1, (b) Severity level – Which informs the perceived threat severity level as per the criteria defined by CVSS scores to measure the overall health report of the resources controlled by the business entity.

Compliance Rate (CR) = $(X_{t-1} – X_t) / X_{t-1}$ where X_t is no of errors observed w.r.t security assurance policy at time t and X_{t-1} is at time t-1. CR represents the observed rate of change w.r.t identified non-compliance factors of security policy. The CR represents the compliance rate against the security policy pertaining to the most recent time window. Time window represents the gap between two successive vulnerability analysis attempts. Time window is a configurable parameter that defines how frequently the security log and other critical information have to be collected from the critical resources to be monitored.

Average Compliance Rate = $1/n \sum CRi$ where 'CR$_i$' is the value of the compliance rate at a given instant of a time window and 'n' is the no of such time windows chosen in a given period (e.g., in the last 24 hours, in the last 30 days etc). Severity Level is one of 'Critical', 'High', 'Medium', 'Low', 'None' labels which is decided based on Base Score. The base score is computed as shown below. The base score and other formulae computation are done in accordance with section 7.1 of CVSS specification.

Impact =	
If Scope is Unchanged	6.42 * ISS
If Scope is Changed	7.52 * (ISS – 0.029) – 3.25 * (ISS – 0.02) ^{15}
Exploitability =	8.22 * AttackVector * AttackComplexity * Privileges Required * User Interaction
Base Score =	
If Impact <= 0	0, else
If Scope is Unchanged	Roundup(Minimum[(Impact + Exploitability),10])
If Scope is Changed	Roundup(Minimum[1.08 * (Impact + Exploitability),10])

(Source: https://www.first.org/cvss/v3.1/specification-document)

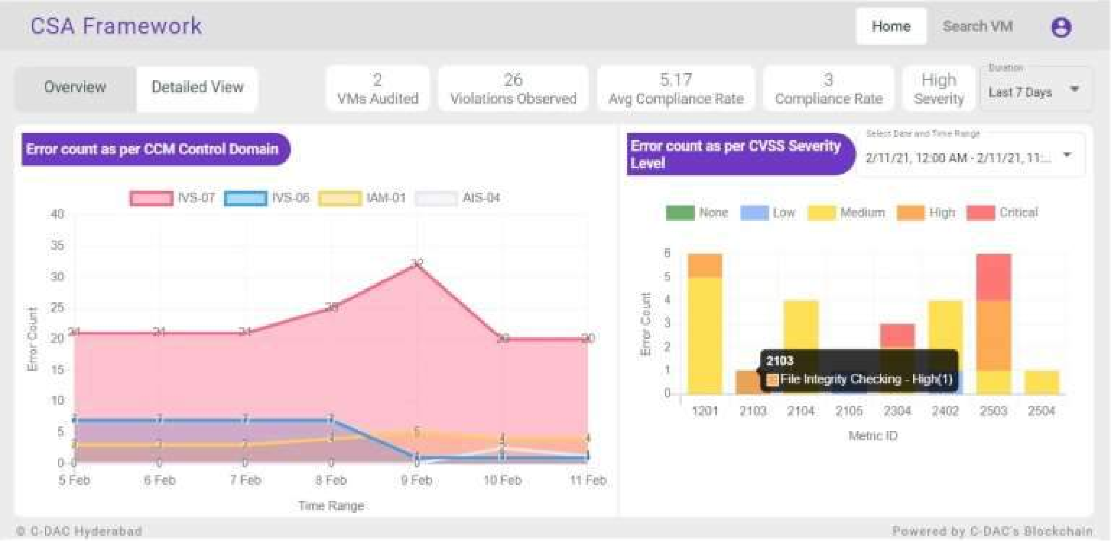
The Impact Sub-score (ISS) is calculated as $[1 - [(1 - confidentiality) * (1 - integiry) * (1 - Availability)]$

The severity level is decided based on base score value ranges as follows.

Rating	CVSS Score
None	0.0
Low	0.1 – 3.9
Medium	4.0 – 6.9
High	7.0 – 8.9
Critical	9.0 – 10.0

IV. Implementation & Results

We have implemented an 'Agent' code using "logstash" for each control that is described in the security policy framework depicted in Table 1 above. The log information from each resource is routed through an intermediate gateway developed using Node.JS server-side program which acts as a client to the HLF Blockchain network. The Node server upon receiving the encrypted JSON object sends the same unaltered to the Blockchain network. The smart contract later decrypts the JSON Object and stores the same in the Blockchain network. The flow of user interaction with the system is depicted in Fig. 3. below. A Dashboard component also has been developed using the Angular programming framework which can be accessed from the web by end-users who are members of service providers, customer, and auditor organizations. The dashboard component displays the no of violations against the security policy framework as described in Table I in a graphical format. It also displays various metrics such as no of resources audited in the platform, the average compliance rate over a period of time (e.g., last 24hrs, 7 days, 30 days). We also wanted to understand the effectiveness of the security policy compliance rate being computed by the system. To observe this we relied upon measuring severity levels in accordance with CVSS scores. If the security policy has any impact, then it should get reflected in observed severity levels.



CSA Framework

2 VMs Audited

26 Violations Observed

5.17 Avg Compliance Rate

3 Compliance Rate

High Severity

Duration: Last 7 Days

Select Date and Time Range: 2/11/21, 12:00 AM - 2/11/21, 11:...

2103
File Integrity Checking - High(1)

© C-DAC Hyderabad

Powered by C-DAC's Blockchain

Fig. 3. Display of resources being monitored and their stats

Fig. 3. above depicts the snapshot of the dashboard where no of resources audited, compliance rate, severity level etc. The dashboard also provides a detailed error report comprising an error description, its impact, and a mitigation plan for each and every violation that is observed by the system.

In Fig. 3., it can be observed that the left side graph shows the observed events information which are considered as violations as per the underlying security policy. The graph shows the details in accordance with the CCMv3 guidelines of Cloud Security Alliance. The right side of the graph

Copyright © 2024 Global IT Research Institute (GiRI)

displays the no of violations observed corresponding to each category of events in terms of their CVSS score.

Fig. 4., depicts the user interface screen wherein the end-user can observe the different types of errors or violations that have been monitored by the underlying system, its severity level, and also the mitigation plan. From this screen one can observe how many controls with which the resources being monitored are compliant can be observed by clicking on the chart icon on the top right corner.

The screen can be used to observe the error description of each violation that is observed in a resource, its impact as well as mitigation plan also.

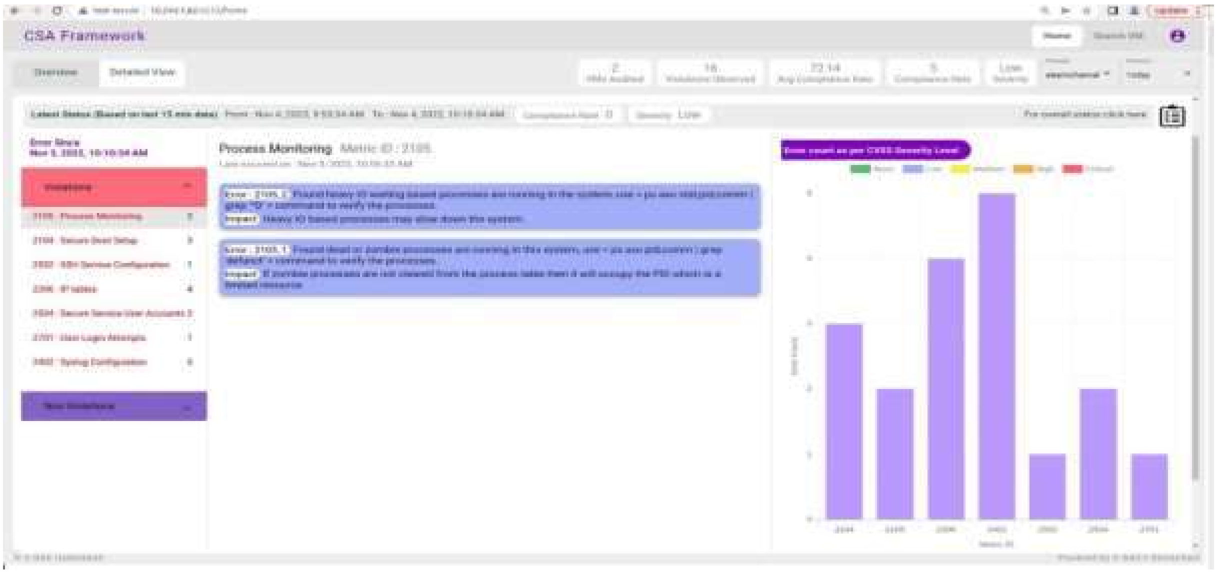


Fig. 4. Observed e/vents and corresponding event details

A. Use Case Scenario

The use case scenario will give a clear picture of how the organizations can get benefitted using this software. The software can be used in a multi-institute or multi-center-project environment wherein different stakeholders from different geographical locations work in a collaborative manner. Multi-institutional or multi-center-project-based resources can be monitored for security vulnerability based on a pre-defined security policy framework and maintain relevant data in a Blockchain network, hosted by relevant stakeholders, spread across different geographical locations for data analytics. Since the data is maintained in Blockchain all relevant stakeholders are ensured of the uniqueness and correctness of data at their premises and use such data for security analysis purposes in an independent manner. Hence, Institutions working in a multi-center, multi-project, or collaborative manner with other institutes and having the requirement of 24x7 monitoring of critical resources for security breaching based on a pre-determined security policy framework that applies universally among the participating entities can use this model or system. One such example scenario could be the Cyber Insurance domain where the cyber insurer requires a mechanism wherein monitoring of resources can be done in a seamless manner even in an untrusted network environment. Another such example could be where a consortium project is being executed by multiple stakeholders in a consortium manner while utilizing resources located at different places under different network administration teams.

The functional requirements of deployment in this case is as follows.

- ✓ Blockchain Network: As far as this use case is considered Blockchain network represents Hyperledger Fabric-based Blockchain network architecture. In this architecture, the network administrator can define a consortium of logical organizations that would like to share data among members of respective organizations based on pre-defined signing policy and membership.

- ✓ Logical Organization: An organization from the perspective of a Blockchain network is one that binds peer nodes and users through a common membership. We can use this concept to create several logical organizations that group peer nodes and users, belonging to different physical institutes/centres/project groups. Once the logical organizations are defined then data received from Blockchain client applications can be shared among the participants of those organizations only.

- ✓ Channel: In order to share the data between the logical organization's peers of respective organizations have to join the specific channel which is nothing but a logical subnetwork that supports data exchange among its members only. Hence, a channel has to be created so that peers of different organizations can join the channel.

- ✓ Critical Resource: Each critical resource that has a public and private key certificate pair and is capable of sending its security log information in a specific data format through a secured transmission medium has to be identified. The critical resource then sends the data in the desired format to the blockchain client application in the respective centre/project group/institute through an agent program. The Blockchain client application then submits the data to the Blockchain network.

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

- ✓ Blockchain Client Node: A Blockchain client application accepts the data in the specified format from a critical resource and validates its authenticity and integrity based on the resource's certificates. The client application then submits the same as a blockchain transaction to the Blockchain network along with information like channel configuration, user details, and request parameters. The data to be inserted will be forwarded to all the peer nodes that are part of the logical organizations that are participating in the Blockchain network as a consortium. The Blockchain client also receives requests from another entity other than critical resources i.e., the User application to fetch data from the underlying Blockchain network corresponding to a limited time period like the last 24 Hours, last 7 days, etc., in a specific format.

- ✓ Blockchain Peer Node: Each transaction that is submitted by different client applications will be received by one of the peer nodes in each logical organization that is being administered by the Blockchain network. Upon successful processing of transaction data through a Smart Contract program deployed in each peer node, the data will eventually be added to the existing peer ledger in all the Blockchain nodes. Each peer will run the same version of the smart contract. At the same time, each peer can also run multiple smart contracts each bearing a specific version number. Each center/institute/project group can designate one peer node as a member of the blockchain-based logical organization.

- ✓ User App: The user application can be accessed by authorized users with valid certificates representing respective stakeholders such as Cyber Insurer, and System/Network Administrators representing logical organizations of the Blockchain network. Upon receiving the request for the resource's current security status from the end users through the user app, the Blockchain client application issues a query request to the underlying Blockchain platform to fetch the relevant information and then runs a machine learning model or data analytics program to ascertain the overall threat perception level and security compliance rate. Once the threat perception has been computed, the information will be sent as a response to the user's request. The authorized users then can view information like the number of critical resources that are being monitored across different institutes/project groups/centers, the average and overall compliance rate of specific resources being monitored, detailed information of captured events, their impact, and mitigation plan, threat perception level in accordance with CVSS score. Only registered Users Applications can communicate with the Blockchain client node.

- ✓ Ordering Service Nodes: The user application initially sends data to endorsing nodes (Blockchain Peer Nodes) in each organization and collects the read/write set upon executing a Smart Contract on endorsing node. The data will not be committed at this point in time. The User application after collecting read/write set information from each peer will send the same to the ordering service nodes for ordering transaction data and bundle all transactions in a Block in accordance with the pre-defined block size limit. Ordering service nodes after the creation of blocks will broadcast them to all the Blockchain peer nodes in each organization which commits the same upon successful revalidation of transaction data to observe whether any discrepancies are there in the transaction data from the time the read/write sets are collected earlier and before committing them in the same peer node. Blocks will then commit the data either as valid or invalid depending on the validation results. The Ordering Service nodes can be deployed at each participating Institute/Project Group/Centre or one of them can host the ordering service nodes. The same ordering service nodes can be shared and used for multiple project groups/institutes/centers.

In Fig. 5., it is envisaged that two Institutes A and B have joined hands to execute a collaborative project X whose resources are located in both A & B. The architecture is not limited to two only but can be used for many collaborating agencies. Both A & B would like to monitor the security compliance report or threat perception level of all the resources. Based on the above-depicted deployment architecture now it is possible for A and B to run their respective Blockchain client applications to receive data from their respective institute's critical resources and forward them to the Blockchain network.

The Blockchain network in this architecture has two logical organizations namely Org A and Org B which binds one peer node each (more peers can be run to ensure high availability if needed) along with users. A blockchain channel also has been created and made all peers of respective logical organizations members of the same for sharing data among the peers of those organizations only.

Once the Blockchain network receives data from different client applications they forward them to all the peers of both Org A and Org B through Ordering Service nodes and the data will be stored in their respective peer ledgers permanently.

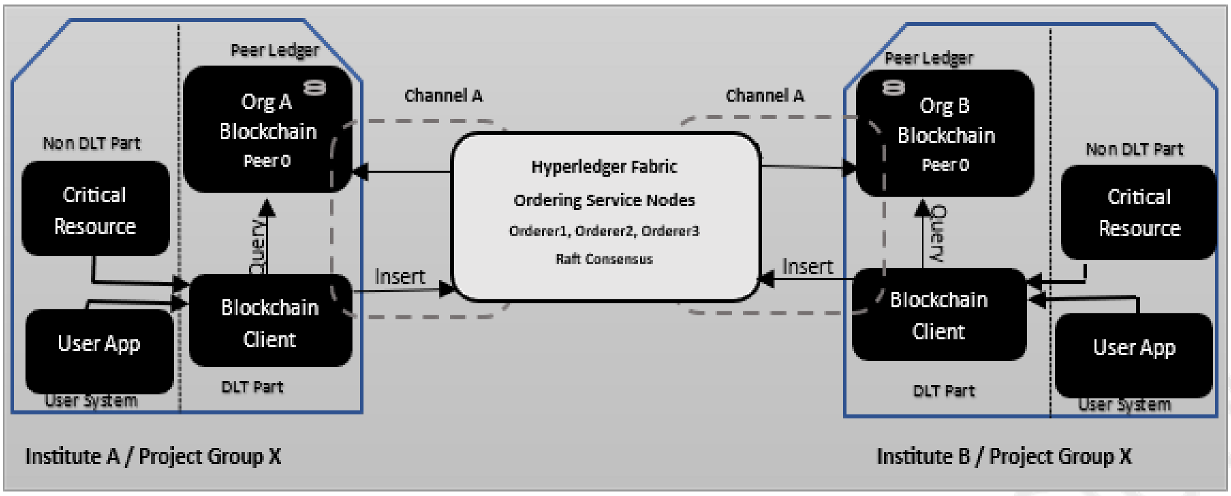


Fig. 5. Security Assessment Blockchain Architecture

The User Apps running at institutes A & B can have a unique representation of the data pertaining to Project X at any given point of time and can run their own analytics programs to infer security compliance information and threat perception level independently. Thus, this way the architecture supports a decentralized network with data replication capabilities while ensuring each participant relies upon a unique representation of data and facilitates each user app to run data analytics independently.

The architecture can support multiple consortiums with multiple channels to delineate the data and data analytics into separate groups and monitor independently.

B. Security and Identity Management

Blockchain Peer and Orderer Node: A peer and orderer node can join the Blockchain network by establishing its membership within an organizational context and generating a public and private key for the peer and orderer nodes.

User Registration: Users with 'READER' and 'WRITER' roles can be registered with the system. READERS are allowed for querying the data and WRITERS are allowed for inserting data into the Blockchain network. User accounts with the WRITER role are used during critical resource registration time and the READER role is used for those user accounts which are meant for accessing the User App components.

Non-DLT System & User System: These are the components which are not related to Blockchain network but are nothing but data producers or consumers. The data source's (or critical resource being monitored) integrity protection is ensured by way of computing the hash of all the fields in the data source object and encrypting it using the data source's private key. The name and password used during key pair generation must be a registered user with the role of 'WRITER'. Since the data source's public key is shared with the Blockchain client (BCTClient), the client can verify the authenticity and integrity of the data once it receives the same and it can insert data into the Blockchain network.

User App: This component's IP address MUST be registered with the BCTClient in order to accept requests from only registered applications. Apart from that the request object coming from User App MUST also provide a username and password for authentication purposes. When the request is from a registered User App and the user is a registered user of the Blockchain network then only the request will be processed further.

Privacy: Blockchain peers, clients, and orderers of different organizations can exchange data among themselves provided they all join the same Blockchain 'channel'. Members of different channels cannot exchange information between themselves. This way by binding peers, clients, and orderers of different organizations that needs to exchange data among themselves only data privacy is achieved.

Event API: The user app that is installed at the respective stakeholder's premises can invoke an API call for retrieving data from the blockchain system periodically to compute overall and individual threat perception levels.

C. Performance of the tool

The software module i.e., the intermediate gateway in our overall architecture, used for retrieval of data from the Blockchain network and to verify security policy conformance, should be designed in such a manner that it can withstand the vast amount of data that is required to be processed and improve the system response time.

It is observed that if logs are collected at 10min intervals in a day, a total of 144 iterations will be required to store the log information in the Blockchain platform. We observed that with 14 control data that are collected 144 times a day and each JSON object size of 594bytes on average, approximately a total of 1.14MB of data would be stored from each resource that is being monitored. For 30 days 34.27MB of data will have to be processed corresponding to a particular resource. After observing this, we designed the below data structure to cope up with data generated by multiple resources. In our experiment, we used nearly 8VMs for monitoring purposes and improved the system response time. The end-user, upon selecting the duration for which the analysis has to be performed in the dashboard, data pertaining to that period will be retrieved from the Blockchain network and maintained in accordance with the below data structure.

```
{
        metric_id: val    # id of the control name
        result: [error1, error2, error3, error4 ….] #stores id
        of every error that is observed
        timestamp: [ts1, ts2, ts3 ….]  #stores timestamps at
        which the errors have been observed
        machine_id: xxxx    #id of the resource
        hostname: val   #name of the host
}
```

By default, the data collection interval is set to 10min in our approach. In this format instead of processing individual JSON objects representing each log entry, we rearranged similar entries based on the time of their occurrence while responding to the request from the dashboard component. This has enhanced the system response time considerably.

Since the data is collected from different resources for a specific period it makes the data analysis more effective as events can be related and inferences can be drawn accordingly. We used the pseudo-code explained above for computing threat perception scores by related various events that occurred during the specified period.

D. Storage Management

It is important to understand how the storage space is utilized in each peer node in the Blockchain network. Whenever a peer node confirms a transaction block that it has received from the ordering service node in the Hyperledger network, the peer node stores the same in its ledger. The ordering service ensures an atomic broadcast of the blocks once they are ordered as explained above. Usually, when a peer node in the Blockchain network goes down for any reason and when it rejoins the same network, it will try to contact the other peer nodes to obtain the missed entries in its peer ledger to be in sync with the blockchain network.

Here, the question arises whether we should ensure that all peer nodes must be homogeneous in terms of their storage capabilities. What happens if we use nodes in different logical organizations with different storage spaces?

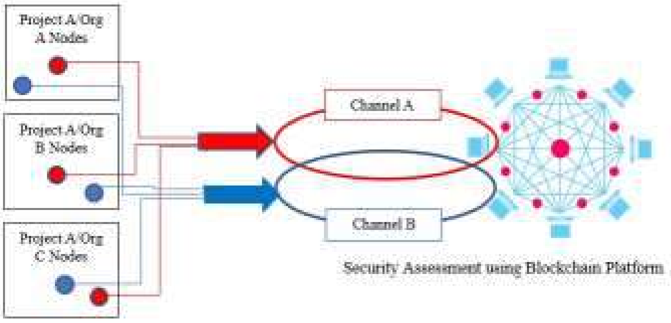


Fig. 6. Schematic diagram of transition from one channel to another channel to cope with available storage space in blockchain nodes

It is explained that from a single resource, approximately 1.14MB of log data has to be maintained in peer ledgers of Blockchain nodes in the network. Apart from that the same data will get replicated in other nodes of the Blockchain network. When multiple resources are monitored those many nodes data has to be stored in the Blockchain node.

When the storage space gets filled in the respective blockchain nodes it is not possible to replace the one which got filled first compared to other nodes. This is because if we replace a node that gets filled first with another node the newly added node will start trying to be sync with other blockchain nodes which are members of the same channel. This situation necessitates that all nodes of Blockchain nodes must be homogenous. This is a serious limitation as it will restrict blockchain usage in a multi-institute or multi-center-project environment where ensuring homogenous nodes is a difficult phenomenon.

We studied the impact of such a requirement and its practical applicability. In general, no production system or network can be brought down temporarily and hence we need to devise an alternative mechanism.

In this regard, we conducted a trial experiment in our lab following the below steps which resulted in a smooth transition of the existing production-grade blockchain network that could handle the above storage management-related issue and also eliminated the requirement of strictly using homogenous nodes across all the organizations.

As shown in Fig. 6., we maintained a single peer node in each logical organization of the Blockchain network and made the signing policy an IMPLICIT MAJORITY so that members of any of the two organizations, when they sign the transaction data, would be sufficient. We also made another node in each logical organization ready with new storage space available and made them members of a new 'channel' that binds only these newly added peers of respective organizations as members. Now we have two channels (please refer to what purpose a channel serves above), one which was there from the beginning and is being used currently. The newly created channel contains new peers of respective organizations. Once this setup is made (for which we need not bring down the production network), we deployed the same smart contract that is used for validating transaction data (the log data) in the existing channel on to the new channel also. Once the arrangements are completed, we made all subsequent requests from client applications to divert their future request submission to the new channel. Since new channel and smart contract combination is made available on newly added peers, they maintain their own peer ledgers on respective nodes. This way we handled the smooth transition of the production network from one channel to another channel with the same smart contract. This solution not only helped in dealing with storage management effectively but also eliminated the need for worrying about having homogeneous nodes only across the Blockchain network.

E. Comparative Analysis

Table II below depicts the comparative analysis of developed solutions with different security analysis tools. Other tools given in Table II have certain shortcomings when compared to our solution as both of them generate a report in the system in which those tools are deployed. We used freeware versions. Whereas our solution provides a web interface to view the health report of all resources being monitored in one place. Moreover, we also have the CVSS score that displays the severity level to alert the concerned stakeholders. Another important feature of the developed solution is that it automatically collects data at regular intervals and sends them to the Blockchain platform for permanent storage in a tamper-evident manner. This way, we have designed and implemented a unified security policy framework with a 24x7 monitoring and alert system. The system facilitates a unique state of representation of security log information due to the usage of Blockchain because of which inferences can be drawn beyond any doubts on data integrity.

TABLE II
COMPARATIVE ANALYSIS OF DEVELOPED SOLUTION WITH OTHER SIMILAR INITIATIVES

Controls	Lynis	CIS-CAT	Our Solution
File Integrity	No (Depends on auditd)	No (depends on auditd)	Yes (Own Implementation)
Continuous monitoring	No	No	Yes
Reports	Yes (On VM where it runs)	Yes (On VM where it runs)	Sends information to Blockchain
Impact Details	Partial	Yes	Yes
Firewall rules for all open ports	No	Yes	Yes
File System Base control	Yes (Doesn't say whether it is right?)	Yes (Say's whether it is right)	Yes (Say's whether it is right)
No login shell for system users/network services	No	Yes	Yes
Presentation	Text	Yes (html)	Yes (html)

V. CONCLUSIONS

We presented a unified and comprehensive security assessment framework that is supported by Blockchain Technology. The security policy framework was developed based on inputs drawn from key takeaways from top threats in the cloud platform, existing cloud best practices/standards such as Cloud Security Alliance, and implementation techniques to provide an early detection mechanism using CIS Benchmarks etc. In this attempt, we demonstrated how Blockchain technology supported the data provenance related to security aspects in the decentralized network and log management in a time-stamped and tamper-evident manner based on consensus among the relevant stakeholders. We also demonstrated an inbuilt mechanism for monitoring the effectiveness of the security policy framework by way of computing severity levels in various resources. During the course of using the Blockchain network the role of blockchain nodes, and challenges w.r.t storage management have been studied carefully and the system has been designed accordingly to provide better performance.

ACKNOWLEDGMENT

The author acknowledges all their colleagues who supported them in coming out with the results and implementation of the software and MeitY, Government of India.

REFERENCES

[1] G Ramachandra et., "A Comprehensive Survey on Security in Cloud Computing", *Procedia Comput. Sci.*, vol. 110, no. 2012, pp. 465–472, 2017.

[2] Monjur Ahmed, Mohammad Ashraf Hussain, "Cloud computing and security issues in the cloud", *International Journal of Network Security and Applications.*, Vol.6(1), 2014, pp. 25-36

[3] Z. Zheng, S. Xie, H. Dai, X. Chen and H. Wang. (2017) "An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends," *IEEE International Congress on Big Data (BigData Congress),* Honolulu, HI, pp. 557-564.

[4] Technical report from Cloud Security Alliance, "*Top threats to cloud computing: Deep Dive – A case study analysis for 'The Treacherous 12: Top Threats to cloud computing*' and a relative security industry breach analysis.

[5] *Cloud Controls Matrix* from Cloud Security Alliance. Available: https://cloudsecurityalliance.org/research/cloud-controls-matrix/ on July 16, 2020

[6] *CloudWatch User Guide for monitoring Amazon instances health information.* Available: https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/WhatIsCloudWatch.html

[7] *Lynis Documentation* Available: https://cisofy.com/documentation/lynis/

[8] *CIS-CAT Lite tool* Available: https://www.cisecurity.org/blog/introducing-cis-cat-lite/

[9] X. Liang, S. Shetty, D. Tosh, C. Kamhoua, K. Kwiat and L. Njilla, "ProvChain: A Blockchain-Based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability," 2017 *17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, Madrid, 2017, pp. 468-477, doi: 10.1109/CCGRID.2017.8.

[10] *CIS Benchmarks for Ubuntu Linux* Available: https://www.cisecurity.org/cis-benchmarks/

[11] *Hyperledger Architecture, Volume 1 (2017) Introduction to Hyperledger Business Blockchain Design Philosophy and Consensus*. Available: https://www.hyperledger.org/wp-content/uploads/2017/08/Hyperledger_Arch_WG_Paper_1_Consensus.pdf

[12] *Common Vulnerability Scoring System v3.1: Specification Document* Available: https://www.first.org/cvss/v3.1/specification-document

[13] D. Ongaro, J. Ousterhout. (2014) "In search of an understandable consensus algorithm", *Proc. USENIX Conf. USENIX Annu. Tech. Conf. (USENIX ATC)*, pp. 305-320.

[14] C. Yang, L. Tan, N. Shi, B. Xu, Y. Cao and K. Yu, "AuthPrivacyChain: A Blockchain-Based Access Control Framework With Privacy Protection in Cloud," in *IEEE Access*, vol. 8, pp. 70604-70615, 2020, doi: 10.1109/ACCESS.2020.2985762.

[15] Park JH, Park JH. Blockchain Security in Cloud Computing: Use Cases, Challenges, and Solutions. *Symmetry*. 2017; 9(8):164. https://doi.org/10.3390/sym9080164

[16] S. Singh, Y.-S. Jeong, and J. H. Park, "A survey on cloud computing security: Issues, threats, and solutions," *J. Network and. Computer. Applications*., vol. 75, pp. 200–222, Nov. 2016.

[17] B. Duncan, D. J. Pym and M. Whittington, "Developing a Conceptual Framework for Cloud Security Assurance," *2013 IEEE 5th International Conference on Cloud Computing Technology and*

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

Science, Bristol, 2013, pp. 120-125, doi: 10.1109/CloudCom.2013.144.

[18] Sachin Shetty, Val Red, Charles Kamhoua, Kevin Kwiat and Laurent Njilla "Data provenance assurance in the cloud using Blockchain", *Proc SPIE 10206, Disruptive Technologies in Sensors and Sensor Systems*, 102060I (2 May 2017); Available: https://doi.org/10.1117/12.2266994

[19] *Channels — hyperledger-fabricdocs main documentation* Available: "https://hyperledger-fabric.readthedocs.io/en/release-2.2/channels.html"

[20] *Transaction Flow — hyperledger-fabricdocs* main documentation. Available: "https://hyperledger-fabric.readthedocs.io/en/release-2.2/channels.html"

[21] *Ledger — hyperledger-fabricdocs main documentation* Available: "https://hyperledger-fabric.readthedocs.io/en/release-2.2/ledger/ledger.html"

N Satyanarayana lives in Hyderabad, India, and is born in 1976. He did his Master of Technology a post-graduation degree in computer science from Jawaharlal Nehru Technological University, Hyderabad, India. Prior to this, he completed his Master of Computer Application post-graduate degree in computer applications from Sri Venkateswara University, Tirupati, India in the year 1999.

He is currently working as Joint Director in the Centre for Development of Advanced Computing (CDAC), Hyderabad, India. He has been working for CDAC for the past twenty years and played an instrumental role in developing various applications in the areas like eLearning, Peer to Peer, Network Management, and Blockchain Technology.

Mr. N Satyanarayana published several papers at National and International conferences in various areas such as Peer to Peer Computing, Network Management, e-Learning, and Blockchain. His current research interest includes Blockchain consensus algorithms and reference architectures. He is also contributing towards best practices/standards in the respective fields of his work being a member of the technical committees of the Bureau of Indian Standards.

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

A Horizontal Federated Learning Approach to IoT Malware Traffic Detection: An Empirical Evaluation with N-BaIoT Dataset

Phuc Hao Do[*,***], Tran Duc Le[**], Vladimir Vishnevsky[****], Aleksandr Berezkin[*], Ruslan Kirichek[*, ****]

[*]*The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Saint-Petersburg, Russia*

[**]*University of Science and Technology – The University of Danang, Da Nang, Viet Nam*

[***]*Danang Architecture University, Da Nang, Viet Nam*

[****] *V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia*

haodp@dau.edu.vn, letranduc@dut.udn.vn, vishn@inbox.ru, berezkin.aa@sut.ru, kirichek@sut.ru

***Abstract* —The increasing prevalence of botnet attacks in IoT networks has led to the development of deep learning techniques for their detection. However, conventional centralized deep learning models pose challenges in simultaneously ensuring user data privacy and detecting botnet attacks. To address this issue, this study evaluates the efficacy of Federated Learning (FL) in detecting IoT malware traffic while preserving user privacy. The study employs N-BaIoT, a dataset of real-world IoT network traffic infected by malware, and compares the effectiveness of FL models using Convolutional Neural Network, Long Short-Term Memory, and Gated Recurrent Unit models with a centralized approach. The results indicate that FL can achieve high performance in detecting abnormal traffic in IoT networks, with the CNN model yielding the best results among the three models evaluated. The study recommends the use of FL for IoT malware traffic detection due to its ability to preserve data privacy.**

***Keyword* — IoT, abnormal traffics, malware detection, federated learning, AI model**

I. INTRODUCTION

As a result of the rapid growth of the Internet of Things (IoT) technology, IoT devices have become an integral part of people's daily life. However, it inevitably introduces some network security challenges [1][2]. IoT devices are easy targets for malicious attacks such as malware attacks due to their heterogeneity and vulnerability. The prevalence of privacy and security concerns is rising due to the continual expansion of IoT devices and the resulting exposure of more and more private data online. It is essential to monitor IoT networks to prevent malicious cyberattacks on IoT devices [3]. By studying the traffic of IoT devices, network intrusion detection in the IoT ecosystem may be enhanced, and cyberspace security can be guaranteed [4][5][6].

In recent years, the expansion of deep learning has played a significant role in advancing IoT intrusion detection research [7]. Deep neural networks are employed to automatically identify dataset features, thereby reducing the feature engineering workload and improving data processing performance without requiring human intervention. Despite its benefits, the application of the Internet anomaly detection approach to the IoT is not straightforward due to the large amount of data required. Most existing machine learning or artificial intelligence solutions rely on a single server to collect data from various IoT devices and build global models [8]. However, this approach is not always effective, particularly when device actions involve sensitive or private information that could severely impact environmental security and privacy if disclosed to unauthorized parties in IoT networks.

In the context of preserving information privacy and integrity, Federated Learning (FL) [9] has gained increasing relevance. FL involves decentralizing the training model among several nodes or clients that use local data. Each decentralized node trains a distinct model on its data and distributes the model parameters (not the private data) to others through a central entity known as a server or via a peer-to-peer methodology [10]. The model parameters are then aggregated to produce a singular and global model. After several iterations, each client obtains a global model by

Manuscript received on Jan. 19, 2023. This work was supported by the Russian Science Foundation within of scientific project No. 22-49-02023 and it is also supported by a follow-up of the invited journal the accepted & presented paper entitled "A Horizontal Federated-Learning Model for Detecting Abnormal Traffic Generated by Malware in IoT Networks" of the 25th International Conference on Advanced Communication Technology (ICACT2023)

P. H. Do, he is a Ph.D student in the Department of Communication Networks and Data Transmission at Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Russia. (email: haodp@dau.edu.vn).

T. D. Le, he is a professor in Information Technology Faculty, The University of Danang - University of Science and Technology, Danang, Vietnam (email: letranduc@dut.udn.vn).

V. Vishnevsky, he is a professor at V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Russia. (email: vishn@inbox.ru).

Aleksandr Berezkin, he is working at the Bonch Bruevich Saint Petersburg State University of Telecommunications as the Associate Professor of Department of Programming Engineering and Computer Science. (email: berezkin.aa@sut.ru).

Dr. Sc. Ruslan Kirichek, he is a professor in the Department of Communication Networks and Data Transmission at Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Russia. (corresponding author, phone: +7-921-970-0160, email: kirichek@sut.ru).

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

aggregating their unique models. This strategy naturally supports data privacy because data is not shared with external identities.

In the paper, we aim to build a horizontal Federated-Learning model to detect abnormal traffics, specifically DDoS attack traffic generated by malware, such as Botnet in IoT networks. To achieve this, we plan to utilize popular models like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU).

In this study, we utilize the N-BaIoT [11] dataset for model training due to its real-world origin, as opposed to other datasets that simulate network traffic. Our primary objective is to evaluate the efficacy of Federated Learning (FL) in comparison to traditional centralized artificial intelligence models. To achieve this goal, we address research questions such as whether FL can accurately detect malware or malicious traffic in IoT networks and which training models are suitable when applied to this decentralized approach.

Hence, the primary contributions of this study are as follows:

- Employing feature selection and feature engineering techniques to preprocess the N-BaIoT dataset and select features that have the most considerable impact on the model's accuracy and performance;
- Deploying and evaluating various deep learning models, including CNN, LSTM, and GRU, in the FL approach with different dataset portions to determine their compatibility and select the most suitable model for the FL approach;
- Conducting a comparative analysis between the FL model and traditional deep learning models using the same dataset, thus assessing the feasibility of FL in detecting abnormal traffic in IoT networks.

The results of this study will be included in the Draft Recommendation ITU-T Q.TSRT_IoT “Test specifications for remote testing of Internet of Things using the probes”.

II. RELATED WORKS

Several researchers have conducted a thorough investigation to find a more trustworthy anomaly detection strategy for IoT networks. To solve the issue of constrained device resources, the study in [12] created a unique, lightweight attack detection method that uses the Support Vector Machines (SVM) algorithm to recognize hazardous assaults in the IoT space. To offer a distributed intrusion detection system, Ferdowsi et al. recommended anomaly detection utilizing distributed Generative Adversarial Networks (GANs) [13]. Each device may protect user privacy while applying a detection model to its data since it is not maintained centrally. A centralized IoT intrusion detection system based on fog computing was suggested in the publication [14]. Two cascading recurrent neural networks (RNNs) are used in the design, each with its hyperparameters and attack-specific tuning. The system administrator is notified if any of the two RNNs determines that an input instance is malicious. Time-series data from IIoT sensors were utilized by the authors of [15] to identify abnormalities using an attention-based convolutional neural network with LSTM [16].

In recent years, FL has risen to prominence in cybersecurity. This IoT security paradigm has already been used in several works. In this context, the study proposal [17] underlined the problem with traditional AI-based solutions' lack of data privacy. However, a private dataset was used for the examination. Despite sharing similar objectives, the research mentioned in [18, 19] mainly focused on industrial IoT devices. They looked at application samples and sensor readings rather than network data.

The authors introduced the Federated Averaging (FedAVG) approach in [20], which is a strong basis for many FL-based investigations. In this strategy, a central server coordinates the training of a global model across several clients using their datasets. The server is responsible for averaging the client-sent models' parameters and sending the resulting global model back to the clients. Until a termination condition is met, this process is repeated. FL has significantly advanced since several researchers [21] examined the most current advancements in this area.

Two more robust model aggregation functions were suggested in [22]. They are based explicitly on the coordinate-wise mean and median of the clients' models supplied to the server. The authors of [23] suggested resampling as a way to lessen the variability in the distribution of the models that the clients supplied. It should be used before employing a powerful aggregating function. When used with models trained on non-IID datasets, it seeks to lessen any negative consequences that such a function could have.

The paper [24] focuses on the challenging task of optimizing neural architectures in the federated learning framework. The authors describe recent work on federated neural architecture search, which involves searching for optimal neural architectures across multiple clients in a distributed manner. They categorize these approaches into online and offline implementations, as well as single- and multi-objective search approaches. They also explain the different types of federated learning, including horizontal, vertical, and hybrid.

The authors [25] propose a Federated Learning (FL) based IoT Traffic Classifier (FLITC) that uses Multi-Layer Perception (MLP) neural networks to classify traffic data while keeping local data on IoT devices. FLITC sends only the learned parameters to the aggregation server, reducing communication costs and latency. The main idea behind FLITC is to train a shared model on distributed devices without exchanging raw data, thereby preserving privacy and reducing communication overhead. The method is expected to improve the accuracy and efficiency of IoT traffic classification while maintaining data privacy.

The study [26] proposes a federated-learning traffic classification protocol (FLIC) for Internet traffic classification without compromising user privacy. FLIC aims to achieve accuracy comparable to centralized deep learning for Internet application identification without privacy leakage. It can classify new applications on-the-fly when a participant joins the learning process with a new application, which has not been done in previous works. The authors implemented the FLIC prototype using TensorFlow, allowing clients to gather packets, perform on-device training, and exchange training results with the FLIC server. They demonstrated that federated learning-based packet classification achieves an accuracy of 88% under non-independent and identically distributed (non-IID) traffic across clients. When a new application was added dynamically as a client participating in the learning process, an accuracy of 92% was achieved.

The study [27] focuses on the security challenges posed by the deployment of IoT devices and proposes a federated learning-based edge device identification (FedeEDI) method to control access and manage internal devices. The authors note that external attackers often exploit vulnerable IoT devices to gain access to the target's internal network and cause security threats. They review existing literature on deep learning-based algorithms for edge device identification and highlight the limitations of centralized learning-based EDI (CentEDI) methods that train all data together. The authors propose a federated learning-based approach that addresses data security concerns and is suitable for deployment on edge devices.

Overall, the studies presented in this section highlight the potential of federated learning as a viable solution to the challenges posed by the deployment of IoT devices.

III. THE HORIZONTAL FEDERATED LEARNING

A. Horizontal FL

Federated learning can be divided into three categories: horizontal federated learning, vertical federated learning, and federated transfer learning [28].

Horizontal federated learning is designed for scenarios where participating clients' datasets share the same feature space but contain different samples. The term "horizontal" originates from instance-distributed learning, as illustrated in Fig. 1a, where datasets are horizontally partitioned across data samples and allocated to clients. In federated learning, data can be considered horizontally partitioned when different clients generate data with the same attributes (features) but distinct samples. Similarly, as indicated by the part surrounded by the two dashed lines in Fig. 1b, the data can be considered horizontally partitioned in federated learning when different data are generated on different clients that have the same attributes (features). For example, two hospitals in different regions may have distinct patients but perform the same tests and collect the same personal information (e.g., name, age, gender, and address). There are three main differences between instance-distributed learning and horizontal federated learning [29]:

- Data are typically independent and identically distributed (IID) in distributed learning but may be non-IID in horizontal federated learning. Designers can manually allocate subsets of client data to be IID in distributed learning to enhance convergence, while in horizontal federated learning, the central server has no access to raw data, which is usually non-IID on different clients link.springer.com.
- Horizontal federated learning involves a large number of connected clients, whereas instance-distributed learning often does not have as many workers. Too many workers may worsen distributed training performance when the total data amount is fixed.
- Global model update mechanisms differ. In instance-distributed learning, a deep neural network synchronously updates the global model once local gradients of mini-batch data are calculated. This approach is not suitable for horizontal federated learning due to communication constraints.

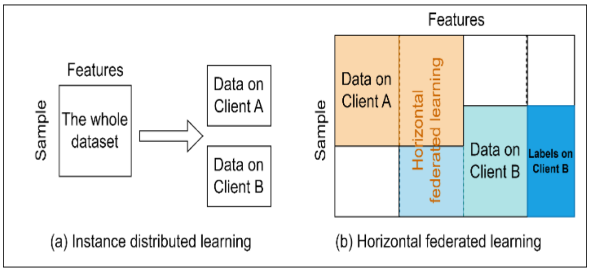


Fig. 1. (a) Instance distributed learning
(b) horizontal federated learning

Horizontal federated learning faces three main challenges compared to centralized learning: reducing communication resources, improving convergence speed, and ensuring no private information leakage.

In contrast, vertical federated learning is applicable when datasets share the same sample space but have different feature spaces. Vertical federated learning is similar to feature distributed learning, where the central server acts as a coordinator to compute the total loss instead of aggregating uploaded weights.

Federated transfer learning, on the other hand, leverages vertical federated learning with a pre-trained model trained on a similar dataset for solving a different problem.

Typical horizontal federated learning (Fig. 2) algorithms, such as the FedAvg, consist of the following main steps:

- Initialize the global model parameters on the server and download the global model to every participating (connected) client.
- Every connected client learns the downloaded global model on its own data for several training epochs. Once completed, the updated model parameters or gradients (gradients here mean the difference between the downloaded model and the updated model) would be sent to the server. Note that the clients may have different amounts of training data and unbalanced computational resources. As a result, the server is not able to receive the uploads from different clients at the same time.
- The server aggregates the received uploads (synchronously or asynchronously) to update the global model.
- Repeat the above two steps until convergence.

From the above steps, we can find that the central server can only receive model weights or gradients of the participating clients and has no access to any local raw data. Therefore, users' privacy is immensely protected in horizontal federated learning.

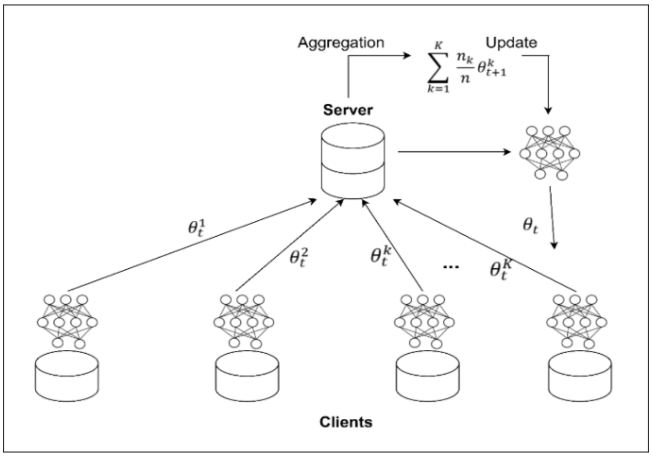


Fig. 2. Flowchart of federated learning. θ is the global model parameters, n_k is the data size of client k, K is the total number of clients and t is the communication round in federated learning. We initialize global model parameters randomly at the beginning of the communication round and use updated model parameters afterward

In this study, the use of this model is suitable the fact that we use a single dataset, which is N-BaIoT. Thus the features are the same.

The Fig 3 displays the approach used in this study, which consists of the dataset, data processing, data aggregation, divide "attribute class", and classifier. Standardization and minimum-maximum normalization min-max normalization) are also used as data preparation techniques.

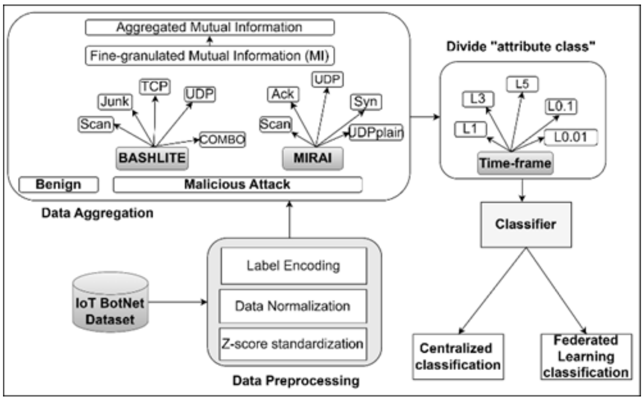


Fig. 3. The research flow

B. Dataset

There are many public datasets related to IoT security as follows:

- N-BaIoT [1]: This dataset tackles the shortage of available botnet datasets, particularly for the Internet of Things. It implies actual traffic data, collected from nine commercial IoT devices that were actually infected with Bashlite and Mirai malware.
- MedBIoT [2]: It is collected from actual and simulated Internet of Things devices in a medium-sized network (i.e., 83 devices). The data collection is divided into categories based on the kind of traffic (malware or normal traffic), making it simple to label the data and extract characteristics from the raw pcap files.
- Bot-IoT [3]: The BoT-IoT dataset was developed using a realistic network environment constructed at the Cyber Range Lab of the UNSW Canberra Cyber Center. The dataset consists of Service Scan, DDoS, DoS, Keylogging, and Data exfiltration attacks. In addition, DDoS and DoS attacks are categorized based on the protocol employed.
- TON-IoT [4]: This dataset contains heterogeneous data sources gathered from Telemetry datasets of IoT and IIoT sensors, Windows OS, Ubuntu, and network traffic datasets. The datasets were acquired from the Cyber Range and IoT Labs' realistic and wide network.
- IoT-23 [5]: IoT-23 contains 20 malware captures (scenarios) conducted on IoT devices, and 3 benign IoT traffic captures. The authors ran a particular malware sample on a Raspberry Pi for each malicious scenario, which utilized multiple protocols and carried out distinct tasks.

The N-BaIoT dataset is collected separately for each device, so it is very suitable to implement the FL empirical model. Therefore, we will use it for our research. IoT devices in the N-BaIoT dataset were targeted by the Bashlite and Mirai botnet attack families. Each file has 115 features as well as a class label. The dataset has also been built to support binary and multi-class classification. The target class labels are: "TCP attack," "benign" for detection, and "Mirai" or "Bashlite" attack types for multi-class classification.

The N-BaIoT dataset comprises feature headers that describe various aspects of network traffic data, which can be grouped into categories based on their purpose and scope. The headers related to stream aggregation include H, which provides statistics summarizing recent traffic from the packet's host (IP); HH, which summarizes recent traffic from the packet's host to the packet's destination host; HpHp, which summarizes recent traffic from the packet's host+port (IP) to the packet's destination host+port; and HH_jit, which summarizes the jitter of the traffic from the packet's host to the packet's destination host. Additionally, the time-frame decay factor Lambda determines how much recent history of the stream is captured in these statistics, with values such as L5, L3, L1, and others.

The dataset also provides statistics extracted from the packet stream, including the weight of the stream, which can be viewed as the number of items observed in recent history. The mean represents the average value of the stream, while the standard deviation is denoted by std. The root squared sum of the two streams' variances is represented by radius, while the root squared sum of the two streams' means is denoted by magnitude. Moreover, an approximated covariance between two streams is represented by cov, and an approximated covariance between two streams is denoted by *pcc*.

Overall, the N-BaIoT dataset provides a comprehensive set of feature headers that can be used to analyze various aspects of network traffic data, which can be useful for developing and evaluating intrusion detection systems in IoT networks.

In this study, we will focus on multi-classification. We use part of the dataset to proceed with the experiment. The multi-class classification will be evaluated.

[1] https://www.kaggle.com/datasets/mkashifn/nbaiot-dataset
[2] https://cs.taltech.ee/research/data/medbiot/
[3] https://ieee-dataport.org/documents/bot-iot-dataset
[4] https://research.unsw.edu.au/projects/toniot-datasets
[5] https://www.stratosphereips.org/datasets-iot23

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

C. Preprocessing

Although data preparation [30] is difficult and time-consuming [31], its importance has been demonstrated for speeding the training process and enhancing its effectiveness. Consequently, this study employs label encoding, min–max normalization, and standardization as pre-preprocessing procedures.

1) Label Encoding

The class label has 11 different category values (one "Benign" class and ten subclasses of attack type). As a result, before these attributes are used with the models, they are converted into numerical values. There are several methods for converting categorical values, including one-hot encoding [32], ordinal encoding [33], similarity encoding [34], entity embedding [35], and multi-hot encoding. Among them, one-hot encoding and ordinal encoding are the most widely employed. This study uses the one-hot encoding technique for encoding categorical values.

2) Normalization and Standardization

Suppose the columns in a dataset contain values with varying ranges. In that case, the performance of both regression and classification models is negatively impacted. Mahfouz et al. in [36] demonstrated how this issue causes the performance of the models to decrease when uneven scales of features are observed in a dataset. It is required to determine the acceptable range for the insignificant and dominating values to address such problems. The two most used methods are min-max normalizing and z-score standardization:

- The following equation is used to apply min-max normalization to change the values of the dataset's feature values into the range [0, 1]:

$$X_{normalized} = \frac{X - X_{min\_value}}{X_{max\_value} - X_{min\_value}} \quad (1)$$

where $X_{normalized}$ represents the normalized value, X_{min_value} and X_{max_value} are the intended interval's boundary range, which is [0, 1], and X is the initial value that would be altered inside those ranges.

- Z-score standardization is used to rescale dataset features, reflecting the characteristics of a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

$$X_{normalized} = \frac{X - \mu}{\sigma} \quad (2)$$

3) Data aggregation

After encoding the target class, the "benign" class was added to the N-BaIoT dataset, which has 115 characteristics and 10 class labels. To improve the performance of the classification process, we can implement some feature selection methods. In the scope of the paper, we choose the Mutual Information (MI) method for the feature selection process of the input data.

An aggregated MI with multiple rank aggregation functions is developed and evaluated for the multi-class dataset. The concept of aggregated MI is explained as follows:

- Calculate the information gain score for each feature, f_i, in dataset D relative to class type $c \in C$. The features are then ranked based on the aggregator functions listed below. Calculate the information gain score for each feature, *fi*, in dataset D relative to class type $c \in C$. After that, the features are ranked based on the aggregation methods (Min, Max, Mean). Only part of preserved features are supplied to the classifiers, and the total performance is assessed.
- List of aggregators:
  - ○ *Min:* Chooses the class type c_i as the target class and takes the relevance score with the lowest value.
  - ○ *Max:* Chooses the class type c_i as the target class and takes the relevance score with the highest value.
  - ○ *Mean:* Chooses the class type c_i as the target class and takes the relevance score with the mean value.

4) Dividing attribute class

In this research study, we aim to evaluate the classification performance of the N-BaIoT dataset by dividing its attributes into different classes based on the time-frame property. The dataset includes time-frames such as L5, L3, L1, L0.1, and L0.01, which correspond to the traffic capture time. We will utilize these time frames to create attribute sets for each subclass. Specifically, Table I shows the attribute set for the time frame L5, which contains 23 attributes. We will follow the same procedure for the remaining time frames. This approach allows us to assess the classification performance of each class containing different attributes, providing insights into the effectiveness of the dataset's features for anomaly detection in IoT networks.

TABLE I
SOME PROPERTIES OF THE TIME-FRAME (L5)

No	Attribute	No	Attribute
1	MI_dir_L5_weight	13	HH_L5_pcc
2	MI_dir_L5_mean	14	HH_jit_L5_weight
3	MI_dir_L5_variance	15	HH_jit_L5_mean
4	H_L5_weight	16	HH_jit_L5_variance
5	H_L5_mean	17	HpHp_L5_weight
6	H_L5_variance	18	HpHp_L5_mean
7	HH_L5_weight	19	HpHp_L5_std
8	HH_L5_mean	20	HpHp_L5_magnitude
9	HH_L5_std	21	HpHp_L5_radius
10	HH_L5_magnitude	22	HpHp_L5_covariance
11	HH_L5_radius	23	HpHp_L5_pcc
12	HH_L5_covariance		

D. Deep Learning Techniques Applied to Abnormal Traffics Detection

1) Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN) [37]: CNNs were initially developed for image recognition tasks, but researchers have adapted them to process textual data in phishing detection with great success. CNNs utilize convolutional layers to automatically learn features and patterns from the email content by applying multiple filters to different regions of the input text. These filters can capture local patterns, such as word groupings or specific textual structures, which can indicate phishing attempts.

The formula for CNN classification varies depending on the specific architecture and design of the CNN model. However, a general formula for CNN classification can be broken down into the following steps:

- Convolutional Layers: The data input is passed through one or more convolutional layers. Each convolutional layer applies a set of filters to the data input, creating feature maps that capture different aspects of the data.
- Activation Function: After each convolutional layer, an activation function is applied to introduce non-linearity to the output feature maps. The most commonly used activation function is ReLU (Rectified Linear Unit).
- Pooling Layers: After the activation function, the feature maps are passed through one or more pooling layers. Pooling layers downsample the feature maps by taking the maximum or average value within a specific window size. This reduces the size of the feature maps and makes the model more computationally efficient.
- Flatten: After the final pooling layer, the feature maps are flattened into a one-dimensional vector.
- Fully Connected Layers: The flattened feature vector is then passed through one or more fully connected layers. Each fully connected layer applies a set of weights to the input vector and outputs a new vector of a specified size.
- Softmax: The final fully connected layer uses the softmax function to convert the output vector into a probability distribution over the different classes in the classification task. The class with the highest probability is then selected as the predicted class.

The above steps can be represented as a mathematical formula for a basic CNN classification model as follows:

$$y = sm(W_2 * relu\left(W_1 * pool\left(conv(x)\right) + b_1\right) + b_2) \quad (3)$$

where *x* is the data input, conv represents the convolutional layers, the pool represents the pooling layers, W_1 and b_1 represent the weights and biases of the first fully connected layer, RELU represents the activation function, W_2 and b_2 represent the weights and biases of the final fully connected layer, and *sm* is softmax function, softmax function represents the final activation function that outputs the probability distribution over the different classes.

2) Long Short-Term Memory (LSTM)

LSTM [38] is a recurrent neural network (RNN) that uses feedback to remember portions of the input and make predictions. RNNs are intended to process sequential input and have thus found widespread use in speech recognition and machine translation. The well-known problem of vanishing gradients affects the long-term memory of conventional RNNs. It restricts their ability to make predictions based on the most recent data in the sequence. LSTM (Fig. 4) overcomes the problem of vanishing gradients and can thus handle longer sequences (long-term memory). LSTM can extract context from a succession of features. Using a gate function, it can add or delete information from the hidden state vector, preserving vital information in the hidden layer vectors.

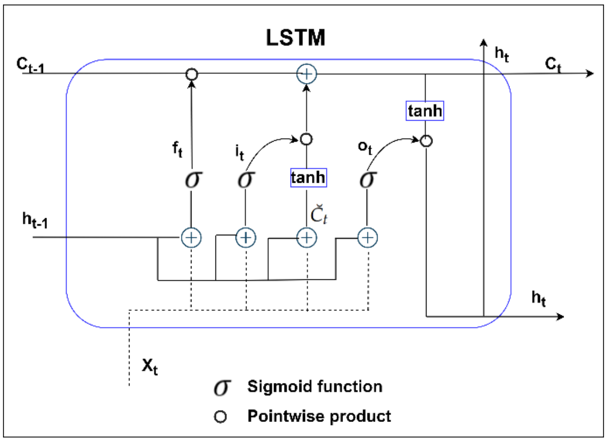


Fig. 4. LSTM Model

There are three gate functions in LSTM: the forget, the input, and the output. The forget gate is used to regulate how much information from C_{t-1} is maintained throughout the computation of C_t, and i_t the forget vector may be written as follows:

$$f_t = \sigma\left(U^f x_t + W^f h_{t-1} + b_f\right) \quad (4)$$

where U^f, W^f, and b_f are the forget gate's parameters, x_t is the input vector in step *t*, and h_{t-1} is the hidden state vector in the previous step. The input gate determines the amount of x_t information added to C_t and may be stated as follows:

$$f_t = \sigma\left(U^f x_t + W^f h_{t-1} + b_f\right) \quad (5)$$

where U^i, W^i, and b_i are the input gate's parameters and C_t may be determined by using both the input gate's vector i_t and the forget gate's vector f_t as shown below:

$$f_t = \sigma\left(U^f x_t + W^f h_{t-1} + b_f\right) \quad (6)$$

Where $\check{C}_t = \tanh(U^c x_t + W^c h_{t-1} + b_C)$ reflected the information represented by the vector of the hidden layer. Note that * represents the Hadamard product (element-wise). The output gate regulates the output in C_t, and:

$$o_t = \sigma(U^o x_t + W^o h_{t-1} + b_o), h_t = o_t * \tanh(C_t) \quad (7)$$

where U^o, W^o, and b_o are the output gate parameters and C_t is the internal state at time step t.

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

3) Gated Recurrent Unit (GRU)

A GRU model [39] (Fig. 5) is also an RNN and LSTM variant. However, GRU has only two gates: the update and the reset gates. Due to its simplicity and ease of training, GRU is superior to LSTM, in which data transmitted to the output is decided by the gates, which are two vectors. The update gate helps the model to calculate how much information from the past must be transmitted to the future. For step *t*, the update gate Z_t is computed using the following formula:

$$z_t = \sigma(U^z x_t + W^z h_{t-1}) \quad (8)$$

where U^z, W^z are the weights of the update gate and h_{t-1} stores data for the previous $(t-1)$ units. The reset gate is used to determine how much of the past data to forget, which is computed using the following formulas:

$$r_t = \sigma(U^r x_t + W^r h_{t-1}) \quad (9)$$

where U^r, W^r are the reset gate's weights and h_{t-1} holds information for the previous $(t-1)$ units. Using the reset gate, the current memory content will preserve the essential information from the past:

$$c_t = \tanh(U^c x_t + r_t * W^c h_{t-1}) \quad (10)$$

where U^c, W^c are the weights. Note that ∗ represents the Hadamard product (elementwise). The final step involves calculating the vector h_t that contains the information for the current unit and transmits it along the network:

$$h_t = z_t * h_{t-1} + (1 - z_t) * c_t \quad (11)$$

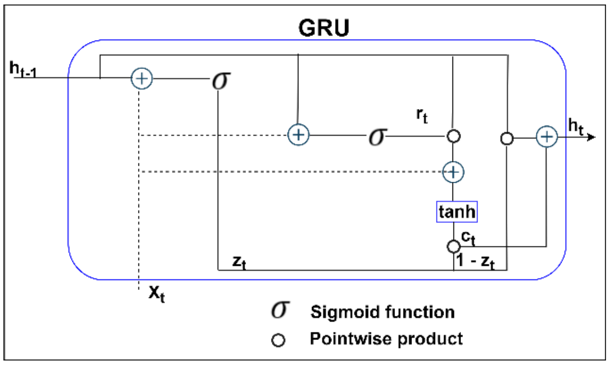


Fig. 5. GRU Model

A GRU network acquires a long spatial (or temporal) sequence with less computational complexity than a conventional encoder-decoder. GRU can handle longer sequences than regular RNNs because it can solve the vanishing gradient problem with gating techniques. To build context between features spread out across a large area or assess the degree of interconnections between features, GRU and LSTM may be used for sequences of spatial data.

4) Experiment Flow for Centralized Approach

We should recall that we use the N-BaIoT dataset, in which abnormal traffic is caused by malware, especially Botnets. Therefore, here we use "Botnet" as a general concept to indicate the source of abnormal traffic.

The Fig 6 shows the operational flow of the experimental process of Botnet detection in the centralized approach. Data is centralized in one place. First, we perform data processing and feature selection from the dataset data to process the data and select the attributes that are good enough to include in the training model. The dataset is divided into 70% for training, 10% for validation, and 20% for testing.

After that, we will apply three models, CNN, LSTM, and GRU, to detect abnormal traffic generated by Botnet.

The results are obtained, evaluated, and compared between these models.

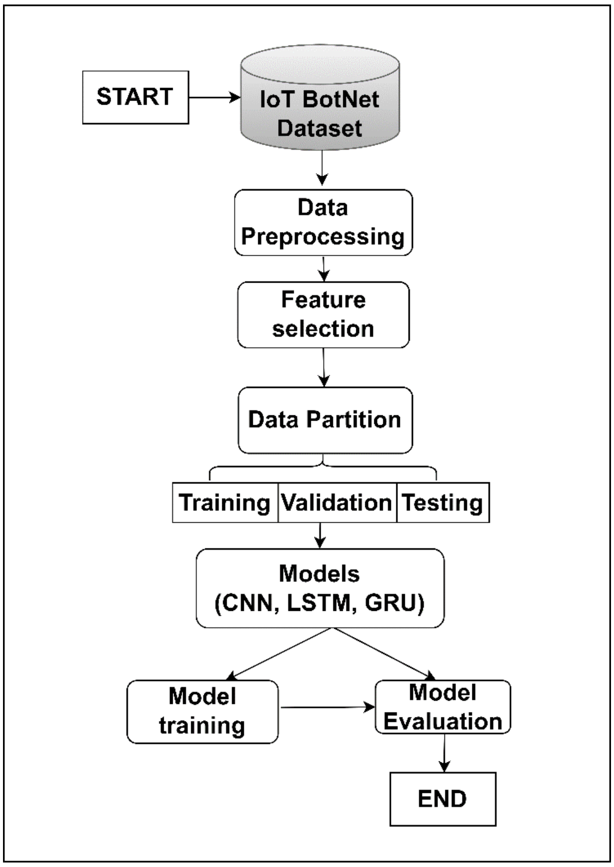


Fig. 6. Experiment Flow for Centralized Approach

E. Federated learning Model

This section describes the architecture of the FL approach. We will present the components and how they interact during the training and testing of models. In addition, it illustrates how the framework is implemented for our validation use case, which uses the N-BaIoT dataset. The Fig 7 depicts the framework architecture, which consists of K clients, each of which owns data from a single device and a server that coordinates the FL process. And the Fig 8 shows the back-and-forth interaction between the client-server and the preprocessing and feature selection process.

1) Model training component

Two FL techniques derived from the well-known FedAVG, Mini-batch aggregation, and Multi-epoch aggregation, are considered. The primary distinction between FedAVG and the other algorithms is that FedAVG considers the aggregation function as a parameter. As a result, the server can experiment with aggregating methods other than average. In this paper, we will implement the multi-epoch aggregation method. The model is trained for all E epochs simultaneously before

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

being transmitted to the server for multi-epoch aggregation.

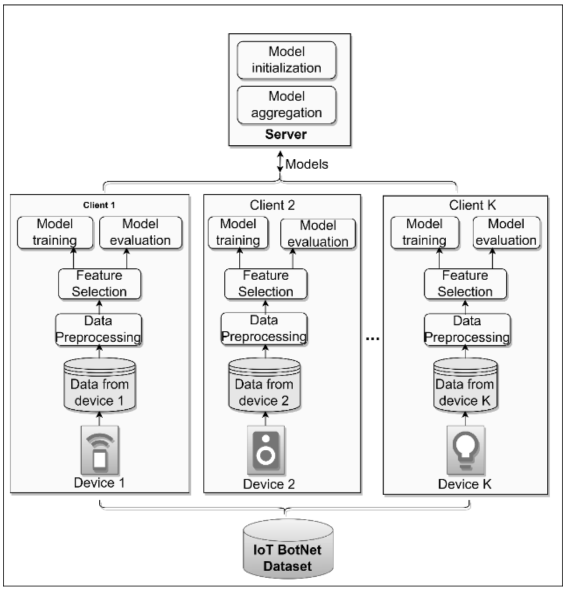


Fig. 7. FL framework architecture and its components

According to [20], averaging models may produce arbitrarily poor results due to the objective's non-convexity. This issue is significantly more likely to arise with multi-epoch aggregation since the models are trained individually for a much more extended time before being combined. To minimize this issue, multi-epoch aggregation training is performed for T = 30 rounds with a decreasing learning rate.

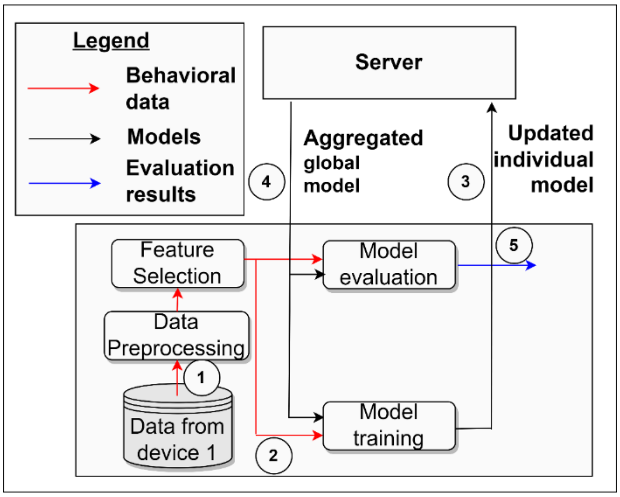


Fig. 8. Detailed view of the training process in the FL

2) *Server component*

The server is responsible for organizing the clients' training activities. In particular, it initializes the model from scratch. It compiles the models that the clients send into a so-called global model. The server is responsible for initializing the initial model's weights. All clients will have access to the first model when it is finished, and training may begin. It is important to note that every client starts with the same model. The server must combine the updated parameters from each client after receiving them to create the new global model parameters.

F. Evaluations

Using the N-BaIoT dataset, this experiment aims to evaluate how well our approach discovers abnormal IoT traffic caused by malware. It is crucial to compare the federated learning technique with conventional solutions to ensure that it is appropriate for IoT networks.

This research study evaluates the effectiveness of three well-known models, CNN, LSTM, and GRU, using three different approaches for anomaly detection in IoT networks. The first approach is a centralized approach that uses all attributes of the dataset. The second approach is Federated Learning with all attributes, while the third approach is Federated Learning with the division of attributes based on the time-frame characteristics of the dataset. The time-frame characteristics divide the attributes into five groups, corresponding to L1, L3, L5, L0.1, and L0.01, based on the time intervals of the collected streams. The evaluation provides insights into the performance of the different models and training approaches for anomaly detection in IoT networks, which can inform the development of more accurate and efficient intrusion detection systems.

The Fig 9 shows the accuracy of the centralized model based on all attributes of the dataset. The CNN model achieves the highest accuracy of 90.9%, while the LSTM and GRU models reach their highest accuracies of 88.39% and 90.66%, respectively. However, it is important to note that the LSTM and GRU models have more variability in their accuracies across epochs than the CNN model. For example, the LSTM model's accuracy fluctuates between 46.08% and 88.39%, while the GRU model's accuracy ranges from 56.5% to 90.9%. Furthermore, the accuracy of the LSTM model drops significantly after epoch 14, whereas the CNN and GRU models continue to perform well. The GRU model also shows a significant improvement in accuracy from epoch 1 to epoch 3, indicating that it may require more epochs to converge. Overall, the results suggest that the CNN model is the most reliable and consistent performer, while the LSTM and GRU models may require further optimization and fine-tuning to achieve optimal accuracy.

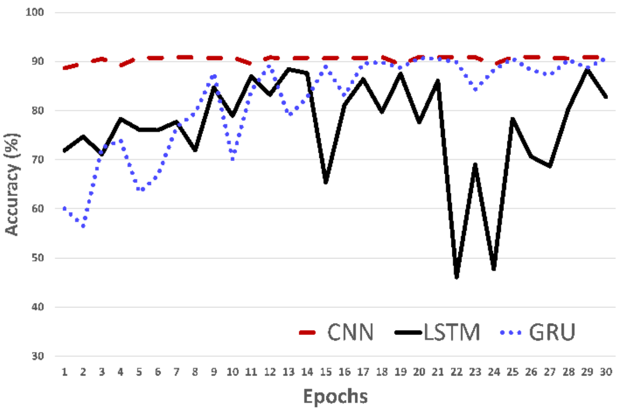


Fig. 9. The performance of the centralized learning model based on all attributes of the dataset

The Fig 10 shows the loss value when training the model with centralized data corresponding to CNN, GRU, and LSTM algorithms. It is interesting to note that the loss values of the LSTM and GRU models are much higher than those of the CNN model, especially in the earlier epochs. This could be due to the fact that the LSTM and GRU models are more complex than the CNN model, and therefore require more

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

training epochs to converge to a similar level of accuracy. It is also worth mentioning that the loss values of the LSTM model show significant fluctuations throughout the training process, with some epochs having much higher losses than others. This could be an indication that the LSTM model is more sensitive to the specific data samples used for each epoch of training. Overall, the CNN model achieves the lowest loss values consistently across all epochs, followed by the LSTM model and then the GRU model. This suggests that the CNN model is the most effective at reducing the error between predicted and actual values during training.

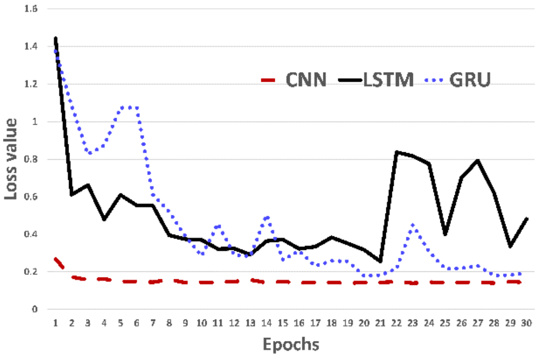


Fig. 10. Loss value of centralized learning model based on all attributes of the dataset

The Federated Learning approach's results of model training based on all attributes of the dataset with distributed data on each client are shown in Fig 11.

In Fig. 11, with the FL approach, the CNN model performs the best with the highest accuracy scores in most of the epochs. It starts with an accuracy of 85.017% in the first epoch and reaches up to 90.831% in the 25th epoch before slightly decreasing in the last few epochs.

On the other hand, the LSTM model has significantly lower accuracy scores compared to the CNN model in most of the epochs. It starts with an accuracy of 33.997% in the first epoch and gradually improves before reaching its peak at 88.803% in the 29th epoch. However, it shows a significant fluctuation in its performance throughout the epochs, with some epochs having a sharp drop in accuracy.

Similarly, the GRU model also has lower accuracy scores compared to the CNN model, but it performs better than the LSTM model in most of the epochs. It starts with an accuracy of 35.42% in the first epoch and gradually improves before reaching its peak at 89.56% in the last epoch. However, like the LSTM model, it also shows significant fluctuation in its performance in some epochs. Overall, the CNN model seems to perform better than the LSTM and GRU models in terms of classification accuracy on this dataset.

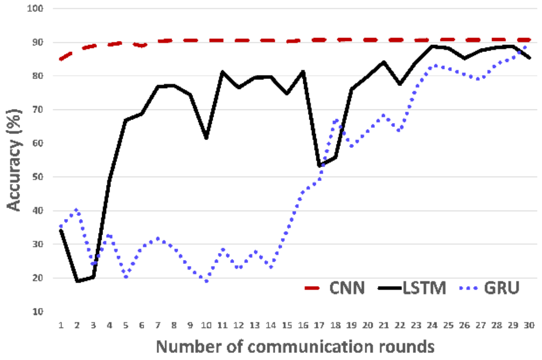


Fig. 11. Classification performance of the FL model based on all attributes of the dataset

The Fig 12 shows the loss value when training distributed data with the FL approach using all attributes. We can see that the loss values of CNN decrease gradually with each epoch, indicating that the model is improving in performance. On the other hand, the loss values for LSTM and GRU are more erratic, with occasional spikes in loss value. This could be due to the more complex nature of these models.

It is also worth noting that the loss values for LSTM and GRU are generally higher than those of CNN, suggesting that CNN is better suited for this particular task. Overall, the results suggest that CNN outperforms LSTM and GRU in terms of loss value.

Observing the accuracy of the models when deployed for centralized and distributed data, we see that the accuracy does not differ too much. The value fluctuates around 1% for accuracy. However, with centralized data, data privacy cannot be guaranteed. So in the case of ensuring data privacy, especially in IoT networks, the Federated Learning model will respond better.

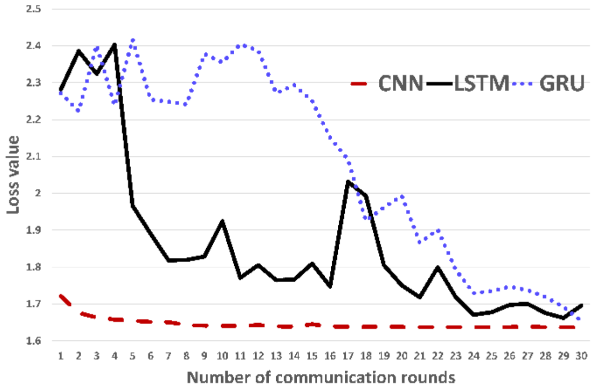


Fig. 12. Loss value of FL model based on all attributes of the dataset

The Fig 13 compares the training times of the centralized and FL-based approaches. The results show that the training time of the model when the data is concentrated is much better. Obviously, the distributed data training adds transmission, computation, and configuration update time for all clients so that the execution time will be longer.

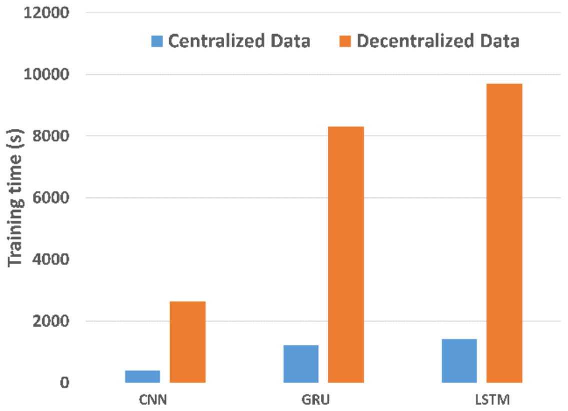


Fig. 13. Time training between centralized and decentralized based on all attributes of the dataset

Table II presents a comparison of three different algorithms, CNN, LSTM, and GRU, based on the metrics of precision, recall, F1-score, and accuracy, for centralized learning based on a time frame of L0.01.

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

According to the table, the CNN algorithm outperforms the LSTM and GRU algorithms in all four metrics, achieving the highest precision (0.94), recall (0.90), F1-score (0.87), and accuracy (0.90) compared to LSTM and GRU. LSTM achieves the second-best performance with a precision of 0.84, recall of 0.89, F1-score of 0.85, and accuracy of 0.89. GRU, on the other hand, performs the worst among the three algorithms, achieving a precision of 0.78, recall of 0.76, F1-score of 0.73, and accuracy of 0.76. The results suggest that the CNN algorithm is the most suitable for centralized learning based on a time frame of L0.01, as it achieves the highest accuracy and the best balance between precision and recall.

TABLE II
COMPARASION TABLE OF THE ALGORITHMS FOR CENTRALIZED LEARNING BASED ON TIME FRAME (L0.01)

Metrics	CNN	LSTM	GRU
Precision	0.94	0.84	0.78
Recall	0.90	0.89	0.76
F1-score	0.87	0.85	0.73
Accuracy	0.90	0.89	0.76

In Fig 14, the classification results using the FL model and based on the set of attributes of the time-frame L0.01 of the dataset, the results show that the CNN model achieves the highest accuracy among the three models with an average of 89.44% across all 30 epochs. LSTM and GRU models achieve significantly lower accuracy compared to CNN, with LSTM averaging 75.89% and GRU averaging 76.92% accuracy across all epochs.

It is also observed that the accuracy of all three models tends to increase over the course of the 30 epochs, with fluctuations in some epochs. However, the overall trend shows an increasing trend, with CNN having the most consistent increase in accuracy over time.

Additionally, it is worth noting that the accuracy of the LSTM and GRU models experiences more fluctuations compared to CNN. This may suggest that CNN is more stable and reliable than LSTM and GRU in this particular classification task.

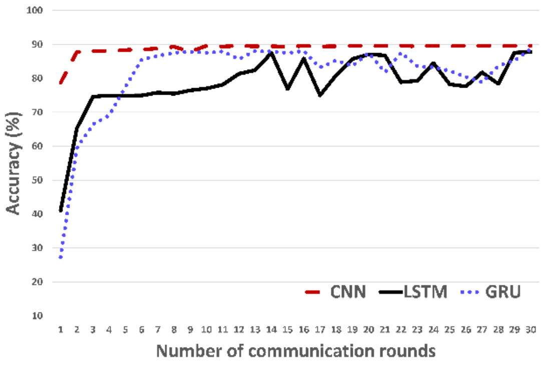


Fig. 14. Classification performance of the FL model based on class attribute L0.01 of the dataset

Table III shows the Precision, Recall, F1-score, and accuracy measures of CNN, LSTM, and GRU algorithms when training centralized data using only L0.1 time-frame algorithms. i.e. use a time frame of 0.1(s) to collect the stream information. The results show that the CNN algorithm gives the best results with an accuracy of 90%, while the LSTM algorithm gives relatively low results, only about 78%. As for the GRU algorithm, the classification result is 88%. Based on the experimental results, using only the attributes of the time frame L0.1 gives quite similar results to using all the attributes of the data set. However, with the L0.1 time-frame attribute set, the LSTM algorithm gives quite low results, while the CNN and GRU algorithms give many good results.

TABLE III
COMPARASION TABLE OF THE ALGORITHMS FOR CENTRALIZED LEARNING BASED ON TIME FRAME (L0. 1)

Metrics	CNN	LSTM	GRU
Precision	0.93	0.69	0.89
Recall	0.90	0.78	0.88
F1-score	0.87	0.70	0.85
Accuracy	0.90	0.78	0.88

In Fig 15, the classification results using the FL model and based on the attribute set of the time-frame L0.1 of the data set, the results show that the CNN algorithm gives the best results and reaches the value close to 90 % and converge very quickly starting from round 2. As for the LSTM algorithm, the results fluctuate quite a lot through each round, the learning ability of the LSTM model for this L0.1 time-frame attribute set Not good. Experiments show that using the attribute set of this time-frame L0.1 also gives quite similar results to using all the attributes of the data set.

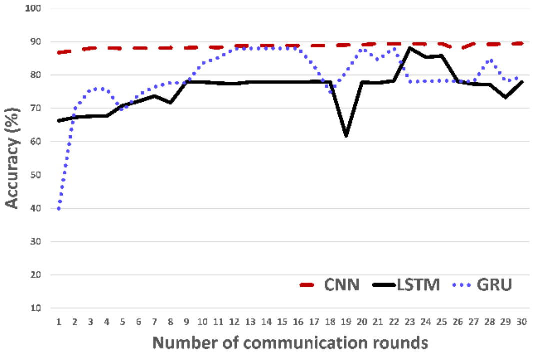


Fig. 15. Classification performance of the FL model based on class attribute L0.1 of the dataset

Table IV presents the evaluation metrics of Precision, Recall, F1-score, and accuracy for the CNN, LSTM, and GRU algorithms when trained on centralized data using L1 time-frame algorithms. L1 refers to a time frame of 1 second to gather the stream's information. According to the findings, the CNN algorithm performed the most accurately, with an accuracy rate of 89%, while the LSTM and GRU algorithms also produced satisfactory results, with accuracy rates of 88% and 86%, respectively. Moreover, these outcomes indicate that utilizing only the L1 time-frame attributes can yield comparable results to those obtained by using the complete set of dataset attributes.

TABLE IV
COMPARASION TABLE OF THE ALGORITHMS FOR CENTRALIZED LEARNING BASED ON TIME FRAME (L1)

Metrics	CNN	LSTM	GRU
Precision	0.91	0.83	0.82
Recall	0.89	0.88	0.86
F1-score	0.85	0.85	0.82
Accuracy	0.89	0.88	0.86

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

In Fig 16, the performance of the Federated Learning (FL) model is evaluated based on the L1 time-frame attributes of the dataset. The results indicate that the CNN algorithm outperforms the LSTM and GRU algorithms, achieving an accuracy of 90% with a fast convergence rate, starting from round 4. In contrast, the GRU algorithm's performance fluctuates considerably throughout each round, demonstrating poor learning ability for the L1 time-frame attribute set. Furthermore, the study suggests that using this L1 time-frame attribute set can yield results similar to those obtained when using all the attributes of the dataset.

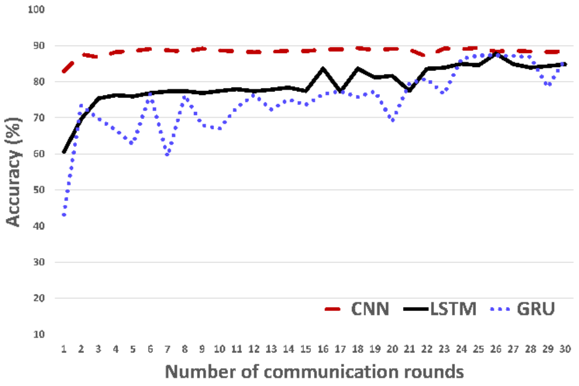


Fig. 16. Classification performance of the FL model based on class attribute L1 of the dataset

Table V presents the evaluation metrics of the CNN, LSTM, and GRU algorithms when trained on centralized data using only the L3 time-frame attribute set, which collects information from a time frame of 3 seconds.

The experimental results indicate that the CNN and LSTM algorithms achieve the highest accuracy of 88%, while the GRU algorithm yields comparatively low results of around 78% when trained on the L3 time-frame attribute set. These findings suggest that using the L3 time-frame attribute set can yield similar results to those achieved with the entire dataset.

TABLE V
COMPARASION TABLE OF THE ALGORITHMS FOR CENTRALIZED LEARNING BASED ON TIME FRAME (L3)

Metrics	CNN	LSTM	GRU
Precision	0.90	0.83	0.73
Recall	0.88	0.88	0.78
F1-score	0.85	0.84	0.70
Accuracy	0.88	0.88	0.78

In Fig 17 depicts the classification results obtained using the FL model and the L3 time-frame attribute set of the data set. The results indicate that the CNN algorithm achieves the highest accuracy of 90%, with rapid convergence starting from round 6. Conversely, the GRU algorithm produces fluctuating results throughout each round, indicating poor learning ability for this L3 time-frame attribute set.

Empirical findings indicate that adopting the L3 time-frame attribute set achieves comparable results to using the complete attribute set of the data.

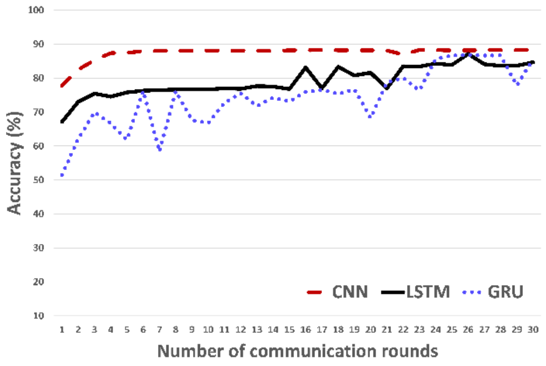


Fig. 17. Classification performance of the FL model based on class attribute L3 of the dataset

Table VI presents the evaluation metrics of the CNN, LSTM, and GRU algorithms when trained on centralized data using only the L5 time-frame attribute set, which collects information from a time frame of 5 seconds.

Based on the evaluation metrics presented in Table VI, the CNN, LSTM, and GRU algorithms were compared when trained on centralized data using only the L5 time-frame attribute set. The results indicate that the CNN algorithm achieved the highest accuracy of 87%, while the LSTM algorithm achieved an accuracy of 82%. However, the GRU algorithm gave relatively lower results with an accuracy of only 77%. It is noteworthy that the experimental results demonstrate that using only the L5 time-frame attributes produces quite similar results to using all the attributes of the data set.

TABLE VI
COMPARASION TABLE OF THE ALGORITHMS FOR CENTRALIZED LEARNING BASED ON TIME FRAME (L5)

Metrics	CNN	LSTM	GRU
Precision	0.87	0.66	0.79
Recall	0.87	0.77	0.82
F1-score	0.83	0.70	0.79
Accuracy	0.87	0.77	0.82

The Fig 18 illustrates the classification results obtained by the FL model based on the L5 time-frame attribute set of the data set. The results indicate that the CNN algorithm provides the highest accuracy of 90% and exhibits the fastest convergence rate, starting from round 6. Conversely, the GRU algorithm's results exhibit significant fluctuations across each round, indicating that the model's learning ability for this L5 time-frame attribute set is insufficient. Based on the experimental results, it can be inferred that employing only the L5 time-frame attribute set yields results that are comparable to using all attributes of the dataset.

Based on the experiments conducted, the results indicate that the use of attribute sets based on time-frames yields comparable outcomes to using the entire dataset. Furthermore, a minor variation in accuracy was observed between centralized and distributed data models, fluctuating around 1%. However, the centralized approach is not adequate for ensuring data privacy, which is crucial in IoT networks. Therefore, the Federated Learning model is a more appropriate approach to preserving data privacy.

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

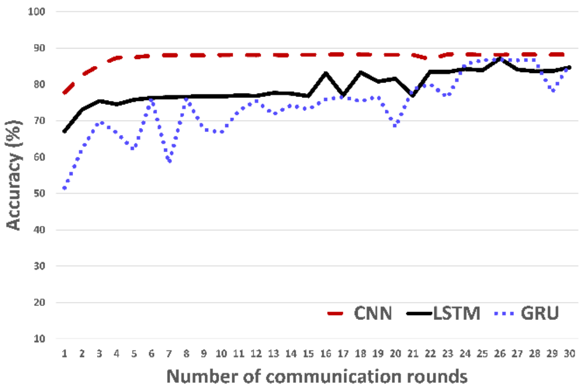


Fig. 18. Classification performance of the FL model based on class attributes L5 of the dataset

IV. Conclusion

Based on the results obtained from the N-BaIoT dataset, this study recommends using the Federated Learning (FL) approach for detecting IoT malware abnormal traffic while preserving user privacy. Two different methods were compared, i.e., the Federated Learning method, where each device owner trains a separate model, and the Centralized approach, where each device owner trains a single, isolated model using a centralized dataset. Although the Centralized approach still has slightly higher accuracy in detecting abnormal IoT traffic than FL, the difference is insignificant. However, the Centralized approach model fails to ensure the security and privacy of confidential information, unlike the FL model. Furthermore, when comparing the CNN, LSTM, and GRU models applied in the FL approach, CNN yields the best results and is suitable for simple FL models.

In future studies, our research aims to further optimize the Federated Learning model, specifically focusing on the global parameter to enhance its accuracy in detecting abnormal IoT traffic. Additionally, we plan to improve the processing time to achieve more efficient and effective results.

Acknowledgment

The study was financially supported by the Russian Science Foundation within of scientific project No. 22-49-02023 ”Development and study of methods for obtaining the reliability of tethered high-altitude unmanned telecommunication platforms of a new generation.

References

[1] Fan, Xiaochen, et al. "BuildSenSys: Reusing building sensing data for traffic prediction with cross-domain learning." *IEEE Transactions on Mobile Computing* 20.6 (2020): 2154-2171.

[2] Alrawi, Omar, et al. "Sok: Security evaluation of home-based iot deployments." *2019 IEEE symposium on security and privacy* (sp). IEEE, 2019.

[3] Kulik, V., & Kirichek, R. (2018, November). The heterogeneous gateways in the industrial internet of things. *In 2018 10th International congress on ultra modern telecommunications and control systems and workshops* (ICUMT) (pp. 1-5). IEEE.

[4] Kumar, J. Sathish, and Dhiren R. Patel. "A survey on internet of things: Security and privacy issues." *International Journal of Computer Applications 90.11* (2014).

[5] Butun, Ismail, Patrik Österberg, and Houbing Song. "Security of the Internet of Things: Vulnerabilities, attacks, and countermeasures." *IEEE Communications Surveys & Tutorials 22.1* (2019): 616-644.

[6] Radhakrishnan, Divya. "Internet of things: Privacy and security issues: A qualitative study about the internet of things and its privacy and security challenges from a developer's perspective." (2021).

[7] Tsimenidis, S., Lagkas, T., & Rantos, K. (2022). Deep learning in IoT intrusion detection. *Journal of network and systems management*, 30, 1-40.

[8] Aldweesh, A., Derhab, A., & Emam, A. Z. (2020). Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. *Knowledge-Based Systems*, 189, 105124.

[9] Kairouz, Peter, et al. "Advances and open problems in federated learning." *Foundations and Trends® in Machine Learning* 14.1–2 (2021): 1-210.

[10] Otoum, Safa, Ismaeel Al Ridhawi, and Hussein T. Mouftah. "Blockchain-supported federated learning for trustworthy vehicular networks." *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020.

[11] Meidan, Yair, et al. "N-baiot—network-based detection of iot botnet attacks using deep autoencoders." *IEEE Pervasive Computing* 17.3 (2018): 12-22.

[12] Arbex, Gustavo Vitral, et al. "IoT DDoS Detection Based on Stream Learning." *2021 12th International Conference on Network of the Future* (NoF). IEEE, 2021.

[13] Ferdowsi, Aidin, and Walid Saad. "Generative adversarial networks for distributed intrusion detection in the internet of things." *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019.

[14] Almiani, Muder, et al. "Deep recurrent neural network for IoT intrusion detection system." *Simulation Modelling Practice and Theory* 101 (2020): 102031.

[15] Xia, Qinyu, Shi Dong, and Tao Peng. "An Abnormal Traffic Detection Method for IoT Devices Based on Federated Learning and Depthwise Separable Convolutional Neural Networks." *2022 IEEE International Performance, Computing, and Communications Conference* (IPCCC). IEEE, 2022.

[16] Sim, Khe Chai, et al. "Domain Adaptation Using Factorized Hidden Layer for Robust Automatic Speech Recognition." *Interspeech*. 2018.

[17] Almiani, Muder, et al. "Deep recurrent neural network for IoT intrusion detection system." *Simulation Modelling Practice and Theory* 101 (2020): 102031.

[18] Vladimirov, Sergey, and Ruslan Kirichek. "The IoT identification procedure based on the degraded flash memory sector." *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*. Springer, Cham, 2017. 66-74.

[19] Lou, Yang, et al. "Predicting network controllability robustness: A convolutional neural network approach." *IEEE Transactions on Cybernetics* 52.5 (2020): 4052-4063.

[20] Amanullah, Mohamed Ahzam, et al. "Deep learning and big data technologies for IoT security." *Computer Communications* 151 (2020): 495-517.

[21] Sim, Khe Chai, et al. "Domain Adaptation Using Factorized Hidden Layer for Robust Automatic Speech Recognition." *Interspeech*. 2018.

[22] Siniosoglou, Ilias, et al. "A unified deep learning anomaly detection and classification approach for smart grid environments." *IEEE Transactions on Network and Service Management* 18.2 (2021): 1137-1151.

[23] Larriva-Novo, Xavier, et al. "An IoT-focused intrusion detection system approach based on preprocessing characterization for cybersecurity datasets." *Sensors 21.2* (2021): 656.

[24] Zhu, Hangyu, Haoyu Zhang, and Yaochu Jin. "From federated learning to federated neural architecture search: a survey." *Complex & Intelligent Systems* 7 (2021): 639-657.

[25] Abbasi, Mahmoud, Amir Taherkordi, and Amin Shahraki. "FLITC: A Novel Federated Learning-Based Method for IoT Traffic Classification." *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 2022.

[26] Mun, Hyunsu, and Youngseok Lee. "Internet traffic classification with federated learning." *Electronics* 10.1 (2020): 27.

[27] He, Zhimin, et al. "Edge device identification based on federated learning and network traffic feature engineering." *IEEE Transactions on Cognitive Communications and Networking* 8.4 (2021): 1898-1909.

[28] Wen, Jie, et al. "A survey on federated learning: challenges and applications." *International Journal of Machine Learning and Cybernetics* (2022): 1-23.

[29] Zhu, Hangyu, Haoyu Zhang, and Yaochu Jin. "From federated learning to federated neural architecture search: a survey." *Complex & Intelligent Systems* 7 (2021): 639-657.

[30] Do, Phuc Hao, et al. "An Efficient Feature Extraction Method for Attack Classification in IoT Networks." *2021 13th International*

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). IEEE, 2021.

[31] Cohen, Patricia, Stephen G. West, and Leona S. Aiken. Applied multiple regression/correlation analysis for the behavioral sciences. *Psychology press*, 2014.

[32] Conrod, Patricia J., et al. "Effectiveness of a selective, personality-targeted prevention program for adolescent alcohol use and misuse: a cluster randomized controlled trial." *JAMA psychiatry* 70.3 (2013): 334-342.

[33] Cerda, Patricio, Gaël Varoquaux, and Balázs Kégl. "Similarity encoding for learning with dirty categorical variables." *Machine Learning* 107.8 (2018): 1477-1494.

[34] Guo, Cheng, and Felix Berkhahn. "Entity embeddings of categorical variables." *arXiv preprint* arXiv:1604.06737 (2016).

[35] Zhou, Yuyang, et al. "Building an efficient intrusion detection system based on feature selection and ensemble classifier." *Computer networks* 174 (2020): 107247.

[36] Mahfouz, Ahmed, et al. "Ensemble classifiers for network intrusion detection using a novel network attack dataset." *Future Internet* 12.11 (2020): 180.

[37] Yamashita, Rikiya, et al. "Convolutional neural networks: an overview and application in radiology." *Insights into imaging* 9 (2018): 611-629.

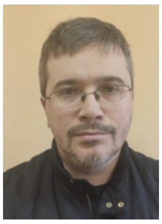
[38] Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv preprint* arXiv:1412.3555 (2014).

[39] Volkov, A., Khakimov, A., Muthanna, A., Kirichek, R., Vladyko, A., & Koucheryavy, A. (2017, June). Interaction of the IoT traffic generated by a smart city segment with SDN core network. *In International Conference on Wired/Wireless Internet Communication* (pp. 115-126). Springer, Cham.

Phuc Hao Do received his MS degree in Computer science from the University of Danang - University of Science and Technology in 2017. He is currently a Ph.D. student in the Department of Communication Networks and Data Transmission at the Bonch-Bruevich Saint- Petersburg State University of Telecommunications, Russia. His research interests include AI, ML, D and its application in different fields like network, blockchain.

Dr. Tran Duc Le acquired his degree of Ph.D. at Admiral Makarov State University of Maritime and Inland Shipping, Russia in 2018. He has been working in Information Technology Faculty, The University of Danang - University of Science and Technology, Danang, Vietnam since 2019. His research areas include the Internet of Things, Security Analytics, Malware Analysis.

Dr. Sc. Vladimir M. Vishnevsky received the Engineering degree in applied mathematics from the Moscow Institute of Electronics and Mathematics, Russia, in 1971, the Ph.D. degree in queuing theory and telecommunication networks and the D.Sc. degree in telecommunication networks from the V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences (ICS RAS), in 1974 and 1988, respectively. He became a Full Professor with ICS RAS in 1989 and the Moscow Institute of Physics and Technology in 1990. He was an Assistant Head of the Institute of Information Transmission Problems of RAS from 1990 to 2010 and an Assistant Head of laboratory with ICS RAS from 1971 to 1990. He is currently the Head of Telecommunication Networks Laboratory, ICS RAS. He is a member of Expert Councils of Russian High Certifying Commission and Russian Foundation for Basic Research, member of IEEE Communication Society, International Telecommunications Academy and New York Academy of Science. He has authored over 300 papers in queuing theory and telecommunications. He is a Co-Chair of IEEE conferences - ICUMT, RTUWO, and the General Chair of DCCN conference. His research interests lie in the areas of computer systems and networks, queuing systems, telecommunications, discrete mathematics (extremal graph theory, mathematical programming) and wireless information transmission networks.



Dr Aleksandr Berezkin, is working at the Bonch Bruevich Saint Petersburg State University of Telecommunications as the Associate Professor of Department of Programming Engineering and Computer Science. Science interest are Computer Vision and Machine Learning. In 2009, he defended his thesis with the topic "Model and method of decoding error correction based on neural network". Now he is doctoral student at the Department of Programming Engineering and Computer Science.

Dr. Sc. Ruslan Kirichek is working at the Bonch Bruevich Saint Petersburg State University of Telecommunications as the head of Department of Programming Engineering and Computer Science. He was born in 1982 in Tartu (Estonia). He graduated Military-Space Academy A.F. Mozhaiskogo and the Bonch-Bruevich St. Petersburg State University of Telecommunications in 2004 and 2007, respectively. He received Ph.D. at the Bonch-Bruevich St. Petersburg State University of Telecommunications in 2012 and Dr.Sc. at the Povolzhskiy State University of Telecommunications and Informatics in 2018. From 2008 to 2013 he worked as a senior researcher at the Federal State Unitary Enterprise "Center-Inform". Since 2012 he has been working as the Head of the Internet of Things Laboratory at the Bonch-Bruevich Saint Petersburg State University of Telecommunications. Since 2017 he has been working as ITU-T Q12/11 Rapporteur in “Testing of Internet of things, its applications and identification systems”. Since 2023 he has been working as the Rector of the Bonch-Bruevich Saint Petersburg State University . He is a General Chair of the International Conference “Internet of Things and Its Enablers” (inthiten.org).

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

A Deep learning Framework for Cultural Heritage Damage Detection for Preservation; Based on the case of Heunginjimun and Yeongnamnu in South Korea

Sang-Yun Lee*, Daekyeom Lee**

**Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea*

*** SEASON Co., Ltd., Sejong, Republic of Korea*

syllee@etri.re.kr, daek29@season.co.kr

***Abstract*—In general, there are many restrictions on investigations for safety inspection due to the uniqueness of cultural heritages. Methods such as visual inspection and non-destructive inspection, which are mainly used as inspection methods, are regularly carried out, but there are limitations on time and cost. This is insufficient to identify and respond quickly when an abnormal symptom appears in cultural heritage. As a basic study of system development for rapid abnormal detection of architectural, cultural properties through Deep Learning, this paper organized a Deep Learning framework for detecting tilt in buildings for the roof of Heunginjimun Gate (Korea Treasure No. 1) and Yeongnamnu Pavilion (Korea Treasure No. 147). A framework was developed using a Convolutional Neural Network (CNN). As a result of an application, EfficientnetB0 and EfficientnetB2 models showed excellent accuracy in detecting the tilt of the roof of Heunginjimun with an average accuracy of 99.66% and 99.69%, respectively. In addition, EfficientnetB0, EfficientnetB2, and Shufflenet_v2 models showed excellent accuracy in detecting tilt of the roof of Yeongnamnu with 98.81%, 99.80%, and 98.48% accuracy. Additionally, the Grad-CAM experiment was conducted as a basis for whether the model made the proper judgment to confirm the criteria for determining abnormal detection according to the results of each model. These findings quickly detect abnormalities occurring in cultural heritages from the perspective of cultural heritage management and preservation, enabling rapid response, and are valuable for research on artificial intelligence technology related to cultural heritages.**

***Keyword*—Conventional Neural Network, Cultural Heritage, Grad-CAM, Preservation**

I. INTRODUCTION

In recent years, disastrous incidents have occurred during the attempts to preserve cultural assets, such as the tilting of Cheomseongdae in Gyeongju due to the earthquake in Pohang in 2016, the burning of Sungnyemun Gate, the attempted arson of Heunginjimun Gate in 2018, and the collapse of the Gongsanseong Fortress Wall due to torrential rain in July 2020.

According to the data on the status of emergency repair of cultural properties by cause of damage that occurred for 6 years in the data of 'Cultural heritage in statistics (2021, 2022)' published by the Cultural Heritage Administration of South Korea, total 319 cases of damage occurred for 6 years, storm and flood damage (typhoon, strong wind, heavy rain), biological damage (termites, pests), others (collapse, fall, unknown cause, etc.), cold waves, and earthquakes accounted for 168 cases, 48 cases, 36 cases, 19 cases, and 16 cases, respectively [1, 2].

In line with the situation above, the cost of maintenance and repair of cultural properties is also increasing. According to the current status of cultural asset repair and maintenance by the Cultural Heritage Administration of the Republic of Korea, the amount spent on cultural property repair was about 432 billion won in 2018, about 526 billion won in 2019, about 570 billion won in 2020, about 577 billion won in 2021, and about 574 billion won in 2022. It shows a gradually increasing trend [3].

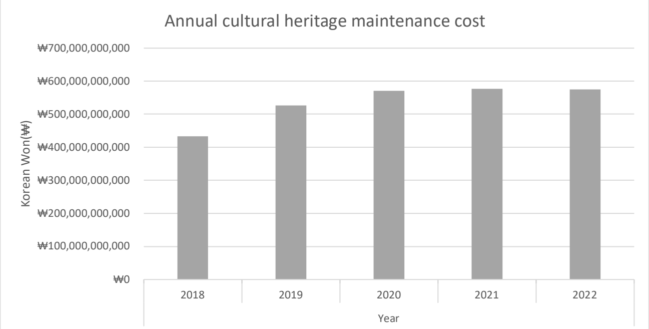


Fig. 1. Annual cultural heritage maintenance cost in South Korea (Reorganized based on KOSIS, 2022)

South Korea's cultural heritages are valuable assets that embodies human culture, so it is imperative that they be

Manuscript received Dec. 13, 2022. This paper was funded by the government (Cultural Heritage Administration) in 2023 with the support of the National Research Institute of Cultural Heritage (Grant ID: No.2021A01D06-001), and a follow-up of the invited journal to the accepted & presented paper entitled " Damage Detection and Safety Diagnosis for Immovable Cultural Assets Using Deep Learning Framework" of the 25th International Conference on Advanced Communication Technology (ICACT2023).

Sang-Yun Lee, he is a Ph. D. and working for Electronics and Telecommunications Research Institute (ETRI) as a principal researcher in Daejeon, 34129, Republic of Korea (corresponding author to provide phone: +82-10-2995-0928; E-mail: syllee@etri.re.kr)

DaeKyeom Lee, he is working ofr Season Co., Ltd in Sejong-si, 30127, Republic of Korea. He is now leading the AI Convergence Technology Research Team. (phone: +82-10-3287-9100; E-mail: daek29@season.co.kr)

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

handed down to future generations. Therefore, regular inspection and maintenance work is required to sufficiently maintain the structural stability of aging cultural properties due to issues such as climate change, Artificial activity, and deterioration. In addition, since damage to cultural heritage is irreversible once incited, efforts should be made to prevent them from being destroyed or damaged. However, due to the unique qualities of cultural heritage, there are many restrictions on safety inspections, so inspection is usually conducted by methods such as visual inspection and other non-destructive methods. Methods such as visual inspection and non-destructive inspection are usually performed on a regular basis, but there are limitations on time and cost. In addition, when abnormal symptoms appear in cultural properties, identifying them and taking immediate initial responses is not enough. In accordance to this issue, attempts are being made to combine cultural heritage conservation and artificial intelligence technology to respond to cultural heritage damage [4-14]. As a representative case, Mishra et al implemented a case study on Dadi - Poti tom in New Deli Hauz Khas Village. They developed a You Only Look Once (YOLOv5) real-time object detection algorithm and ResNet 101-based R-CNN through customized defect detection and localization. A study was conducted using the model to detect four types of defects: discoloration, brick exposure, cracks, and delamination [4]. Mansuri and Patel used the R-CNN model to build an automatic visual inspection system in Surat, India, and British and Dutch cemeteries as well have found 'breakthroughs', 'exposed brick', 'cracks' spalling exposed bricks, and cracks existing. A study was conducted to detect these three types of defects in heritage structures [5]. Yu et all generated data and used deep networks to carry out the Dunhuang cultural heritage protection project known as Thousand Buddha Caves [7]. The papers above performed machine learning mainly on stone structures, unlike this paper, which is for wooden buildings.

In this paper, Deep Learning video data taken by CCTV of the roof of Heunginjimun (Treasure No. 1), and Yeongnamnu (Treasure No. 147), were used for Deep Learning. In the case of Heunginjimun, it is the only one among the gates of the capital city surrounding Hanyang, the capital of the Joseon Dynasty, that maintains the form of Onseong (a double wall surrounding the gate to protect it). Like Sungnyemun, which is National Treasure No. 1, the gatehouse has multiple levels (double-layer), and like Gyeongbokgung Palace, the existence of Jabsang (Small figures on the roof. People believed that they protected the building from evil spirits, especially fire spirits.) on the roof indicates that the building holds significant historical value [15,16]. Also, in the case of Yeongnamnu, it is a pavilion-style building that is a representative form of South Korean traditional architecture. It is also a representative wooden structure of the late Joseon Dynasty. It is a cultural property with high preservation value, as it is being called one of the three significant pavilions in Korea [17,18,19]. In addition, as a preceding study, [20] proposed a Deep Learning Framework that can detect fine gradients for Heunginjimun. However, unlike this paper, since the experiment was performed with only a single subject of Heunginjimun, this paper additionally included Yeongnamru as a subject, and confirmed the results with Grad-CAM to verify that Deep Learning is performed well.

By using the data from the aforementioned cultural heritages, we will identify an optimal Deep Learning model and propose a building abnormality detection system by constructing a framework. The Deep Learning framework developed in this study is believed to be of great help for the preservation and transmission of cultural assets. The structure of this thesis is as follows. In Chapter 2, the data set composition and preprocessing process for generating abnormal data were explained, and in Chapter 3, the methodology used in this experiment and the parameter setting of the used model were described. In Chapter 4, the experimental results of each model and Grad-CAM results for verify were presented, and finally, the conclusion is made in Chapter 5.

II. DATA COLLECTION AND PREPROCESSING

A. The composition of the dataset

In order to construct a dataset for use in the damage detection Deep Learning framework of cultural heritage, images of cultural heritage filmed through CCTV were used. CCTV images installed at the front of Heunginjimun and Yeongnamnu were used, and the data recorded by CCTV are converted into AVI video format in IRAS (IDIS CCTV monitoring SW). The converted AVI video was divided into 600 frame units, saved as JPG pictures, and input to the pre-trained model. The dataset constructed using CCTV images as above was divided into five categories: Clear-Day, Clear-Night, Rainy-Day, Rainy-Night, and Cloudy-Day by environment. The composition of each environment dataset consists of 10000 pieces of data and consists of 5000 pieces of the training set, 2500 pieces of the validation set, and 2500 pieces of the test set. The ratio of the experiment's train set, validation set, and test set was 50% 25% 25%. Table. 1 shows the composition and ratio of each data set of Heunginjimun and Yeongnamnu.

TABLE I
THE COMPOSITION AND RATIO OF THE DATASET

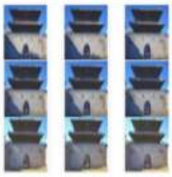
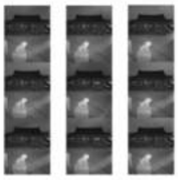
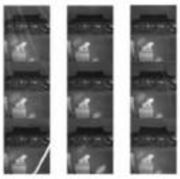
Heunginjimun								
Dataset	Normal/Abnormal (Ratio)	Train/Validation/ Test (Quantity)	Train set (Quantity)		Validation set (Quantity)		Test set (Quantity)	
			Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Clear-Day	5:5	5000/2500/2500	2500	2500	1250	1250	1250	1250
Clear-Night								
Rainy-Day								
Rainy-Night								
Cloudy-Day								
Yeongnamnu								
Dataset	Normal/Abnormal (Ratio)	Train/Validation/ Test (Quantity)	Train set (Quantity)		Validation set (Quantity)		Test set (Quantity)	
			Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Clear-Day	5:5	5000/2500/2500	2500	2500	1250	1250	1250	1250
Clear-Night								
Rainy-Day								
Rainy-Night								
Cloudy-Day								

B. Configuring the data set's Environments

In order to predict the tilt of the roof part, prediction experiments were conducted in various environments. We tried to reflect on the most common environments according to our data. The roof of Heunginjimun was divided into five environments: Clear-Day, Clear-Night, Rainy-Day, Rainy-Night, and Cloudy-Day. Also, it was divided into Day and

Night based on the time when CCTV converted to night camera. Yeongnamnu also divided the environment into same. Table. 2 shows examples of data for each environment.

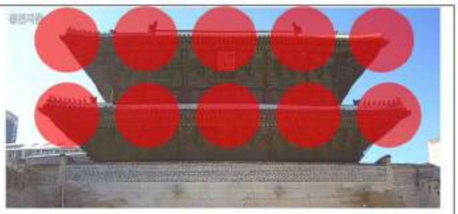
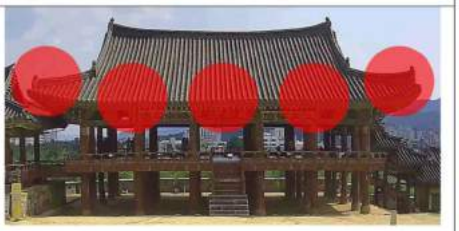
TABLE II
EXAMPLES OF DATASETS BY ENVIRONMENT

Data Example		
Environment	Heunginjimun	Yeongnamnu
Clear-Day		
Clear-Night		
Rainy-Day		
Rainy-Night		
Cloudy-Day		

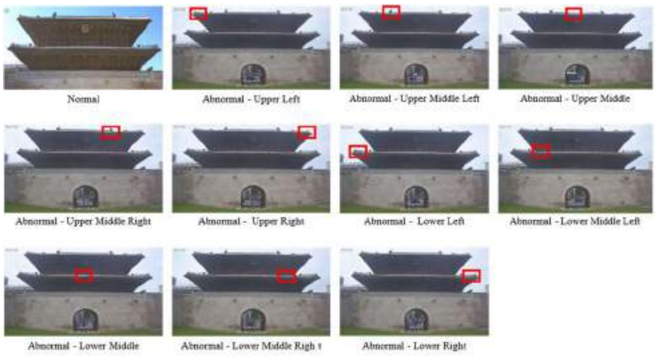
C. Generation of Abnormal data

As with the normal data we needed for Deep Learning, an example of abnormal data was needed to explore the abnormal state of architecture. However, once a cultural heritage is damaged, it is often difficult to restore it to its original state. Even if it can be restored, it takes a tremendous amount of money and time, so collecting enough abnormal data for machine learning is very difficult. Therefore, as an alternative method, in the case of abnormal data, the Adobe Photoshop action function was used to generate abnormal data for each scenario in each environment by stretching the eaves. In the case of the roof of Heunginjimun, it is composed of two layers, and displacements were applied to a total of 10 areas. The area where the displacement was applied is marked by a red circle. Unlike Heunginjimun, the roof of Yeongnamnu is composed of a single layer, and displacements were applied to five areas accordingly. As with Heunginjimun, the area marked with a red circle is the area where the displacement was used. Table. 3 shows the displacement points of Heunginjimun and Yeongnamnu.

TABLE III
DISPLACEMENT POINTS FOR GENERATING ABNORMAL DATA

Heunginjimun	
Yeongnamnu	

The samples of the abnormal data set generated by applying displacement to the points of 10 areas of the roof of Heunginjimun and 5 areas of the roof of Yeongnamnu are shown in Fig. 2 and Fig. 3 below. When checking the abnormal data compared to the original, it can be seen that the tilt of the displaced area (the red rectangle in Fig. 2 and Fig. 3) is slightly different.



Normal

Abnormal - Upper Left

Abnormal - Upper Middle Left

Abnormal - Upper Middle

Abnormal - Upper Middle Right

Abnormal - Upper Right

Abnormal - Lower Left

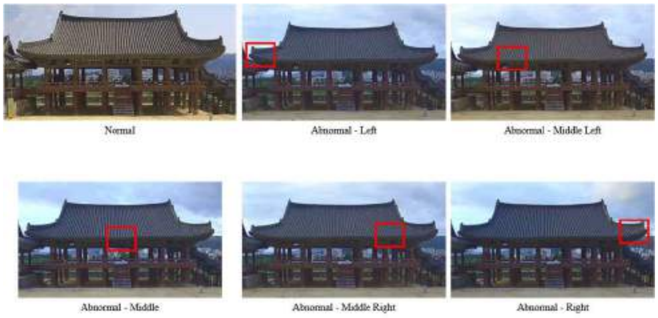
Abnormal - Lower Middle Left

Abnormal - Lower Middle

Abnormal - Lower Middle Righ t

Abnormal - Lower Right

Fig. 2. Samples of abnormal images of Heunginjimun



Normal

Abnormal - Left

Abnormal - Middle Left

Abnormal - Middle

Abnormal - Middle Right

Abnormal - Right

Fig. 3. Samples of abnormal images of Yeongnamnu

III. METHODOLOGY

A. Convolutional Neural Network (CNN)

Convolution Neural Network (CNN) is a technique announced by LeCun and Bengio in 1995 [21]. It is an algorithm that shows excellent performance in image recognition or voice recognition among Deep Learning methods, and its structure is designed according to neurobiological principles [22-25]. In the case of CNN, it has the ability to analyze the specific characteristics of the object in question regardless of the location and the direction in which it is placed. CNN consists of a convolution layer that

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

extracts features from an input image and a pooling layer that compresses the extracted information. Convolution and pooling are repeated to extract the features [25]. Furthermore, CNN uses a filter in the convolution step to extract the features of an image. As shown in Fig. 4, the Filter composed of matrices in the Image Matrix scans the image from top left to bottom right.

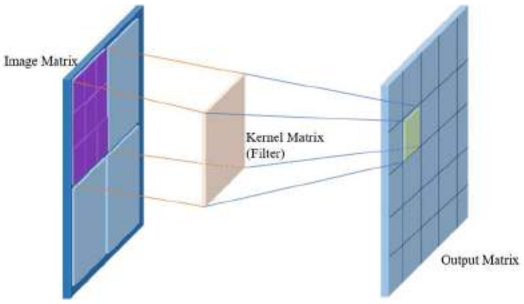


Fig. 4. Feature extraction in convolution

Subsequently, after going through the feature extraction step in which convolution and pooling are repeated, the classification step is passed as shown in Fig.5. The feature matrix extracted in the flattening step is flattened into vectors. Then the sorted vectors go through a fully connected layer and an activation function such as softmax or sigmoid is ordered to output a class and classify the image and create a result

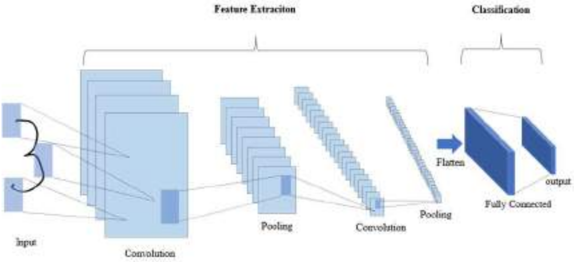


Fig. 5. Convolutional Neural Network Architecture

B. Experimental Environment

This experiment was performed on Ubuntu 20.040.2. and under the LTS environment. The specifications and other packages of the devices used in this study are as follows.

- Memory 64GB
- CPU 24 core
- GPU 1080(GeForce GTX 1080ti)
- CUDA 11.3
- Python 3.9
- efficientnet-pytorch==0.7.1
- matplotlib==3.6.0
- nnpack==0.1.0
- opencv-python==4.6.0.66
- Pillow==9.2.0
- scikit-learn==1.1.2
- tqdm==4.64.1
- numpy==1.23.3
- decord==0.6.0
- imgaug==0.4.0
- ttach
- tqdm

C. Framework design and implementation

A Deep Learning framework was designed to find the tilt that occurs on the roof of Heunginjimun and the roof of Yeongnamnu. The Deep Learning framework is shown in Fig. 6

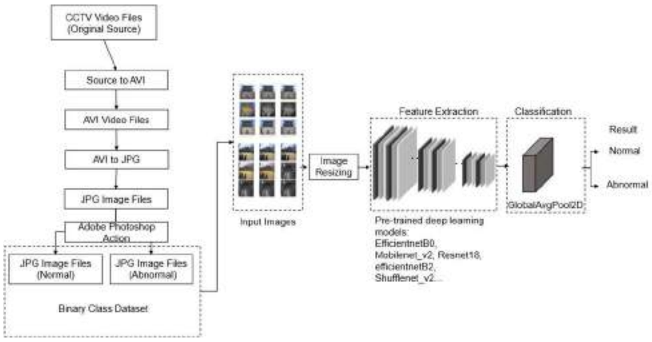


Fig. 6. The framework for cultural heritage damage detection

The Deep Learning framework first converts the video generated by CCTV filming cultural heritage into an AVI video format, cuts them in frame units, and extracts them into JPG images. The extracted image is used as normal data, and in the case of abnormal data, it is created by applying distortion to the normal data. This method of applying distortion to normal data uses the action cam function of Adobe Photoshop. Thereafter, an image is adjusted through a resizing procedure to use the data in a neural network model. After image resizing, the dataset is divided into training, testing, and verification sets and used as input values for neural network models. Finally, through the feature extraction part and classification part of the model, it is classified into two classes: normal and abnormal.

D. Model selection and parameter settings

The neural network model used in machine learning loads image data from the dataset created through the preprocessing process and uses normal and abnormal images as input values for the neural network model. For the neural network model used for classification, the most commonly used models were selected: efficientnetB0 [26], mobilenet_v2 [27], resnet18 [28], efficientnetB2, shufflenet_v2 [29], alexnet [30], mnasnet [31], inception_v3[32], densernet161[33], efficientnetb4, and 10 types of pre-learning-based Deep Learning models were used. The dataset used for pre-learning is ImageNet. Learning is conducted to classify it into normal and abnormal classes through the feature extraction and classification parts of the model. Table. 4 shows the parameters used for machine learning. Optimal parameters were used as long as they were versatile for each environment. In the case of the optimizer, it was adopted through Stochastic Gradient Descent (SGD). The initial learning rate was set to 0.002 for Alexanet and 0.05 for EfficientnetB0, EfficientnetB2, Shufflenet_v2, and Mnasnet. The momentum value was set to 0.9, and 10 to 20 epochs were performed.

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

TABLE IV
PARAMETERS OF HEUNGINJIMUN AND YEONGNAMNU

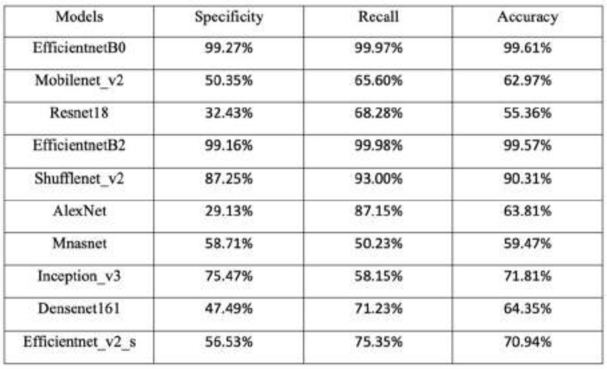
	Heunginjimun				Yeongnamnu		
	Models	Input size	Optimizer	Epochs	Input size	Optimizer	Epochs
Clear-Day	efficientnetB0	224	SGD (lr =0.05, momentum = 0.9)	20	224	SGD (lr =0.05, momentum = 0.9)	10
	mobilenet_v2	224	SGD (lr =0.05, momentum = 0.9)	20	224	SGD (lr =0.05, momentum = 0.9)	10
	resnet18	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	efficientnetB2	260	SGD (lr =0.05, momentum = 0.9)	15	260	SGD (lr =0.05, momentum = 0.9)	10
	shufflenet_v2	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	alexnet	224	SGD (lr =0.002, momentum = 0.9)	15	224	SGD (lr =0.002, momentum = 0.9)	10
	mnasnet	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	inception_v3	299	SGD (lr =0.05, momentum = 0.9)	15	299	SGD (lr =0.05, momentum = 0.9)	10
	densenet161	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	efficientnet_v2_s	384	SGD (lr =0.05, momentum = 0.9)	15	384	SGD (lr =0.05, momentum = 0.9)	20
Clear-Night	efficientnetB0	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	mobilenet_v2	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	resnet18	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	efficientnetB2	260	SGD (lr =0.05, momentum = 0.9)	15	260	SGD (lr =0.05, momentum = 0.9)	10
	shufflenet_v2	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	alexnet	224	SGD (lr =0.002, momentum = 0.9)	15	224	SGD (lr =0.002, momentum = 0.9)	10
	mnasnet	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	inception_v3	299	SGD (lr =0.05, momentum = 0.9)	15	299	SGD (lr =0.05, momentum = 0.9)	10
	densenet161	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	efficientnet_v2_s	384	SGD (lr =0.05, momentum = 0.9)	15	384	SGD (lr =0.05, momentum = 0.9)	20
Rainy-Day	efficientnetB0	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	mobilenet_v2	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	resnet18	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	efficientnetB2	260	SGD (lr =0.05, momentum = 0.9)	15	260	SGD (lr =0.05, momentum = 0.9)	10
	shufflenet_v2	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	alexnet	224	SGD (lr =0.002, momentum = 0.9)	15	224	SGD (lr =0.002, momentum = 0.9)	10
	mnasnet	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	inception_v3	299	SGD (lr =0.05, momentum = 0.9)	15	299	SGD (lr =0.05, momentum = 0.9)	10
	densenet161	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	efficientnet_v2_s	384	SGD (lr =0.05, momentum = 0.9)	15	384	SGD (lr =0.05, momentum = 0.9)	20
Rainy-Night	efficientnetB0	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	mobilenet_v2	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	resnet18	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	efficientnetB2	260	SGD (lr =0.05, momentum = 0.9)	15	260	SGD (lr =0.05, momentum = 0.9)	10
	shufflenet_v2	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	alexnet	224	SGD (lr =0.002, momentum = 0.9)	15	224	SGD (lr =0.002, momentum = 0.9)	10
	mnasnet	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	inception_v3	299	SGD (lr =0.05, momentum = 0.9)	15	299	SGD (lr =0.05, momentum = 0.9)	10
	densenet161	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	efficientnet_v2_s	384	SGD (lr =0.05, momentum = 0.9)	15	384	SGD (lr =0.05, momentum = 0.9)	10
Cloudy-Day	efficientnetB0	224	SGD (lr =0.05, momentum = 0.9)	10	224	SGD (lr =0.05, momentum = 0.9)	10
	mobilenet_v2	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	resnet18	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	efficientnetB2	260	SGD (lr =0.05, momentum = 0.9)	15	260	SGD (lr =0.05, momentum = 0.9)	10
	shufflenet_v2	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	alexnet	224	SGD (lr =0.002, momentum = 0.9)	15	224	SGD (lr =0.002, momentum = 0.9)	10
	mnasnet	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	inception_v3	299	SGD (lr =0.05, momentum = 0.9)	15	299	SGD (lr =0.05, momentum = 0.9)	10
	densenet161	224	SGD (lr =0.05, momentum = 0.9)	15	224	SGD (lr =0.05, momentum = 0.9)	10
	efficientnet_v2_s	384	SGD (lr =0.05, momentum = 0.9)	15	384	SGD (lr =0.05, momentum = 0.9)	10

IV. CONCLUSION OF EXPERIMENTAL RESULTS

A. *Performance results of each model*

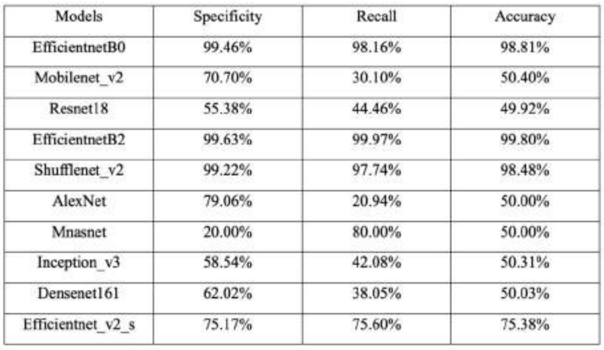
Using datasets for each environment, a normal and abnormal classification experiment was conducted on the roof of Heunginjimun and the roof of Yeongnamnu. The average values are shown in Table. 5 and Table. 6 by synthesizing the evaluation index results of each model that performed environmental classification. In the case of Heunginjimun, among ten machine learning models, EfficientnetB0, EfficientnetB2, and Shufflenet_v2 models showed excellent values of 99.61%, 99.57%, and 90.31%, respectively, based on the accuracy value. In the case of Yeongnamnu, among ten machine learning models, EfficientnetB2, EfficientnetB0, and Shufflenet_v2 showed values of 99.90%, 98.81%, and 98.48% in that order.

TABLE V
CLASSIFICATION EVALUATION METRICS VALUE OF THE MODELS FOR ROOF OF HEUNGINJIMUN



Models	Specificity	Recall	Accuracy
EfficientnetB0	99.27%	99.97%	99.61%
Mobilenet_v2	50.35%	65.60%	62.97%
Resnet18	32.43%	68.28%	55.36%
EfficientnetB2	99.16%	99.98%	99.57%
Shufflenet_v2	87.25%	93.00%	90.31%
AlexNet	29.13%	87.15%	63.81%
Mnasnet	58.71%	50.23%	59.47%
Inception_v3	75.47%	58.15%	71.81%
Densenet161	47.49%	71.23%	64.35%
Efficientnet_v2_s	56.53%	75.35%	70.94%

TABLE VI
CLASSIFICATION EVALUATION METRICS VALUE OF THE MODELS FOR ROOF OF YEONGNAMNU



Models	Specificity	Recall	Accuracy
EfficientnetB0	99.46%	98.16%	98.81%
Mobilenet_v2	70.70%	30.10%	50.40%
Resnet18	55.38%	44.46%	49.92%
EfficientnetB2	99.63%	99.97%	99.80%
Shufflenet_v2	99.22%	97.74%	98.48%
AlexNet	79.06%	20.94%	50.00%
Mnasnet	20.00%	80.00%	50.00%
Inception_v3	58.54%	42.08%	50.31%
Densenet161	62.02%	38.05%	50.03%
Efficientnet_v2_s	75.17%	75.60%	75.38%

B. *Grad-CAM result of optimal model*

To verify the machine learning experiment, Gradient-weighted Class Activation Mapping (Grad-CAM) visualization was performed to find the part that is the grounds of the Machine Learning results. Visualization results were shown in Table. 7 and Table .8. In the case of Heunginjimun Gate, it can be seen that the characteristics of the 10 areas with displacement are also reflected in the Grad-CAM results, which serves as the basis for abnormal judgment. Likewise, Yeongnamnu also, it was confirmed that the area of the five areas to which the displacement was applied was referred to as the basis for abnormal judgment.

TABLE VII
HEUNGINJIMUN GRAD-CAM RESULTS

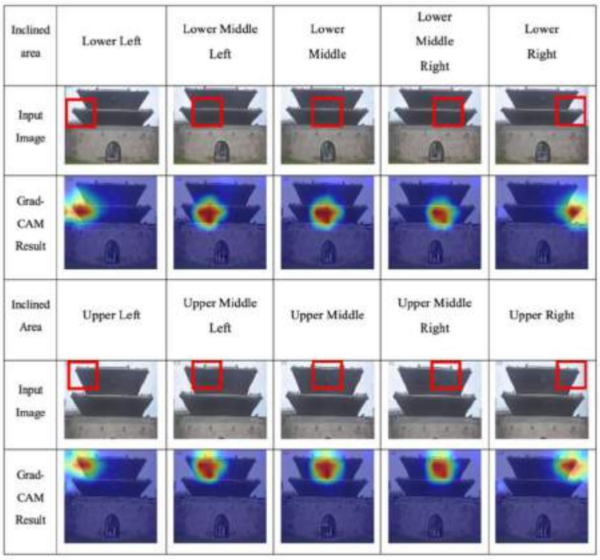
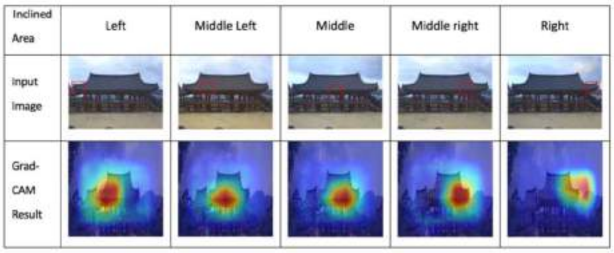

Inclined area
Lower Left
Lower Middle Left
Lower Middle
Lower Middle Right
Lower Right
Input Image
Grad-CAM Result
Inclined Area
Upper Left
Upper Middle Left
Upper Middle
Upper Middle Right
Upper Right
Input Image
Grad-CAM Result

TABLE VIII
YEONGNAMNU GRAD-CAM RESULTS


Inclined Area
Left
Middle Left
Middle
Middle right
Right
Input Image
Grad-CAM Result

V. CONCLUSION

In this paper, basic research was conducted on the development of a system for detecting anomalies in architectural, cultural properties so that abnormal symptoms can be identified or prompt initial responses can be made for the preservation of cultural properties. Heunginjimun and Yeongnamnu, treasures of the Republic of Korea, were used as subjects of basic research. CCTV images of Heunginjimun and Yeongnamnu were collected to build a dataset to determine the roof's displacement. A Deep Learning model was trained with the built data to test the abnormal detection performance so that the model can detect the degree of tilt corresponding to damage to cultural heritage.

The experimental results for 10 models on the roof of Heunginjimun were as follows. Summarizing only the top 3 models with optimal results, the EfficientnetB0, EfficientnetB2, and Shufflenet_V2 models showed excellent values of 99.61%, 99.57%, and 90.31% based on the accuracy. Specificity also showed 99.27%, 99.16%, and 87.25% values for EfficientnetB0, EfficientnetB2, and Shufflenet_V2. Finally, based on the recall, EicientnetB0, EfficientnetB2, and Shufflenet_V2 showed values of 99.97%, 99.98%, and 87.15% in order. In the case of the roof of Yeongnamnu, as with the roof of Heunginjimun, a comparative experiment was conducted on 10 models. Based on the accuracy value, the EfficientnetB0, EfficientnetB2, and Shufflenet_V2 models showed values of 98.81%, 99.80%, and 98.48%, and based on the specificity, they showed values of 99.46%, 99.63%, and 99.22%. In addition, based on the recall, values of 98.16%, 99.97%, and 97.74% were shown. As a result, the Efficientnet series models of EfficientnetB0 and EfficientnetB2 showed optimal performance for tilt detection.

Then, using Grad-CAM, we tried to check whether the EfficientnetB0 model, which represents the Efficientnet algorithm, produced results based on appropriate judgment grounds. As a result of Grad-CAM, it was confirmed that the model made a judgment based on the change in tilt displacement of the roof. The framework proposed in this experiment was proposed to perform rapid damage detection of cultural heritage and to help apply artificial intelligence, which remains at the introductory stage in managing and preserving cultural heritage. First, a normal dataset was constructed for Deep Learning and abnormal data based on normal data was generated. In addition, in order to find the most effective and universally usable Deep Learning model, we tried to derive a model that is optimal for all environments filmed by CCTV using 10 commonly used models. In conclusion, the Efficientnet series algorithm performed the best in detecting the tilt of cultural properties, which means that the Efficientnet model is the most universally effective model in situations where parameter tuning is limited. However, this study did not reflect seasonal characteristics due to an insufficient data collection period. The five environments covered the most basic environments, so learning by applying more displacement scenarios and mixed environments in future experiments is necessary. Complementing these limitations will further improve the robustness of the framework.

ACKNOWLEDGMENT

This paper was funded by the government (Cultural Heritage Administration) in 2023 with the support of the National Research Institute of Cultural Heritage (No.2021A01D06-001, Development of damage detection and alarm technology based on intelligent image analysis for safety diagnosis of Immovable cultural heritage)

REFERENCES

[1] E. C. Choi, "Cultural heritage in statistics," *Cultural Heritage Administration of the Republic of Korea, 2022*, Available:https://www.cha.go.kr/cop/bbs/selectBoardArticle.do?nttId=85272&bbsId=BBSMSTR_1020&pageIndex=1&pageUnit=10&searchCnd=&searchWrd=&ctgryLrcls=&ctgryMdcls=&ctgrySmcls=&ntcStartDt=&ntcEndDt=&searchUseYn=&mn=NS_03_07_04. Accessed 31 March 2023

[2] H.M. Kim, "Cultural heritage in statistics," *Cultural Heritage Administration of the Republic of Korea, 2023*, Available:https://www.cha.go.kr/cop/bbs/selectBoardArticle.do?nttId=82245&bbsId=BBSMSTR_1020&pageIndex=1&pageUnit=10&searchCnd=&searchWrd=&ctgryLrcls=&ctgryMdcls=&ctgrySmcls=&ntcStartDt=&ntcEndDt=&searchUseYn=&mn=NS_03_07_04. Accessed 31

[3] "Current Status of Maintenance and Maintenance of Cultural Heritage," *Korean Statistical information Service(KOSIS), 2023*, Available:https://kosis.kr/common/meta_onedepth.jsp?vwcd=MT_OTITLE&listid=150. Accessed 01 July 2023

[4] M. Mishra, T.Barman, and G.V.Ramana, "Artificial intelligence-based visual inspection system for structural health monitoring of cultural heritage," *Journal of Civil Structural Health Monitoring*, pp. 1-18, 2022, https://doi.org/10.1007/s13349-022-00643-8

[5] L.E.Mansuri, D.A. Patel,"Artificial intelligence-based automatic visual inspection system for built heritage," *Smart and Sustainable Built Environment,* vol. 11(3), pp. 622-646, 2022, https://doi.org/10.1108/SASBE-09-2020-0139

[6] M. Mishra," Machine learning techniques for structural health monitoring of heritage buildings: A state-of-the-art review and case

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

studies," *Journal of Cultural Heritage*, vol 47, pp. 227-245, 2021, https://doi.org/10.1016/j.culher.2020.09.005

[7] T.Yu, C.Lin, S. Zhang, C. Wang, X. Ding, H. An, and J.Zhang, "Artificial Intelligence for Dunhuang Cultural Heritage Protection: The Project and the Dataset," *International Journal of Computer Vision* vol.130(11), pp.2646-2673, 2022, https://doi.org/10.1007/s11263-022-01665-x

[8] A. Belhi, H. Gasmi, A. Bouras, T. Alfaqheri, A. S. Aondoakaa, A.H.Sadka, and S.Foufou, "Machine learning and digital heritage: the CEPROQHA project perspective," *Springer Singapor,* 2020, [In 4th International Congress on Information and Communication Technology: ICICT 2019 London, 2019, vol.2, pp.363-374], https://doi.org/10.1007/978-981-32-9343-4_29

[9] N. Wang, X. Zhao, P. Zhao, Y. Zhang, Z. Zou, and J. Ou, "Automatic damage detection of historic masonry buildings based on mobile deep learning," *Automation in Construction,* vol.103, pp.53-66, 2019, https://doi.org/10.1016/j.autcon.2019.03.003

[10] D. Bienvenido-Huertas, JE. Nieto-Julián, JJ. Moyano, JM. Macías-Bernal, J. Castro, "Implementing artificial intelligence in h-bim using the J48 algorithm to manage historic buildings," *Int J Archit Herit*, vol.14(8), pp.1148–1160, 2019, https://doi.org/10.1080/15583058.2019.1589602

[11] T. Bakirman, B. Kulavuz, and B. Bayram, "Use of Artificial Intelligence Toward Climate-neutral Cultural Heritage," *Photogrammetric Engineering & Remote Sensing*, vol. 89(3), pp.163-171, 2023, https://doi.org/10.14358/PERS.22-00118R2

[12] I. Garrido, J. Erazo-Aux, S. Lagüela, S. Sfarra, C. Ibarra-Castanedo, E. Pivarčiová, and P. Arias, "Introduction of deep learning in thermographic monitoring of cultural heritage and improvement by automatic thermogram pre-processing algorithms," *Sensors, vol. 21(3), pp.750*, 2021, https://doi.org/10.3390/s21030750

[13] Z. Zou, X. Zhao, P. Zhao, F. Qi, and N. Wang, "CNN-based statistics and location estimation of missing components in routine inspection of historic buildings," *Journal of Cultural Heritage*, vol. 38, pp. 221-230, 2019, https://doi.org/10.1016/j.culher.2019.02.002

[14] T. Sharma, P. Agrawal, and N. K. Verma, "Detection of dust deposition using convolutional neural network for heritage images,"*Springer Singapore*, 2019 [In Computational Intelligence: Theories, Applications and Future Directions-Volume II: ICCI-2017, *pp.*347-359], https://doi.org/10.1007/978-981-13-1135-2_27

[15] E. J. Jeong, "Heunginjimun and Heritages around the Gate,"*Art History & Cutural Heritages*, vol. 5, pp. 79-109, 2016.

[16] S. A. Park, K. W. Min, and J. S. Choi, "Ambient vibration analysis of Heunginjimun," *Journal of The Architectural Institute of Korea: Structure & Construction*, vol. 27(5), pp. 19-26, 2011.

[17] T. G. Eom, S. J. Kim, J. L. Park, H. M. Kang, and W. K. Sim, "Interpretation of Cultural Landscape at the Geumsidang (今是堂) sibigyung (12 Landscapes) in Miryang, Gyungnam," *Journal of the Korean Institute of Traditional Landscape Architecture*, vol. 29(2), pp. 1-18, 2011.

[18] S. L. Ryoo, "A Study on the Changes of the Government Pavilion, Miryang Yeongnamnu in terms of Function and Spatiality," *Journal of the Architectural Institute of Korea Planning & Design,* vol. 34(8), pp. 69-76, 2019, https://doi.org/10.5659/JAIK_PD.2018.34.8.69

[19] H. Y. Lee, "A Study on the Historic Changes of Yungnam-Ru in Historic Periods and Architectural Building Forms," *Journal of architectural history*, vol. 9(1), pp. 7-25, 2000.

[20] S. Y. Lee, H. H. Cho,"Damage Detection and Safety Diagnosis for immobable Cultural Assets Using Deep Learning Framwork," *ICACT2023*, pp. 310-313, 2023.

[21] Y. LeCun, Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks* 3361(10), 1995.

[22] J. Wu, "Introduction to convolutional neural networks," *National Key Lab for Novel Software Technology*, vol. 5(23), pp. 495, 2017, Nanjing University. China.

[23] R. Chauhan, K. K. Ghanshala, and R. C. Joshi, "Convolutional neural network (CNN) for image detection and recognition," *2018 first international conference on secure cyber computing and communication (ICSCCC 2018),* pp. 278-282, 10.1109/ICSCCC.2018.8703316.

[24] M. Jogin, M. S. Madhulika, G. D. Divya, R. K. Meghana, and S. Apoorva, "Feature extraction using convolution neural networks (CNN) and deep learning," *2018 3rd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT)*, pp. 2319-2323], 10.1109/RTEICT42901.2018.9012507.

[25] C. D. James, J. B. Aimone, N. E. Miner, C. M. Vineyard, F. H. Rothganger, K. D. Carlson, and S. J. Plimpton, "A historical survey of algorithms and hardware architectures for neural-inspired and neuromorphic computing applications," *Biologically Inspired Cognitive Architectures* vol. 19, pp.49-64, 2017, https://doi.org/10.1016/j.bica.2016.11.002.

[26] M. Tan, Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *International conference on machine learning 2019*, pp. 6105-6114, 2019, May.

[27] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv, 2017,* preprint arXiv:1704.04861, https://doi.org/10.48550/arXiv.1704.04861.

[28] K. He, X. Zhang, S. Ren, and J. Sun, " Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition 2016,* pp. 770-778, 2016.

[29] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," *2018 Proceedings of the European conference on computer vision (ECCV),* pp. 116-131,2018.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60(6), pp. 84-90, 2017, https://doi.org/10.1145/3065386

[31] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "Mnasnet: Platform-aware neural architecture search for mobile," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019,* pp. 2820-2828.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *Proceedings of the IEEE conference on computer vision and pattern recognition 2019*, pp. 2818-2826.

[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, (2017) "Densely connected convolutional networks," *Proceedings of the IEEE conference on computer vision and pattern recognition 2017,* pp. 4700-4708.

Sang-Yun Lee (B'94–M'96–D'07) was born in South Korea 1971, is a principal researcher at Police Science & Public Safety ICT Research Center of Digital Convergence Research Laboratory in ETRI. He has been working at ETRI since 1999. In 2008, he received Ph.D. in Electronics and Telecommunications Engineering at the University of Hanyang(Rep. of Korea). He has been developing technologies in the fields of Broadcasting Communication, System Software, Embedded Software, Artificial Intelligence, and etc.

His main research interests have been in Computational Sciences. Currently, he is involved in developing a technique to detect displacement using Artificial Intelligence technology for CCTV images of cultural assets. He also leads several projects including disaster management for cultural heritage. Since 2016, He acts as an editor of the Study Group 16 (SG16) in the ITU-T and has been developing international standards. He is an international standards expert at TTA and a member of the Korean ITU-T Research Committee. He has been carrying out more than 10 government-funded projects. He is author of more than 70 scientific papers and has registered more than 10 patents.

Daekyeom Lee (B' 16 – M'18) is a Lead of AI Convergence Technology Research Team at Season. From 2021 to present, he has been working at Season Co., Ltd. He graduated from Dongguk University (Rep. of Korea) with a double major in energy systems and economics in 2016 and received a master's degree in engineering from Sejong University (Rep. of Korea) in 2018. Since 2018, he has conducted research in statistics, data analysis, and artificial intelligence.

His main research interest was big data analysis and data science. Additionally, after joining Season Co., Ltd., he has been working as a team leader developing artificial intelligence technology in terms of data utilization. He mainly works on the application of artificial intelligence technology using manufacturing data, object detection using deep learning, and improvement of learning speed through transfer learning. In addition, he is conducting joint research on technology utilization in cooperation with related research institutes and companies and has authored 8 domestic and foreign journals.

A Study on Connectivity Evaluation Among Peer Groups in Pure P2P Networks

Yutaka Naito*, Takumi Uemura*, Takashige Hoshiai**

**Faculty of Computer and Information Sciences, Sojo University, 4-22-1 Ikeda, Nishi-ku, Kumamoto, 860-0082 Japan*
Sojo University IoT/AI Center, 4-22-1 Ikeda, Nishi-ku, Kumamoto, 860-0082 Japan
naito@cis.sojo-u.ac.jp, t_uemura@cis.sojo-u.ac.jp, hoshiai_takashige@yahoo.co.jp

***Abstract*— Recently P2P (peer-to-peer) network garners attention as a technology for developing a peer group by use of connection of peers which function as autonomously distributed and cooperative units virtualized from computer resources. In this paper, we focus on pure P2P networks which form peer groups by connecting peers without the need for intermediates. We execute performance evaluation by using computer simulation and propose a method to measure peer groups' connectivity utilizing mean number of connected peer groups without using peers' arrival and departure rates on peer groups.**

***Keyword*— P2P, Pure model, Cluster model, Performance evaluation, Peer groups' connectivity**

I. INTRODUCTION

RECENTLY P2P (peer-to-peer) network garners attention as a technology for developing a peer group by use of connection of peers which function as autonomously distributed and cooperative units virtualized from computer resources. JXTA [1], [2], SOBA [3], [4], and SIONet [5]-[8] are representative examples. Lately a P2P technology called blockchain [9] has attracted particular attention.

P2P network is classified into two types: pure model (flat model) [10], in which there is no intermediate to connect peer groups, and cluster model [11], in which there is an intermediate.

In the cluster model, an intermediate always connects peer groups, so there is no fragmentation among peer groups.

On the other hand, in the pure model, no intermediate exists to connect peer groups, so any peer connects peer groups by autonomously connecting with peers in other peer groups. Therefore, when a peer connecting peer groups departs from a peer group, the peer groups are broken up into fragments. As a result, it becomes difficult to share information among peer groups. Thus, in the pure model, it is important to evaluate connectivity among peer groups from the perspective of information sharing among peer groups [12], [13].

In this paper, we define connectivity among peer groups (%) as follows.

$$\frac{\text{Maximum number of connected peers}}{\text{Total number of peers}} \times 100 \qquad (1)$$

In other words, the connectivity indicates how many percent of the peers are connected among all the peers, and no fragmentation among peer groups occurs when the connectivity is 100%.

While it is considered important to quantitatively evaluate the connectivity among peer groups in the pure model, some research papers [10], [12], [14]-[17] report on the evaluation of the connectivity among peer groups. However, in all of them, it is necessary to survey in advance "the arrival rate of peers in a peer group" and "the departure rate from a peer group," which are input parameters in the evaluation model of the connectivity. However, in this case, it is a big burden for the evaluators to obtain the arrival rate and the departure rate in advance by the actual survey.

Therefore, in this paper, we examine the performance of the evaluation model using computer simulations. We propose a method to calculate the connectivity among peer groups by using the mean number of connected peer groups without using the arrival and departure rates of peers as input parameters of the evaluation model [18]-[20].

The rest of this paper is organized as follows. The mechanism of the pure model is explained in section II. Next, in section III, we introduce a Community Coexistence Society (CCS), as an example of the pure model application to realize a sustainable welfare society in Japan. In section IV, we explain the performance evaluation model (simulation model) and scale to obtain the connectivity among peer groups. In section V, we show the results of the performance evaluation by simulation and discuss the results. Section VI concludes the paper and discusses future work.

II. MECHANISM OF THE PURE MODEL

In this chapter, we describe the mechanism for connecting peer groups in the pure model. In figure 1(a) and figure 1(b), we consider a virtualized peer as the minimum unit of autonomous distributed cooperation and a peer group formed as a set of peers.

To deepen the understanding of the pure model, we discuss the cluster model, which is a contrast to the pure model. In the cluster model, peer groups are connected to each other through an intermediate, as shown in figure 1(a). In this method, there is no fragmentation between peer groups unless

Manuscript received Oct. 10, 2022.

This work is a follow-up of the invited journal to the accepted & presented paper entitled "A Study on Connectivity Evaluation Among Peer Groups in Pure P2P Networks" of the 25th International Conference on Advanced Communication Technology (ICACT2023).

Yutaka Naito is with Sojo University, Kumamoto, 860-0082 JAPAN (corresponding author e-mail: naito@cis.sojo-u.ac.jp).

Takumi Uemura is with Sojo University, Kumamoto, 860-0082 JAPAN (e-mail: t_uemura@cis.sojo-u.ac.jp).

Takashige Hoshiai is with Sojo University IoT/AI Center, Kumamoto, 860-0082 JAPAN (e-mail: hoshiai_takashige@yahoo.co.jp).

Copyright ⓒ 2023 GiRI (Global IT Research Institute)

the intermediate goes down, but the operating cost of the intermediate is often burdensome. Meanwhile, in the pure model, as shown in figure 1(b), peer A and peer B, which belong to peer group III, simultaneously participate in (belong to) peer group II and peer group I, respectively, so that peer α and peer Y can be connected through peer A and peer B. As a result, peer group I, peer group II, and peer group III can be connected. On the other hand, when peer A and/or peer B departs the peer group, the peer group may be fragmented, and cooperation and information sharing among peer groups may be disrupted.

Thus, although the pure model has the problem of fragmentation, it is unique in that it connects peer groups in a "loose," "flexible," and "autonomous distributed cooperation" form without the need for an intermediate, and it is widely used in a variety of fields because it does not require the operating costs of an intermediate.

There are many studies on fragmentation among peer groups. As examples of peer's departure from peer groups as a cause of fragmentation, a hardware failure of peers [14] and interruption of file exchange services [15] causes departure of peers from peer groups. Also, blockchain mining cost incurs peers' departure from a peer group and produces fragmentation of peer group [16], [17]. Furthermore, many examples of the application of the pure model concept to inter-operational and inter-organizational collaboration in real-world business models have been reported. For example, there is a reference [21] that reports that when bank employees (peers) multitask among departments (peer groups), such as the teller service department and the loan counselling department at a bank, fragmentation occurs when employees depart from the department. Furthermore, reference [13] reports that hotel employees (peers) simultaneously participate in (belong to) different departments such as the front desk and the guest room cleaning (peer groups) to connect departments and that fragmentation occurs when hotel employees depart the department.

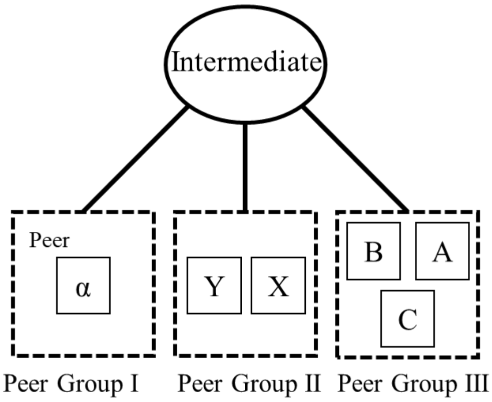


Fig. 1(a). Peer group connecting method (cluster model)

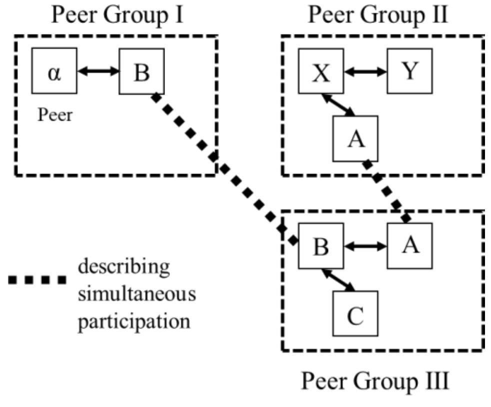


Fig. 1(b). Peer group connecting method (pure model)

Although there is a problem of fragmentation, the trend towards applying the pure model, which does not require high operating costs for intermediates, to inter-organizational collaboration is notable in business categories involving customer service, such as bank tellers and restaurants, and in labor-intensive industries, such as hotels, lodging, nursing care, and so on. It has been reported that 24.4% of hotels and inns have introduced the pure model in recent years and have achieved positive results [22]. While these case studies are reported, in the next chapter we introduce CCS as an example using pure model inter-organizational collaboration in Japan.

III. AN EXAMPLE OF THE PURE MODEL USE

In this chapter, we introduce CCS as an example of use for improving the understanding of the pure model inter-organizational collaboration. In recent years, the number of single-person households in Japan has been increasing due to the aging of the population and the change in customs, such as parents and children living separately to respect privacy, and the isolation of individuals due to the reduction of welfare and medical services such as nursing care caused by the lack of financial resources due to the long-term economic stagnation has become a social problem.

So far, to prevent isolation, local public support organizations have provided information to residents by visiting households and making phone calls to confirm the safety of residents and to introduce them to consultation services. However, the number of users has decreased due to the declining population, and it has also become difficult to secure professional human resources, making it difficult to stably operate local public support organizations.

Therefore, the Japanese government advocates the establishment of a CCS in which information is shared by connecting groups such as public support providers, local businesses, local volunteer groups, and the homes of local residents in the pure model to prevent the isolation of individuals in the community [23].

In the past, public support has been based on information sharing in the form of the cluster model using an intermediate. For example, in the cluster model, as shown in figure 2(a), a facility for the disabled has members such as a staff and disabled persons who receive services. Likewise, in a agricultural corporation, there are a staff and laborers, and in a foreigner support volunteer organization, there are a staff and foreigner receiving support. In general, representative meeting is held for the purpose of cooperation and information sharing among the organizations, for example, once a month. This makes it possible for organizations to cooperate and share information with each other, and there is no fragmentation among the organizations unless the representative meeting is cancelled. Note that the representative meeting, organization, and members correspond to the intermediate, peer group, and peers, respectively, in figure 1(a).

Copyright ⓒ 2023 GiRI (Global IT Research Institute)

However, the method of cooperation and information sharing among organizations through representative meeting has the following issues.

Issue 1: Delay in information transfer

Information on other organizations obtained by a staff member at a representative meeting is transferred from the staff member to the other members, which requires time for information sharing, resulting in a time delay.

Issue 2: Difficulty in creating a sense of ownership

Information about other organizations relayed by staff members tends to become someone else's business for members and members tend to lack a sense of ownership.

Issue 3: Overloading of staff members

Information is gathered by staff members who attend to a representative meeting, and the burden of information transmission is concentrated on them. Mental and physical exhaustion and employee turnover due to the concentration of workload on staff members have already become a social problem in the nursing care and medical fields [24].

In contrast, the peer group connecting method in the pure model, which is being used to realize a CCS, members of a facility for the disabled autonomously and simultaneously participate in or depart from an agricultural corporation or a foreigner support volunteer organization, and voluntarily share information with the members of the participating organizations. For example, as shown in figure 2(b), when disabled person C participates in the facility for the disabled and the agricultural corporation at the same time, and when laborer Y participates in the agricultural corporation and the foreigner support volunteer organization at the same time, the facility for the disabled, the agricultural corporation, and the foreigner support volunteer organization are connected as a result. Thus, collaboration and information sharing among the organizations become possible. Here, note that each member and each organization correspond to a peer and a peer group, in figure 1(b), respectively.

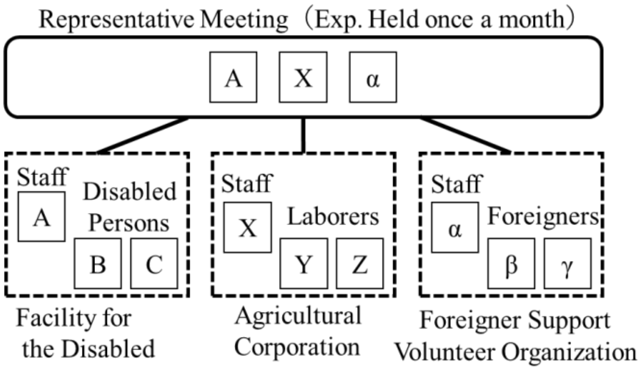

Representative Meeting（Exp. Held once a month）
A
X
α
Staff
A
Disabled Persons
B
C
Staff
X
Laborers
Y
Z
Staff
α
Foreigners
β
γ
Facility for the Disabled
Agricultural Corporation
Foreigner Support Volunteer Organization

Fig. 2(a). Collaboration and information sharing among organizations through representative meeting

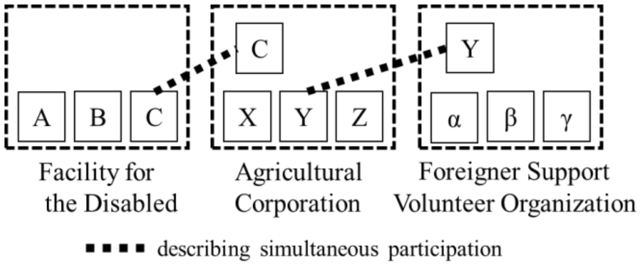

C
Y
A
B
C
X
Y
Z
α
β
γ
Facility for the Disabled
Agricultural Corporation
Foreigner Support Volunteer Organization
▪▪▪▪ describing simultaneous participation

Fig. 2(b). Collaboration and information sharing among organizations in the pure model

This pure model solves the above three issues seen in the cluster model as follows.

Solution to issue 1: Speeding up information transfer

Compared to the representative meeting held periodically, the pure model has the advantage that information sharing is less likely to be delayed because the facilities for the disabled and the agricultural corporation are constantly connected through disabled person C.

Solution to issue 2: Creating ownership of information

Disabled person C, who belongs to a facility for the disabled, also belongs to an agricultural corporation, so he/she can have a sense of ownership of the information about the agricultural corporation.

Solution to issue 3: Reduction of staff workload

The information can be shared within the facility for the disabled by having C, who belongs to the facility for the disabled, simultaneously participating in the agricultural corporation and sharing the information obtained in the agricultural corporation without the involvement of staff A. This type of collaboration between the disabled and the agricultural corporation is widely called "agricultural welfare collaboration" and has been attracting attention in recent years as a place where the disabled can play an active and autonomous role [25]. In realizing a CCS, it is important for various entities including individuals to autonomously participate in the activities of not only one organization but also various organizations called “second place” and “third place,” and the Japanese government is actively promoting the introduction of peer group collaboration in the community using the pure model [23].

On the other hand, a structural problem of the pure model, the connectivity decreases due to the fragmentation among the peer groups. For example, in figure 2(b), when a disabled person (peer) C departs from the agricultural corporation (peer group), the facility for the disabled peer group is separated from the peer group of the agricultural corporation and the foreigner support volunteer organization. In other words, information sharing among peer groups is disrupted, and at the same time, peers A, B, and C are isolated from other peers.

Therefore, in the pure model, quantitative evaluation of connectivity is important to prevent fragmentation. In previous studies, information on participation in and departure from each peer group (organization) by each peer (member) was collected over a long period of time to determine the arrival rate and the departure rate in advance [10], [12], [14]-[17], and computer simulations were used to calculate the connectivity among peer groups, which caused an excessive research cost.

Therefore, in this paper, we examine the performance of the evaluation model using computer simulations. We propose a method to calculate the connectivity among peer groups by using the mean number of connected peer groups without using the arrival and departure rates of peers as input parameters of the evaluation model [18]-[20]. The mean number of connected peer groups is the mean number of peer groups in which any peer participates at the same time.

The reason for using the mean number of connected peer groups as an input parameter for the evaluation model instead of the arrival and departure rates is its convenience. It is relatively easier to obtain the mean number of connected peer groups than to obtain the arrival and departure rates by a measurement survey. For example, in a questionnaire survey, the question "How many groups, on average, did you belong to?" is easier to answer than "How often did you join in and depart from each group?" Hence, the proposed method requires less effort and cost to obtain the information needed to calculate the connectivity.

The next section describes the performance evaluation model used in simulations to calculate the connectivity using the proposed method.

IV. PERFORMANCE EVALUATION MODEL

In this section, we describe a performance evaluation model for quantitatively evaluating the connectivity among peer groups in the pure model.

Figure 3 shows the example of the performance evaluation model. This figure shows peer groups I to III formed by virtualized peers A to D as the minimum unit of autonomous distributed cooperation.

By peer B belonging to peer group I participates in peer group II, peer B participates in peer groups I and II simultaneously. As the same way, by peer C participates in peer groups II and III, peer A and peer D are connected via peer B and peer C. As the result of the connection among peer A, peer B, peer C and peer D, peer group I, peer group II, and peer group III are connected.

Thus, all peers independently participate in all peer groups with an arrival rate λ. Here, figure 3 shows the situation where peer B belonging to peer group I tries to participate in peer group II and peer group III with an arrival rate λ. As a result, it participates only in peer group II. We assume all peers in a peer group are always connected to each other. Therefore, we do not deal with the connection topology between peers in a peer group in this paper.

On the other hand, all peers independently depart from all peer groups to which they belong with a departure rate μ. For example, in figure 3, if peer B departs from peer group II, a fragmentation occurs between peer groups II-III, and peer group I. In this paper, for the sake of simplicity, we assumed that the arrival rate of each peer is all the same. And the departure rate of each peer is also all the same. We will discuss the performance evaluation model having different arrival and departure rate on each peer in the future work.

We explain the parameter and the performance evaluation scale on the performance evaluation model.

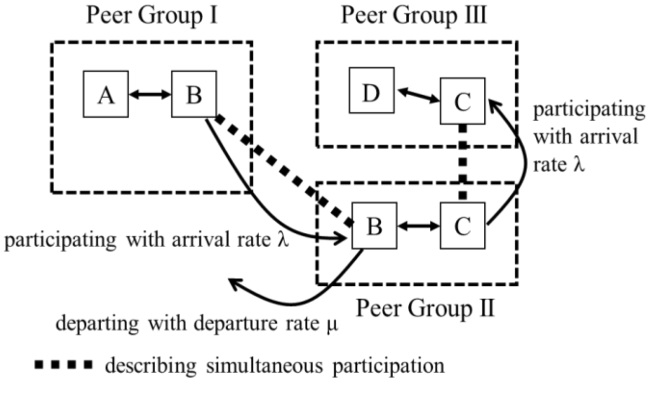

Peer Group I
Peer Group III
A
B
D
C
participating with arrival rate λ
participating with arrival rate λ
B
C
departing with departure rate μ
Peer Group II
▪ ▪ ▪ ▪ describing simultaneous participation

Fig. 3. Performance evaluation model

- The number of peer groups m: Total number of peer groups.
- The total number of peers n: The total number of peers, i.e., the sum of all peers participating in m peer groups. For example, n=4 in figure 3.
- Arrival rate λ: The arrival rate of each peer in each peer group. The mean number of times a peer participates in each peer group per unit of time. It is equal to the reciprocal of the mean time of participation in each peer group.
- Departure rate μ: The mean number of times that a peer departs from each peer group per unit of time. It is equal to the reciprocal of the mean time of sojourn in each peer group.
- Utilization rate ρ: Ratio of arrival rate to departure rate ($\rho = \lambda / \mu$)
- Maximum number of connected peers L: The maximum number of connected peers at a given point in time.
- Connectivity S: Percentage of peers that are connected among peer groups. $E\{L\}$ is the mean value of L, obtained by simulation. S is calculated from the following formula.

$$S = \frac{E\{L\}}{n} \times 100(\%) \qquad (2)$$

- The number of simultaneously participated peer group on each peer K_i: The number of peer groups that the peer p_i (i=1 to n) is participating at the same time at a given point in time.
- The number of simultaneously participated peer group K: The mean number of K_i at a given point in time.
- The mean number of connected peer groups $E\{K\}$: The mean number of peer groups in which a peer participates at the same time. $E\{K\}$ is the mean value of K over the observation period, calculated by simulation.
- Peer's simultaneous participation rate R: The number of peer groups in which a single peer participates at the same time.

$$R = \frac{E\{K\}}{m} \times 100\ (\%) \qquad (3)$$

For example, figure 4 shows a situation where there is a fragmentation among peer groups when m=3 and n=6. In peer group I, the number of connected peers is 2 because the p_1 and p_2 are connected. On the other hand, in peer group II and peer group III, the number of connected peers is 4

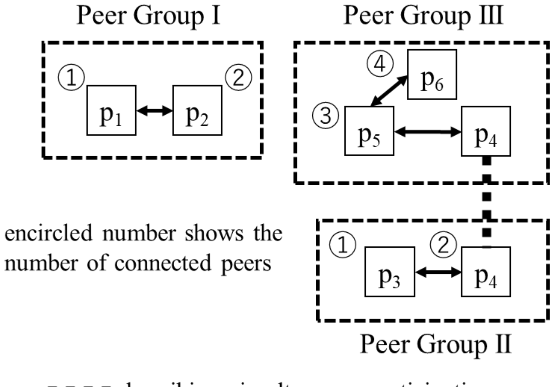

Peer Group I
Peer Group III
①
②
p1
p2
④
p6
③
p5
p4
encircled number shows the number of connected peers
①
②
p3
p4
Peer Group II
▪ ▪ ▪ ▪ describing simultaneous participation

Fig. 4. Example of maximum number of connected peers (L) and number of connected peer groups (Ki).

Copyright ⓒ 2023 GiRI (Global IT Research Institute)

because p3, p4, p5 and p6 are connected. Therefore, the maximum number of connected peers L is 4. Moreover, the number of simultaneously participated peer group on each peer Ki, which is the number of peer groups in which pi participates simultaneously, is K1 =1, K2 =1, K3 =1, K4 =2, K5 =1 and K6 =1.

V. SIMULATION RESULTS AND DISCUSSION

In this section, we clarify that the connectivity S is calculated without using the arrival and departure rates (i.e., participation and departure information) from simulation results.

The convergence condition of the simulation is set to be within 3% difference from the previous $E\{L\}$ value.

Therefore, the convergence condition is given by the following formula where the time until convergence is the observation period.

$$\frac{\left| E\{L\}^{(t+\Delta t)} - E\{L\}^{(t)} \right|}{E\{L\}^{(t)}} < \varepsilon \qquad (4)$$

Where t is any time in the simulation, Δt is the minimum unit time, and ε is the threshold.

The simulation was based on the performance evaluation model shown in figure 3. The simulator is implemented in C language, and the specification of the computer used for the simulation is presented in Table 1. The time required to obtain a single plot (point) in the graph was approximately 10 seconds, and the confidence interval ±5% was obtained by running the simulation three times for each plot in 95% confidence interval.

To ensure sufficient coverage, the number of peer groups m was set to m=10 and m=30, and the total number of peers n was set from n=30 to n=1,500.

Figure 5 shows the simulation results for m=10 and n=30. For both the utilization rate ρ=0.4 and ρ=0.04, the connectivity S does not depend on the departure rate μ, but on ρ, i.e., the ratio of λ to μ. Therefore, once ρ is determined, the connectivity S is independent of the departure rate μ. Therefore, figure 5 describes that robustness to ρ holds for the connectivity S.

Next, figure 6 shows the relationship between the departure rate μ and the mean number of connected peer groups $E\{K\}$ for m=10 and n=30, where $E\{K\}$ is the mean number of peer groups in which a peer is participating at the same time. Figure 6 shows that $E\{K\}$ does not depend on the departure rate μ, but on ρ, i.e., the ratio of λ to μ, for both the utilization rate ρ=0.4 and ρ=0.04. Therefore, once ρ is determined, $E\{K\}$ is uniquely determined. Figure 6 describes that robustness with respect to ρ holds for the mean number of connected peer groups $E\{K\}$.

Next, figure 7 and figure 8 show the simulation results of the departure rate μ and the peer's simultaneous participation rate R for n/m=3 (m=10, n=30 and m=30, n=90) and n/m=50 (m=10, n=500 and m=30, n=1,500), respectively, for the utilization rate ρ = 0.4 and ρ = 0.04. From the respective figures, the peer's simultaneous participation rate R is dependent on the utilization rate ρ, since the same utilization rate ρ results in the same peer's simultaneous participation rate R.

Table 1. Simulation execution environment

Computer	Mac Book Pro (Late2013)
OS	OS X version 10.9.4
CPU	Intel Core i7 2.0 GHz
Memory	8GB (DDR3 1,600 MHz)
HDD	250 GB

Figure 9 is constructed by combining figures 5 and 6. Figure 9 describes that even without obtaining λ and μ, the mean number of connected peer groups $E\{K\}$ provides the connectivity S. For example, if each peer participates in 3 peer groups on average, then the connectivity S is approximately 100 %. This means that the derivation of λ and μ (setting of input parameters) by actual measurement is not necessary.

Finally, figure 10 compares n/m=3 (m=10, n=30 and m=30, n=90) and n/m=50 (m=10, n=500 and m=30, n=1,500) in the relationship between the mean number of connected peer groups $E\{K\}$ and connectivity S. The figure shows that the connectivity S depends on n/m, since the connectivity S is the

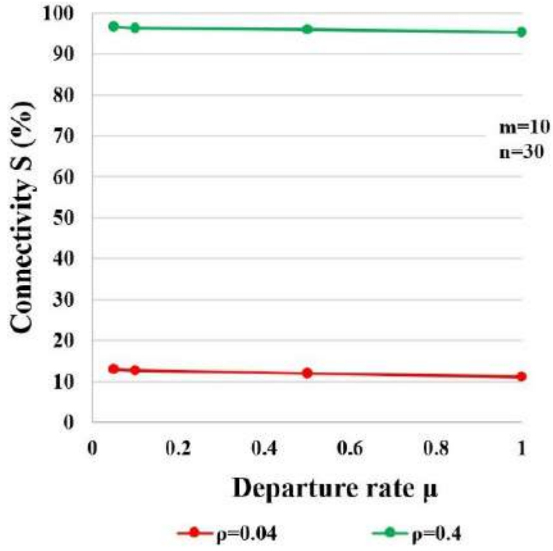


Fig. 5. Robustness to ρ on the Connectivity S

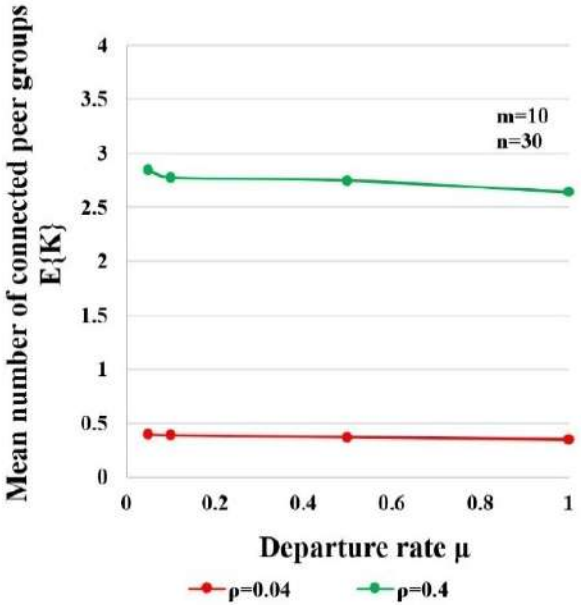


Fig. 6. Robustness to ρ on the Mean number of connected peer groups $E\{K\}$

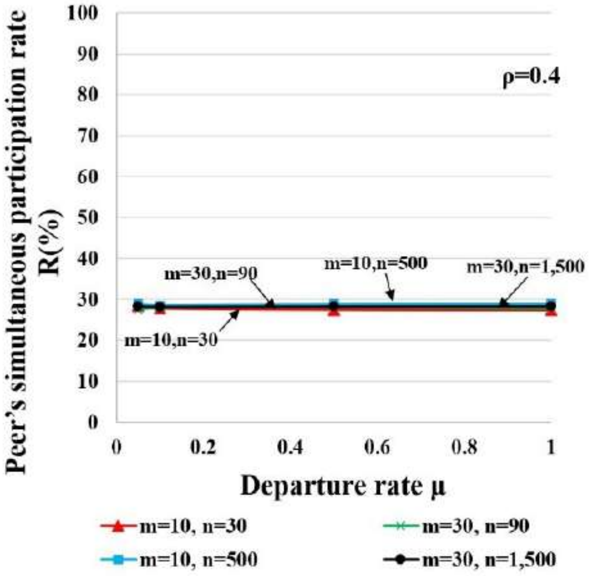


Fig. 7. Robustness to ρ on the peer's simultaneous participation rate R (ρ=0.4)

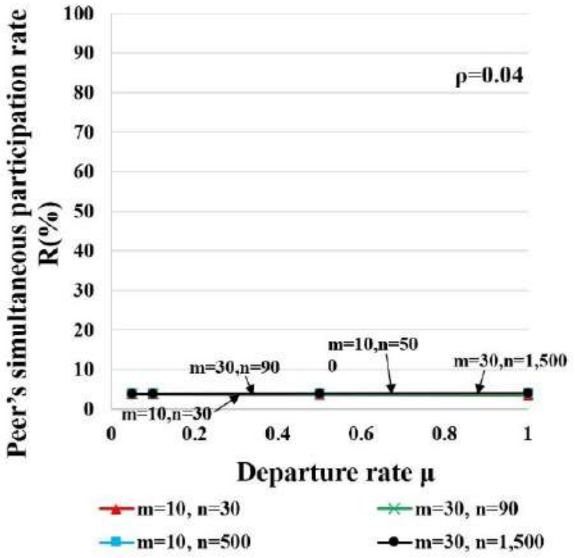


Fig. 8. Robustness to ρ on the peer's simultaneous participation rate R (ρ=0.04)

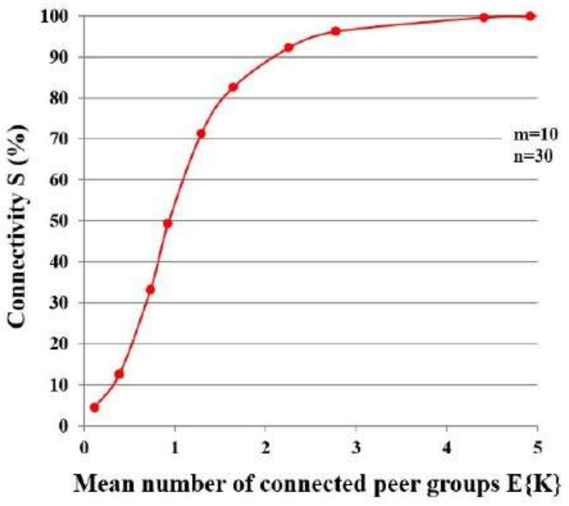


Fig. 9. Mean number of connected peer groups $E\{K\}$ vs. connectivity S

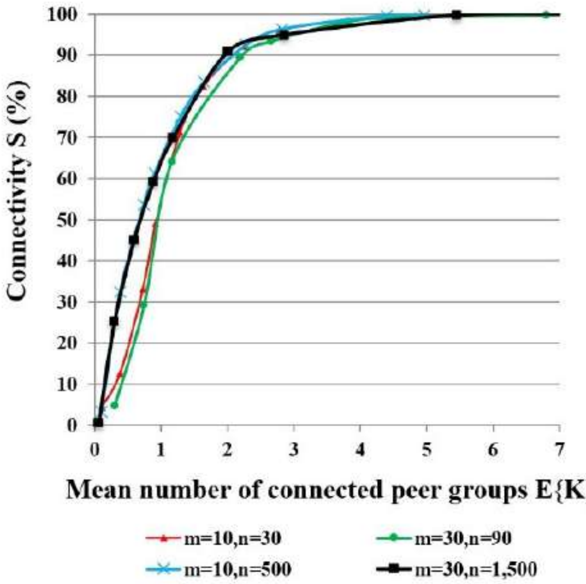


Fig. 10. Mean number of connected peer groups $E\{K\}$ vs. connectivity S (n/m=3, n/m=50)

same when n/m is the same. The larger n/m is, the faster the convergence of the connectivity S to the mean number of connected peer groups $E\{K\}$ becomes.

The simulation results provide the following findings.

- Even without investigating λ and μ as input parameters, the connectivity S is calculated from $E\{K\}$.
- The connectivity S depends not only on ρ, but also on the ratio of n to m (n/m). Therefore, if n/m is the same, the result will be the same.
- The convergence of the connectivity S with respect to the mean number of connected peer groups $E\{K\}$ is faster when n/m is large.
- Once n and m are determined, or once the ratio of n to m is determined, the mean number of connected peer groups $E\{K\}$, is obtained from the connectivity S. For the instance, figure 9 shows that the mean number of connected peer groups $E\{K\}$, is 2 to satisfy 90% of the connectivity S. In other words, a peer should participate in two peer groups in average.
- Once m, S, and $E\{K\}$ are determined, the required number of peers n is derived.
- The peer's simultaneous participation rate R depends on ρ.

VI. CONCLUSION

In this paper, we clarified through simulations that the peer connectivity is calculated without using the arrival rate or the departure rate in the pure model. Specifically, we found that the connectivity can be obtained from the average number of peer groups in which peers participate at the same time. Since the conventional method of evaluating the connectivity requires a large amount of time and effort to derive the necessary arrival and departure rates, we conclude that we can derive connectivity in a relatively short time through the quantitative evaluation of the connectivity.

In the future, we plan to evaluate simulations with a performance evaluation model that considers combination of

Copyright ⓒ 2023 GiRI (Global IT Research Institute)

the peers and peer groups having various arrival and departure rates, i.e., the participation rate λ_{ij} of peer$_i$ (i=1,2,3...n) in peer group$_j$ (j=1,2,3...m) and the departure rate μ_{ij}.

References

[1] L. Gong, "JXTA: a network programming environment," *IEEE Internet Computing*, Vol. 5, No. 3, pp. 88-95, DOI: 10.1109/4236.935182, 2001.

[2] E. Halepovic, and R. Deters, "JXTA performance study," *2003 IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, Vol. 1, pp. 149-154, DOI: 10.1109/PACRIM.2003.1235740, 2003.

[3] N. Yoshida, S. Urashita, Y. Hayashi et al., "SOBA framework: an application framework for broadband network environment," *The 2005 Symposium on Applications and the Internet*, pp. 296-303, DOI: 10.1109/SAINT.2005.59, 2005.

[4] SOBA Project Inc., "What is SOBA?," Available: https://www.soba-project.com.

[5] T. Hoshiai, K. Koyanagi, K. Sukhbaatar Birke, M. Kubota, H. Shibata and T. Sakai, "Semantic Information Network Architecture," *IEICE Transactions on Electronics, Information and Communication Engineers (B)*, Vol. J84-B, No. 3, pp. 411-424, 2001.

[6] T. Hoshiai, *Brokerless model and SIONet*, Telecommunications Association of Japan (Ohmsha), 2003.

[7] T. Hoshiai, "General Theory of P2P [I]: Challenge of Brokerless Model," *IEICE Journal*, Vol.87, No.9, pp.804-811, 2004.

[8] SIONet (NTT Information Communication Glossary), Available: https://www.ntt-review.jp/yougo/word.php?word_id=1928.

[9] J. Kishigami, S. Fujimura, D. Watanabe, M. Ohashi, and A. Nakahira, *Introduction to Blockchain Technology*, Morikita Publishing, 2017.

[10] Y. Kitahashi, Y. Hoshiai, H. Mitomo and T. Hoshiai: "A Proposal of Decentralized Collaboration Architecture on Brokerless Network and its Evaluation," *Transactions of Information Processing Society of Japan*, Vol. 47, No. 8, pp. 2669-2683, 2006.

[11] T. Hoshiai, Y. Kitahashi, Y. Hoshiai, T. Harada and H. Mitomo, "Performance Evaluation on Brokerless Networking Architecture," *IEICE Transactions on Electronics, Information and Communication Engineers (D)*, Vol. J88-D-I, No. 11, pp. 1608-1621, 2005.

[12] R. Cohen, K. Erez, D. Ben-Avraham and S. Havlin, "Resilience of the Internet to random breakdowns," *Phys. Rev. Lett.*, Vol. 85, No. 21, pp. 4626-4628, 2000.

[13] M. Nakauchi, "Facilitators of knowledge transfer among engineers: from the perspective of information acquirers," *Organization Science*, Vol. 48, No. 2, pp. 61-73, 2014.

[14] S. Saroiu, P. K. Gummadi and S. D. Gribble, "A measurement study of peer-to-peer sharing systems," *Proc. Multimed. Comput. Netw*, 2002 (MMCN '02), pp. 156-170, 2002.

[15] K. Leibnitz, T. Hossfeld, N. Wakamiya and M. Murata, "Peer-to-peer vs. client/server: Reliability and efficiency of a content distribution service," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 4516, pp. 1161-1172, 2007.

[16] S. G. Motlagh, J. Misic and V. B. Misic, "Impact of Node Churn in the Bitcoin Network," *IEEE Transactions on Network Science and Engineering*, Vol. 7, No. 3, pp. 2104-2113, 2020.

[17] M. A. Imtiaz, D. Starobinski, A. Trachtenberg and N. Younis, "Churn in the Bitcoin Network: Characterization and Impact," *2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pp. 431-439, DOI: 10.1109/BLOC.2019.8751297, 2019.

[18] Y. Naito, C. Katsuki, N. Suehiro and T. Hoshiai, "Regional activation based on P2P network architecture," *2016 International Symposium on Nonlinear Theory and Its Applications*, pp. 423-426, 2016.

[19] Y. Naito, T. Hoshiai, K. Yoshimi, "Research of Evaluation on Regional Resource Network Model for Innovation Emergence," *Proc. of the 79th National Conference of Japan Society for Information and Management [Autumn]*, pp. 215-218, 2019.

[20] Y. Naito, T. Uemura and T. Hoshiai, "A Study on Connectivity Evaluation Among Peer Groups in Pure P2P Networks," *25th International Conference on Advanced Communication Technology (ICACT)*, pp. 298-302, DOI: 10.23919/ICACT56868.2023.10079579, 2023.

[21] Shinkin Central Bank, "Financial Research Information 2020-8 Trends in multitasking by sales branch staff in Shinkin Banks," *Management Strategy*, Vol. 32, pp. 3, 2020.

[22] Japan Tourism Agency, "Grandia Housen," *Case Studies on Productivity Improvement in the Lodging Industry*, Vol. 3, pp. 16, 2020. Available: http://www.shukuhaku-kaizen.com/wp-content/themes/shukuhaku_kaizen/img/case_studies_2020_01.pdf

[23] Japan Gerontological Evaluation and Research Institute, "Survey and Research Project on Outcome Indicators for Realization of Community Coexistence Society," *Examination of Process Evaluation for Establishment of Comprehensive Support System*, 2020.

[24] Nihon Keizai Shimbun, "Long Working Hours in Nursing Homes, 70% of Nursing Homes Have Two Shifts and Night Shifts Over 16 Hours," 2018. Available: https://www.nikkei.com/article/DGXMZO29431530W8A410C1CR8000/

Yutaka Naito was born in Japan, 1969. He received the Ph.D. degree in Engineering from Graduate School of Engineering, Sojo University, 2022. He is currently engaged in research on P2P network technology for regional revitalization as an assistant professor in the Faculty of Computer and Information Sciences, Sojo University.

Takumi Uemura was born in Japan, 1980. He received the Ph.D. degree in Engineering from Graduate School of Science and Technology, Kumamoto University, 2011. He is currently engaged in research on image processing and pattern recognition as an associate professor in the Faculty of Computer and Information Sciences, Sojo University.

Takashige Hoshiai was born in Japan, 1962. He received the Ph.D. degree in Engineering. He was a visiting researcher at Bell Telephone Laboratories from 1995 to 1997, and proposed the brokerless model in 1998, and invented the semantic information network architecture SIONet, which is the technology to realize the model. In 2011, he proposed Social Community Brand (SCB theory) that utilizes P2P for local revitalization. He is currently conducting research on regional revitalization and the emergence of regional innovation using SCB theory. In addition, he and his team invented a method of innovation emergence based on the concept of the board game GO. He is currently a president of Sojo University IoT/AI Center, a professor of the Faculty of Computer and Information Sciences, Sojo University, an invited researcher of Waseda University, a president of SCB Lab, and a principal of SCB Innovation Academy.

Quick Blocking Operation of IDS/SDN Cooperative Firewall Systems by Reducing Communication Overhead

Akihiro Takai*, Yusei Katsura**, Nariyoshi Yamai*, Rei Nakagawa*, and Vasaka Visoottiviseth***

** Graduate School of Engineering, Tokyo University of Agriculture and Technology, Tokyo, Japan*

*** Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara, Japan*

**** Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom, Thailand*

s224164x@st.go.tuat.ac.jp, katsura.yusei.ky6@is.naist.jp, nyamai@cc.tuat.ac.jp, rnakagawa@go.tuat.ac.jp, vasaka.vis@mahidol.edu

***Abstract*—An Intrusion Detection System (IDS) / Software Defined Networking (SDN) cooperative firewall system has attracted much attention recently because it has many advantages of dynamic network configuration with SDN and scalable IDS hosts. In the IDS/SDN cooperative firewall system, an SDN switch relays traffic between a client and a server and mirrors traffic from a client to an IDS host. The IDS host monitors the mirrored traffic and notifies the SDN switch to block malicious traffic according to the detection of the attack. At this point, malicious packets reach the server until the IDS detects the attack and notifies it. In this paper, we propose a method to speed up mirroring and notification by integrating IDS and SDN switch hosts as a method to shorten the blocking time and compare it with existing methods. The experimental system was constructed using Raspberry Pi3 B+ and 4B boards. As a result, it was confirmed that the proposed method completes the blocking operation faster than the existing method. We also investigated the breakdown of the blocking time to confirm the effect of the proposed method.**

***Keyword*— Firewall, Intrusion Detection System, OpenFlow, Software Defined Network**

I. INTRODUCTION

TO prevent computing devices of the organization from being compromised, a firewall, intrusion detection system (IDS), and intrusion prevention system (IPS) that can detect and block malicious files are essential. For quick and flexible network management and maintenance in large organization networks, the Software-Defined Network (SDN) should be used [1]. Upon integrating SDN with IDS, network administrators can deploy a flexible firewall system, namely the IDS/SDN cooperative firewall system [2], [3]. In this integration, the SDN switch that relays the bidirectional communication traffic of the target network mirrors the traffic to the IDS, and the IDS will detect anomaly packets that may be attacks from an outsider and inform the SDN controller. Once the SDN controller learns about attacks, it will create rules or policies to push to SDN switches to block malicious traffic coming into the network to be protected. In addition, multiple IDS servers can be used to provide load balance between IDS servers. Furthermore, by using the integration of SDN and IDS, not only a flexible firewall, but network administrators can also configure flexible network routes, for example, forwarding anomaly traffic to honeypots for further security analysis.

The large delay required for the period to mirror packets from the SDN switch to the IDS and notify attack detection from the IDS to the SDN controller (the blocking time) allows malicious packets to enter the protected network in the meantime. In the existing IDS / SDN cooperative firewall system, the OpenFlow, which is the leading implementation of the SDN concept, conventionally uses the REST API for notification from IDS [2]. However, because most IDS implementations do not support the REST API, the system must use the log monitoring tools to observe the change of IDS alert logs and configure the SDN controller to block this anomaly traffic via REST API, resulting in increasing the blocking time.

To reduce the blocking time, in the previous work, we have proposed a method to change the notification method from REST API to Syslog, which is faster because of reducing communication overhead [4]. To further reduce the blocking time along with using the fast Syslog notification method, this paper demonstrates a method of integrating IDS hosts and SDN switch hosts, which are detached physical hosts in existing systems, into a single physical host, so that mirroring and notification can be completed within a single physical host. Therefore, combining the integration method and the fast Syslog notification mitigates the communication overhead compared with the existing system of the separate physical hosts and reduces the blocking delay of malicious traffic. To evaluate the effectiveness of the system, we conducted an experiment to compare the blocking time of the existing system in which the IDS and SDN switch hosts are

Manuscript received December 8, 2022. This work is a follow-up of the invited journal to the accepted & presented paper entitled "Quick blocking operation of firewall system cooperating with IDS and SDN" of the 24th International Conference on Advanced Communication Technology (ICACT2022).

Akihiro Takai, he is a master's student in the Graduate School of Engineering at Tokyo University of Agriculture and Technology, Japan. (e-mail: s224164x@st.go.tuat.ac.jp)

Yusei Katsura, he is a Ph.D. student in the Graduate School of Science and Technology at Nara Institute of Science and Technology, Japan. (e-mail: katsura.yusei.ky6@is.naist.jp)

Nariyoshi Yamai, he is a professor in the Graduate School of Engineering, Tokyo University of Agriculture and Technology, Japan. (corresponding author, phone: +81-42-388-7695, e-mail: nyamai@cc.tuat.ac.jp)

Rei Nakagawa, he is an assistant professor in the Graduate School of Engineering at Tokyo University of Agriculture and Technology, Japan. (e-mail: rnakagawa@go.tuat.ac.jp)

Vasaka Visoottiviseth, she is an associate professor in the Faculty of Information and Communication Technology at Mahidol University, Thailand. (e-mail: vasaka.vis@mahidol.edu)

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

detached and the proposed system in which the IDS and SDN switch hosts are integrated. In the experiment, two kinds of experimental systems, the existing method and the proposed method, were constructed with real devices, and a total of four blocking times were measured by the combination of two kinds of notification methods, REST API and Syslog. We make following contributions.

- Proposal of quick blocking operation of integrating IDS hosts and SDN switch hosts using the fast notification method.
- Implementation of the proposed method using Raspberry Pi 4B and 3B+ as devices, Suricata as IDS, and Open vSwitch (OvS) as SDN switch implementation.
- Experimental verification of the reduction of blocking time by host integration in combination with two types of notification methods, REST API and Syslog.

II. Background and Existing Works

A. *Software Defined Network*

In the traditional network, once any network policy changes, network administrators have to carefully configure all switches manually. It requires a great deal of network maintenance. To solve this problem, the SDN is introduced. In the SDN concept, network tasks are divided into the control plane and the data plane. The control plane, which is responsible for making decisions on how packets should be handled, is the task of the SDN controller. On the other hand, the data plane or forwarding plane, which is responsible for handling packets based on the instructions from the control plane, is the task of the SDN switch. Instructions from the SDN controller can be either forwarding or dropping packets on the SDN switch. It can manage or control the forwarding table residing on SDN switches. On the other hand, SDN switches can focus on accelerating the packet forwarding.

By using SDN, administrators can centrally maintain the network policy via the SDN controller. Instead of manually updating each switch, SDN enables the network administrators to distribute the policy evenly across multiple switches. This can simplify network management.

In SDN, there are Northbound and Southbound APIs that are APIs operating between data plane, control plane and application plane. The Northbound APIs are available on an SDN controller and allow applications or the application plane to interact with the controller, which is the control plane. Applications and services are, for example, load-balancers and firewalls. On the other hand, the Southbound APIs are available between the SDN controller in the control plane and other forwarding devices, e.g. switches, and routers, which are the data plane or forwarding plane. By using the Southbound APIs, administrators can adjust the network according to the change requirements.

If you want to submit your file with one column electronically, please do the following:

1) *OpenFlow*

OpenFlow is a well-known technology maintained by the Open Networking Forum (ONF) [5] that implements the SDN concept. The OpenFlow protocol is a set of specifications that provides the southbound interface and defines how the controller interacts with the data plane. The newest version of the OpenFlow Switch specification is version 1.5.1. Each OpenFlow switch will store information about the flow entry received from the controller into the flow table. Each flow entry contains an explicit action to handle each flow.

On an OpenFlow switch, there are two main components: the switch-agent and the data plane. The switch-agent uses the OpenFlow protocol to communicate with one or more controllers. Moreover, it communicates with the data plane using the internal protocol.

Every OpenFlow message begins with the same header structure containing version, type, length, and transaction id (xid). The OpenFlow message types can be, for example, FlowMod, PacketIn, FlowRemoved, PacketOut, and StatsReq. The FlowMod message is used to allow the controller to modify the flow entries on the OpenFlow switch. On the other hand, the PacketIn message type is used by the switch to send incoming packets to the OpenFlow controller. Normally, there are two cases where PacketIn message type is used: (1) there is an explicit action for this behavior specified in the flow entry and (2) the flow does not match any flow entry in the table. Moreover, to make the OpenFlow controller understand that the traffic comes from which switch, each switch will be assigned a Datapath ID.

There are many OpenFlow controller software, for example, Ryu, Treman, Opendaylight, etc. In this research, we selected Ryu as the SDN controller software.

2) *Ryu*

Ryu [6] is a framework that implements the OpenFlow controller and is developed in Python. Once the Ryu controller receives an OpenFlow message from a switch, it will trigger an event. An event handler contains an event class and the state of the switch. States of OpenFlow switch can be, for example, the HANDSHAKE_DISPATCHER, which is the initial state that exchanges the HELO message, and the MAIN_DISPATCHER, which is the normal state. When the controller receives a PacketIn message from the switch, the EventOFPPacketIn event class will be called. For example, with the API set_ev_class(ofp_event.EventOFPPacketIn, MAIN_DISPATCHER), we can define the process when the OpenFlow switch is in the normal operation state and receives a PacketIn message [7].

B. *Related Works*

Here, we surveyed related works that utilize Snort IDS [8] and SDN together to dynamically filter anomaly traffic.

Nam *et al.* studied SDN security enhancement using open-source IDS/IPS Suricata [9][10]. They proposed requirements for implementing SDN security. They mentioned whether SDN solutions, Suricata IDS/IPS, automated intrusion prevention, and mirroring are necessary or not when they want to implement a firewall, network scan detection, abnormal traffic detection, intrusion detection, and intrusion prevention. They concluded that to implement intrusion prevention they need Suricata IDS/IPS, automated intrusion prevention, as well as mirroring because the SDN alone cannot handle this. However, no implementation and performance were shown in their paper.

Hendrawan *et al.* studied the performance degradation when integrating the SDN architecture with an IDS, either the signature-based Snort or the anomaly-based Bro [11]. The throughput, delay, and packet loss, which are required performance metrics of quality of service (QoS), are observed. The results of their experiment on the mininet virtual network [12] showed that integrating SDN with Bro IDS gave better performance than using Snort IDS in all

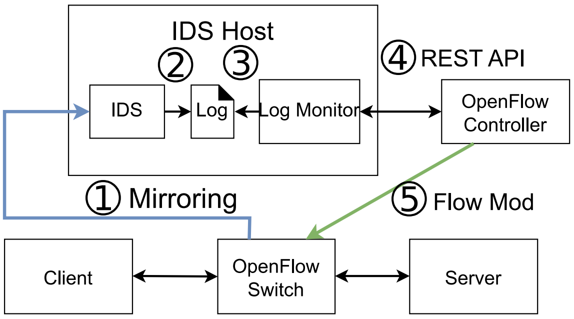


Fig. 1. System architecture of the Original IDS/SDN Cooperative Firewall System.

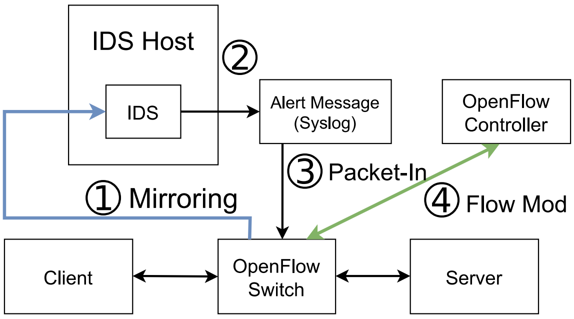


Fig. 2. System architecture of the Quick block operation by Syslog notification.

aspects: throughput, delay, and packet loss. However, the delay is not much different, while the CPU usage and memory usage of Snort are lower than those of Bro.

R. Sutton et al. designed and developed a system that utilized an SDN to divert traffic from some suspicious traffic multiple Snort IDS machines for inspection [13]. The authors developed 'PySnorter,' a Python tool for routing traffic to dedicated Snort machines. Once Snort alerts the malicious traffic, it pulls the information from the alert and generates a REST API command that is then sent to the SDN controller. For the experiments, they implemented a virtualized network using GNS3, which is the graphical version of Network Simulator-3, and used Open vSwitch. However, latency was not measured.

C. *The Original IDS/SDN Cooperative Firewall System*

In firewall systems cooperating with IDS and SDN, the IDS generally do not have the capability to send alert messages directly to the SDN controller, but instead has the capability to send them to an alert handler as Syslog messages or record them into a log file. Typically, Syslog messages are in the common event format (CEF) or log event extended format (LEEF) to facilitate analysis. However, similar to what was proposed in the [12], in the case of a firewall system recording alert messages in a log file and using OpenFlow as an SDN platform, the typical behavior of the firewall system to block malicious packets consists of the following six steps, as illustrated in Fig. 1.

(1) When the OpenFlow switch receives packets from the client that are destined to the server, it will delay the forwarding process to the destination and copy the traffic to IDS via port mirroring.

(2) Once the IDS receives those packets, it will analyze and detect malicious packets. If the IDS finds malicious packets, it will alert and write the information to the log file.

(3) The log monitoring tool continuously monitors that log file.

(4) When the log monitoring tool detects a change in the log file, that is, a new alert message written in the log file, it will use the REST API to send the information of that malicious packet to the OpenFlow controller.

(5) The OpenFlow controller uses the received information to create a flow entry that determines how an OpenFlow switch must behave when it receives packets from that malicious flow. Then, the OpenFlow controller will use the FlowMod message to send that Flow entry to the OpenFlow switch.

(6) The OpenFlow switch updates the flow table based on the information specified in the FlowMod message. Then, it can process packets that are delayed in step 1.

In this case, because most IDS do not support REST APIs, they have to use log monitoring tools to monitor log change and use a command line tool such as 'curl' to transfer REST APIs. Therefore, a processing overhead will occur. Moreover, REST APIs use HTTP methods and run over TCP protocol.

D. *Quick Block Operation by Syslog Notification*

To reduce communication overhead due to REST API, we have proposed the fast notification method for the IDS/SDN cooperative firewall system as illustrated in Fig. 2. There are five steps to perform.

(1) When the OpenFlow switch receives packets from the client that are destined to the server, it will delay the forwarding process to the destination and copy the traffic to IDS via port mirroring.

(2) If the packet matches with any signatures in the IDS, the IDS will send the alert message to the OpenFlow switch by using, for example, the Syslog message or the SNMP trap.

(3) When the OpenFlow switch receives the alert message sent via UDP, it will forward the message as the PacketIn message type to the OpenFlow controller.

(4) The OpenFlow controller uses the received information to create a flow entry that determines how an OpenFlow switch must behave when it receives packets from those malicious flows. Then, the OpenFlow controller will use the FlowMod message to send that Flow entry to the OpenFlow switch.

(5) The OpenFlow switch updates the flow table based on the information specified in the FlowMod message. Then, it can process packets that are delayed in step 1.

Note that this method can reduce one step compared with the existing method mentioned in Section II B, because we send the alerts as the Syslog message directly to OpenFlow switch, but the existing method needs to write the alerts in the log file first and use the log monitoring tools to detect the log change. The characteristics of REST API and Syslog as notification methods can be summarized as follows.

- REST API

The REST API submissions use an external program Swatchdog to monitor logs and curl to send packets. The POST method of the REST API is used to send notifications as well as host blocking. Since TCP is used for communication, it is expected that there will be overhead to establish the connection.

- Syslog

Most IDSs can notify anomalies using Syslog. Syslog can communicate via either UDP or TCP. In an IDS/SDN cooperative firewall system, it is considered better to use UDP for low latency communication.

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

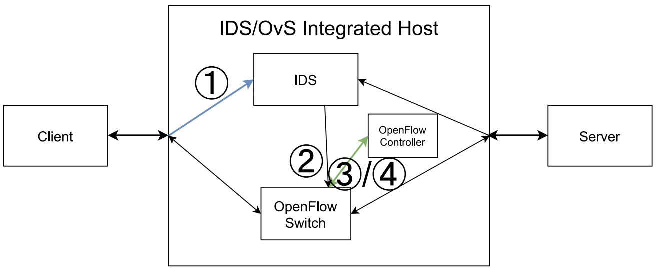


Fig. 3. System architecture of the proposed work.

TABLE I
SOFTWARE SPECIFICATION

Device	Software
OpenFlow controller	Ryu 2.7
OpenFlow switch	Open vSwitch 2.10.1 (OvS)
Intrusion Detection System	Suricata 6.0.6
Log monitoring tools	swatch

III. Proposed Work

To further reduce the blocking time, we adopt an approach to reduce the communication overhead between separated physical hosts of IDS and SDN, as well as that of existing notification method, REST API. Therefore, we propose a method to reduce the blocking time by running an IDS and an SDN switch implementation, such as Open vSwitch, on a single device, thereby reducing the communication overhead in the mirroring in step 1 and the notification in step 4, as illustrated in Fig. 3. In the IDS/SDN cooperative firewall system experimented in the previous study [4], the host acting as an SDN switch and the IDS host are different devices. Therefore, we propose a method to reduce the blocking time by running an IDS and an SDN switch implementation, such as Open vSwitch, on a single device, thereby reducing the communication overhead in the mirroring in step 1 and the notification in step 4. An IDS/SDN cooperative firewall system that integrates an IDS host and an SDN switch host in a single device is called an integrated system, while one that has different hosts for each is called a detached system.

In our implementation, we implement two approaches: (1) the existing detached system and (2) our proposed integrated system. Details of them are described below.

IV. Implementation

For the implementation, as mentioned earlier in Section II, we select Ryu as the OpenFlow controller. For the specification of OpenFlow, we use OpenFlow version 1.3. For the IDS, we select Suricata, which is a famous Open Source Software (OSS). Suricata has a large community and has more than thousands of detection signatures available. Table I summarizes the software and its version used in our experiments.

Moreover, we used two Raspberry Pi 3B+ boards and a Raspberry Pi 4B board to emulate each device. First, Raspberry Pi 3B+ is used to implement a client host and a server host, while Raspberry Pi 4B is used for implementing IDS/OvS integrated host. In the detached system, we use Raspberry Pi 3B+ for the client host, server host and IDS host, while Raspberry Pi 4B is used for implementing OvS host.

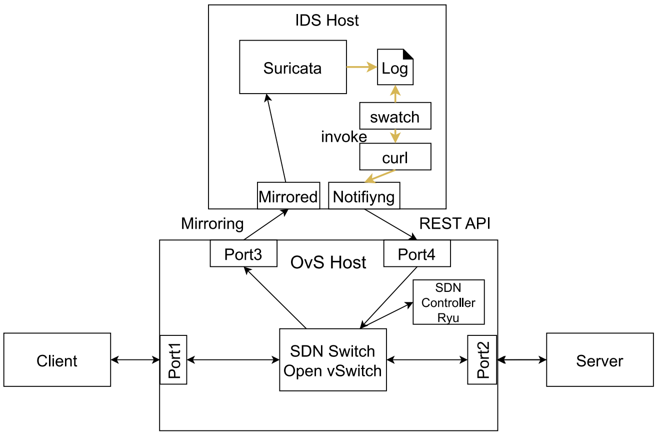


Fig. 4. Implementation of the detached system.

.

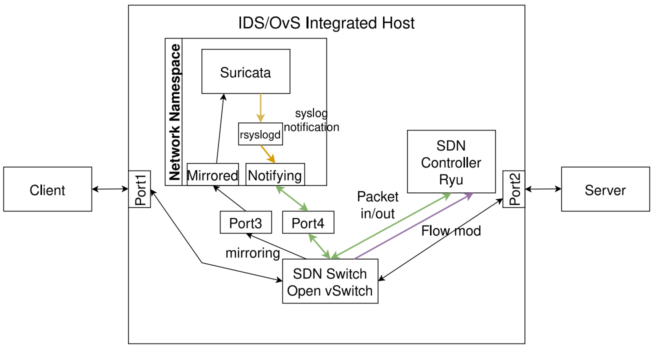


Fig. 5. Implementation of the integrated system.

.

A. Implementation of the Detached System

First, we implement the existing approach that uses Suricata to write alerts to the log file and uses swatch to monitor the log file. As shown in Fig. 4, in this approach Suricata and the swatch are running on the same Raspberry Pi board. Moreover, the Raspberry Pi that emulates both Ryu controller and Open vSwitch is equipped with four ports: Port 1 connecting to the client, Port 2 connecting to the server, Port 3 which is a one-way directional port just to forward packets to Suricata to analyze, and Port 4 which is also a one-way directional port for Ryu to receive REST APIs over TCP from the swatch utility. Moreover, there are only FlowMod messages sent from Ryu to OvS when the Ryu controller wants to modify the flow table on OvS.

B. Implementation of the Integrated System

Fig. 5 shows a diagram of the IDS and SDN switch. A Linux network namespace is running on the IDS/OvS integrated host. Using the Network Namespace function, IP-related processes can be divided into multiple processes within a single Linux unit. Suricata runs within a Network Namespace on the integrated host. This Network Namespace has two Ethernets: one corresponds to port 3 and is used to receive packets from the client and server ports (Ports 1 and 2), and the other corresponds to port 4 and is used by Suricata to notify the outside of Network Namespace.

First, packets arriving from the client are forwarded to the server and simultaneously mirrored to the Network Namespace which the IDS operates. Packets arriving from

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

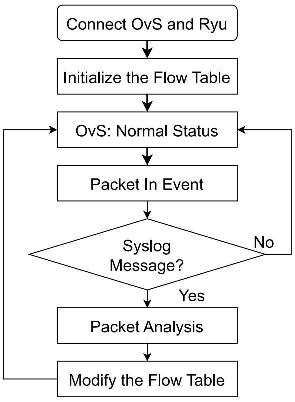


Fig. 6. Flowchart of the process inside the Ryu controller.

TABLE II
FLOW TABLE IN THE INITIAL STATE

Match	Action
in_port 1	OutPut (Port 2, Port 3)
in_port 2	OutPut (Port 1, Port 3)
in_port 4	OutPut (Controller)

TABLE III
FLOW TABLE AFTER GOT ATTACKS

Match	Action
in_port 1	OutPut (Port 2, Port 3)
in_port 2	OutPut (Port 1, Port 3)
in_port 4	OutPut (Controller)
ip_src 192.168.10.21 tcp_dst 80 eth_type = 0x0800	Drop

the server are also forwarded to the client and mirrored to the network namespace. The IDS monitors the packets and notifies the SDN controller across the network namespace using Syslog if any malicious packet is detected. The SDN controller receives the notification and modifies the Open vSwitch flow table to drop the packets from the host that sent the packets deemed to be malicious. that sent the packets deemed to be malicious. This allows the mirroring and notification procedures of the blocking operation to be completed in a single device.

Fig. 6 shows the flowchart of the processes inside our Ryu controller. Once the OvS connects with the Ryu controller, it will initialize the flow table as shown in Table II and set the state of the OvS as normal. When the Ryu controller receives the PacketIn message, it will check whether the message is a Syslog message or not. If so, it will analyze the message and modify the flow table. An example of a flow table that is modified after receiving DDoS attacks is shown in Table III. A new flow entry is added to the flow table specifying the “Drop” action when the packet contains the header that matches the specified condition.

V. EVALUATION RESULTS

For evaluation, we compare our proposed work with the existing approach and the proposed approach described in Section IV A and Section IV B, respectively. To measure the blocking time, we observe the packet timestamp using the tcpdump command. To investigate the breakdown of blocking time, we define the communication overhead (i.e., the blocking time) using the following periods of processing in the IDS/SDN cooperative firewall system.

TABLE IV
BLOCKING TIME OF EACH APPROACH

	REST API	Syslog
Detached System	75.6 ms	9.6 ms
Integrated System	67.3 ms	4.6 ms

1. Mirroring — the time taken to mirror packets from OvS to IDS.
2. Log monitoring — the time it takes for Swatch to detect IDS writes to the log.
3. Start script — time required for Swatch to start the REST API submission script.
4. Notification — the time taken to notify in REST API.
5. Syslog — the time taken for IDS to notify in Syslog.
6. Packet In/Flow Mod — the time it takes for OvS to packet in a Syslog packet and invoke Flow Mod processing.

Both existing and proposed approaches measure the blocking time from the time OvS host receives a malicious packet until SDN controller Ryu receives a Syslog or REST API Packet-in packet and calls the Flow Mod process. Table IV summarizes the blocking time for each method. First, we note the performance differences due to differences in the notification methods. When comparing Figs. 7 and 8, and Figs. 9 and 10, respectively, using Syslog as the notification method, the overhead of log monitoring, REST API sending script invocation, and notification processing required when using REST API was reduced, and the blocking operation could be performed at a speed of approximately 60 ms. In particular, we observed that it was taking a long time to send REST APIs using curl.

Next, we discuss the differences between the results of the detached and integrated systems. Comparing Figs. 7 and 9, and Figs. 8 and 10, respectively, the integrated system completed the blocking operation faster, and the breakdown suggests that this is due to faster mirroring, notification, Packet-in, and Packet-out. Furthermore, comparing the detached and integrated systems in the results using Syslog as the notification method, the time spent for communication was approximately 0.6 ms for the detached system and approximately 0.3 ms for the integrated system. When using the REST API, the time spent for mirroring was approximately 2.2 ms for the detached system and 0.27 ms for the integrated system. These results indicate that the proposed method reduces the communication overhead. Although there are differences in the boards on which the IDS is running, the fact that mirroring and notification are completed within a single host is one of the factors contributing to the reduction in blocking time. To more accurately observe the effect of the proposed method, it is necessary to conduct experiments on a unified board on which the IDS operates.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a method to reduce blocking time by integrating IDS hosts and SDN switch hosts along with the fast notification method and we have confirmed the effectiveness of the proposed method through experiments.

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

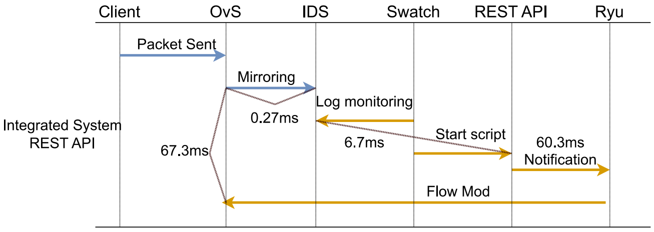


Fig. 7. Blocking time of Integrated System when using REST API.

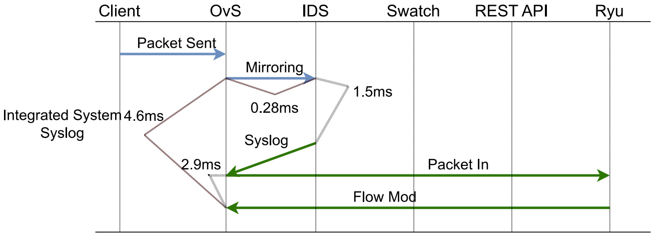


Fig. 8. Blocking time of Integrated System when using Syslog.

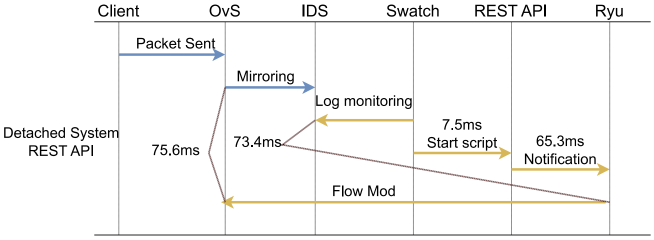


Fig. 9. Blocking time of Detached System when using REST API.

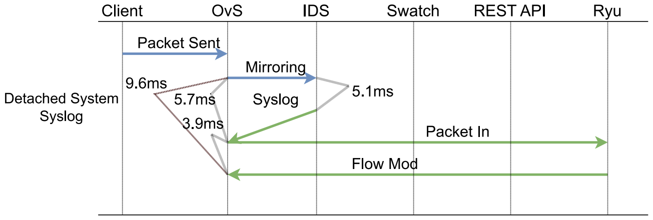


Fig. 10. Blocking time of Detached System when using Syslog.

In the experiments, the blocking time was measured by combining two types of notification methods and two types of IDS/SDN cooperative firewall systems: the integrated system proposed here and the detached system. The experimental results confirm that the notification method using Syslog is faster than the REST API because the Syslog notification method removes the process of monitoring IDS log for REST API. In addition, integrating IDS and SDN hosts reduced the communication overhead between the IDS and SDN, which were physically separated in conventional, and as a result, further reduced the blocking time at no load on the system. We also confirmed that the reduction of communication overhead, which is the goal of the proposed method, contributes to faster blocking operation by examining the breakdown of the blocking time. However, the proposed method and the existing method differ in the host on which the IDS operates, and we have not confirmed how the blocking times of the two systems change when a traffic load is added to them. Since the integrated system is more susceptible to the huge network traffic load than the detached system since IDS and OvS are running on a single host, the performance of the integrated system is expected to be degraded under high load. Therefore, for future work, an immediate verification method is required.

Moreover, the IDS/SDN cooperative firewall system using the IDS/OvS integrated host has an advantage over existing IDS/SDN cooperative firewall systems in terms of parallelization of OvS hosts. This is because the IDS and OvS hosts are integrated, which simplifies the network configuration. Parallelization of IDS and OvS hosts is effective to suppress the increase in shutdown time of IDS/SDN linked firewall system under load environment. In the future, we will construct an OvS host parallelized IDS/SDN cooperative firewall system using an IDS/OvS integrated host and compare its performance with existing parallelized configurations and examine its qualitative cost.

REFERENCES

[1] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, “OpenFlow: enabling innovation in campus networks”, *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp 69–74, April 2008.

[2] P. Zanna, B. O'Neill, P. Radcliffe, S. Hosseini, and M. S. Ul Hoque, “Adaptive threat management through the integration of IDS into Software Defined Networks”, in *Proc. 2014 Int. Conf. and Workshop on the Network of the Future (NOF)*, pp. 1-5, 2014.

[3] Y. Katsura, H. Kimiyama, T. Tsutsumi, N. Yonezaki, J. Ichikawa, and M. Maruyama, “Proposal of real-time brute-force attack detection and blocking system using software switch”, *IEICE Technical Report*, vol. 118, no. 465, NS2018-272, pp. 461-464, 2019 (in Japanese).

[4] Y.Katsura, P.Sakarin, N.Yamai, H. Kimiyama, and V. Visoottiviseth: “Quick blocking operation of firewall system cooperating with IDS and SDN,” in *Proc. 24th Int. Conf. Advanced Communications Technology (ICACT 2022)*, pp.393-398, February, 2022.

[5] The Open Networking Foundation, “SDN Technical Specifications” [Online]. Available: https://opennetworking.org/software-defined-standards/specifications/

[6] Ryu SDN Framework Community, “Ryu SDN Framework” [Online]. Available: https://ryu-sdn.org/

[7] Nippon Telegraph and Telephone Corporation Revision d6cda4f4, “Ryu application API” [Online]. Available: https://ryu.readthedocs.io/en/latest/ryu_app_api.html

[8] Cisco and/or its affiliates, “Snort – Network Intrusion Detection & Protection System” [Online]. https://snort.org/

[9] The Open Information Security Foundation (OISF), “Home – Suricata” [Online]. Available: https://suricata.io/

[10] K. Nam and K. Kim, “A Study on SDN security enhancement using open source IDS/IPS Suricata,” in *Proc. 2018 Int. Conf. Information and Communication Technology Convergence (ICTC)*, pp. 1124-1126, Jeju, South Korea, 2018.

[11] H. Hendrawan, P. Sukarno, and M. A. Nugroho, “Quality of Service (QoS) Comparison Analysis of Snort IDS and Bro IDS Application in Software Define Network (SDN) Architecture,” in Proc. 2019 7th Int. Conf. Information and Communication Technology (ICoICT), pp. 1-7, Kuala Lumpur, Malaysia, 2019.

[12] Open Networking Foundation, “MININET - Open Networking Foundation” [Online]. Available: https://opennetworking.org/mininet/

[13] R. Sutton, R. Ludwiniak, N. Pitropakis, C. Chrysoulas and T. Dagiuklas, “Towards an SDN Assisted IDS,” in *Proc. 2021 11th IFIP Int. Conf. New Technologies, Mobility and Security (NTMS)*, pp. 1-5, Paris, France, 2021.

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

Akihiro Takai was born in Japan in 1999 and received his B.E. from Tokyo University of Agriculture and Technology (TUAT), Japan in 2022. Currently, he is a master’s student at Tokyo University of Agriculture and Technology, Japan. His research interests include computer networks.

Yusei Katsura was born in Japan in 1996 and received his B. Info. Env. degree from Tokyo Denki University, Japan in 2019, and his M. S. degree in computer engineering from Nara Institute of Science and Technology (NAIST), Japan in 2022. He was a research student at Tokyo University of Agriculture and Technology (TUAT), Japan in 2019-2020. He is currently a Ph.D. student at Nara Institute of Science and Technology. His research interests include computer networks.

Nariyoshi Yamai was born in Japan in 1961 and received his B.E. and M.E. degrees in electronic engineering and Ph.D. degree in information and computer science from Osaka University, Japan in 1984, 1986, and 1993, respectively. Currently, he is a professor at the Institute of Engineering, Tokyo University of Agriculture and Technology (TUAT), Japan. His research interests include distributed systems, network architecture, network security, and the Internet.

Rei Nakagawa was born in Japan in 1993, received his B.S. and M.S. degrees from the Tokyo University of Science, Japan, in 2016, and 2018 respectively, and a Ph.D. degree in informatics and engineering from the University of Electro-Communications (UEC), Japan, in 2021. He has been an assistant professor at the Institute of Engineering, Tokyo University of Agriculture and Technology (TUAT), Japan since April 2021. His research interests include network architecture, video streaming technology, software defined network, and information centric networking.

Vasaka Visoottiviseth was born in Thailand in 1975, received her M.E. and B.E. degrees from Tokyo University of Agriculture and Technology (TUAT), Japan in 1999 and 1997, respectively, and received her Ph.D. degree in computer engineering from Nara Institute of Science and Technology (NAIST), Japan in 2003. Currently, she is an associate professor at Mahidol University, Thailand. Her current research interests include mobile and wireless computing, network security, and digital forensics.

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

Automated Vulnerability Assessment Approach for Web API that Considers Requests and Responses

Yuki Ishida*, Masaki Hanada**, Atsushi Waseda**, and Moo Wan Kim***

* Graduate School of Informatics, Tokyo University of Information Sciences, Japan

** Department of Informatics, Tokyo University of Information Sciences, Japan

*** TA Tech., Japan

h22001iy@edu.tuis.ac.jp, mhanada@rsch.tuis.ac.jp, aw207189@rsch.tuis.ac.jp, ykim5jp@ybb.ne.jp

Abstract— **In recent years, Web Application Programming Interfaces (Web APIs) have been extensively used in numerous web applications. However, the number of attacks exploiting Web API vulnerabilities has been rapidly increasing. The Open Web Application Security Project (OWASP) published guidelines known as the OWASP API Security Top 10 to mitigate the risks associated with these vulnerabilities. The guidelines identify the top 10 most critical security risks in Web APIs and provide remediation guidance to help developers. Although developers are required to address these vulnerabilities according to these guidelines, traditional vulnerability assessment tools may not perform adequately when used to assess Web API vulnerabilities. Manually addressing these is difficult because there are a large number of endpoints and parameters in Web APIs using traditional vulnerability assessment tools. To address this issue, we propose a method for automatically conducting Web API vulnerability assessments by utilizing references, requests, and responses for Web APIs. In the evaluation experiment, we showed that the proposed method can detect authorization-related vulnerabilities in the Web APIs of vulnerable testing environments and well-known Content Management Systems, such as Wordpress, Ghost CMS, and Joomla.**

Keywords— **Web API, Vulnerability Assessment, Automation Analysis, Security**

I. Introduction

In recent years, the widespread adoption and use of Web Application Programming Interfaces (Web APIs) have greatly enhanced the convenience of web systems. However, the vulnerabilities associated with Web APIs are increasingly subject to attacks. Akamai Inc. reported that more than 11 billion attacks occurred between January 2020 and June 2021 [1]. Specifically, in June 2021, 113.8 million attack traffic events were observed in a single day, and attacks targeting Web APIs are on the rise. The proactive elimination of Web API vulnerabilities is very important to guarantee a safe and secure cyberspace for general users.

To mitigate the risks associated with these vulnerabilities, the Open Web Application Security Project (OWASP) published guidelines known as the OWASP API Security Top 10 [2]. The guidelines identify the top 10 most critical security risks in Web APIs and provide remediation guidance to help developers. Although developers are required to address these vulnerabilities according to such guidelines, it is difficult to address them manually because of the large number of endpoints and parameters in Web APIs.

To solve these issues, traditional web applications have employed a variety of tools, such as OWASP ZAP (Zed Attack Proxy), for vulnerability assessment. However, because these tools are primarily designed for traditional web applications and content, they may not perform adequately when used to assess Web API vulnerabilities. Specifically, developers are required to manually configure many parameters of traditional tools (e.g., OWASP ZAP) for vulnerability assessment because traditional tools do not consider the logic and characteristics of Web APIs. A lack of understanding of the logic and characteristics of Web APIs causes omissions or mismatches in endpoints or parameters. Consequently, this issue leaves risks of Web API vulnerabilities.

In this study, we propose a method for automatically conducting a Web API vulnerability assessment by utilizing references, requests, and responses for Web APIs to detect authorization-related Web API vulnerabilities. The proposed method first generates endpoints and parameters using Web API references, and requests for validating the Web API are subsequently sent to the generated endpoints. Next, new endpoints and parameters are generated using the responses from the endpoints, and vulnerability assessment requests are sent to the generated endpoints. Finally, when the HTTP status codes of the responses from the endpoints satisfy certain conditions, the proposed method determines that they are valid endpoints and/or parameters and do not contain vulnerabilities.

Manuscript received on Sep. 27, 2023. This work is a follow-up on the accepted and presented paper titled ”An Automated Vulnerability Assessment Approach for WebAPI that Considers Requests and Responses,” at the 24th International Conference on Advanced Communication Technology (ICACT2022).

Yuki Ishida, He is a Ph.D. student in the Graduate School of Informatics, Tokyo University of Information Sciences, 4-1, Onaridai, Wakaba, Chiba, Japan (e-mail: h22001iy@edu.tuis.ac.jp.)

Masaki Hanada is with Department of Informatics, Tokyo University of Information Sciences, 4-1, Onaridai, Wakaba, Chiba, Japan (corresponding author, phone: +81-43-236-1224, e-mail: mhanada@rsch.tuis.ac.jp).

Atsushi Wasedais is with Department of Informatics, Tokyo University of Information Sciences, 4-1 Onaridai, Wakaba-ku, Chiba, Japan (e-mail: aw207189@rsch.tuis.ac.jp)

Moo Wan Kim is with TA-Tech, Japan (e-mail: ykim5jp@ybb.ne.jp)

ISSN: 2288-0003 (Online)

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

TABLE I
Example of RPC API and REST API

Operation	**Type of API**	**HTTP Method**	**Endpoint**	**Request Parameters** (JSON Format)
Read	RPC	POST	/getItem	{”id”:”1”}
	REST	GET	/items/1	None
Create	RPC	POST	/createItem	{”name”: ”Pen”, ”price”: 100}
	REST	POST	/Items	{”name”: ”Pen”, ”price”: 100}

In the evaluation experiment, we showed that the proposed method can detect authorization-related vulnerabilities in the Web APIs of vulnerable testing environments and Content Management Systems (CMSs).

The remainder of this paper is organized as follows. Section II presents an overview of Web APIs. Section III presents the OWASP API Security Top 10. Traditional vulnerability assessment tools are presented in Section IV. Section V describes the details of the proposed method. Section VI presents the experimental environment. Section VII presents the experimental results. Finally, Section VIII concludes the paper.

II. Web APIs

A Web APIs is an API that is accessed using the HTTP protocol and typically uses either Extensible Markup Language (XML) or JavaScript Object Notation (JSON) formats to encode data. Web APIs are extensively utilized to enhance the communication between web systems and a variety of web applications. Currently, despite the lack of a universal standardized format for Web APIs, Remote Procedure Call API (RPC API) and Representational State Transfer API (REST API) have been widely used as representative architectural styles in Web API design.

A. RPC API

The RPC API is based on the remote procedure call (RPC). The representative implementations of the RPC API are XML-RPC and JSON-RPC; whereas JSON-RPC uses a lightweight JSON format to encode data, XML-RPC uses a heavier XML format.

B. REST API

The REST API is a simple method for accessing Web resources. The REST API endpoint is a URL that utilizes HTTP methods such as POST, GET, POST, PUT, and DELETE to execute the CRUD (Create, Read, Update, and Delete) operations. It primarily focuses on providing resources from the server to clients. Similar to the RPC API, the REST API can use either XML or JSON to encode data.

Table I presents an example of HTTP requests to perform read and create operations in RPC API and REST API.

In this study, we targeted the REST API because it has become a mainstream Web API, and the Web APIs of many CMSs are provided by the REST API.

TABLE II
List of OWASP API Security Top 10

Risk ID	Name of Vulnerability items	Category
API 1:2019	Broken Object Level Authorization	Category 1
API 2:2019	Broken User Authentication	Category 1
API 3:2019	Excessive Data Exposure	Category 2
API 4:2019	Lack of Resources & Rate Limiting	Category 3
API 5:2019	Broken Function Level Authorization	Category 1
API 6:2019	Mass Assignment	Category 2
API 7:2019	Security Misconfiguration	Category 3
API 8:2019	Injection	Category 4
API 9:2019	Improper Assets Management	Category 3
API 10:2019	Insufficient Logging & Monitoring	Category 4

III. OWASP API Security Top 10

OWASP is a nonprofit foundation that works to improve the security of web applications. OWASP published guidelines known as OWASP API Security Top 10 to mitigate the risks associated with Web API vulnerabilities; these guidelines aimed to provide guidance on the most important security risks to consider when developing and exposing APIs. The guidelines include recommended countermeasures and outline scenarios in which Web API vulnerabilities may occur.

Table II lists the OWASP API Security Top 10, which can be classified into four primary categories. The *category 1* in Table II pertains to access control vulnerabilities, such as the unauthorized use of Web APIs by general users, owing to inadequate authorization settings intended for administrators. The *category 2* includes vulnerabilities related to data handling. For example, a Web API server may transmit data to the client while assuming that filtering will occur on the Web application side, or it may include undefined data within the request body. The *category 3* highlights vulnerabilities that may occur during the development and operational phases; this includes exposure to problematic APIs utilized during development and issues related to the security configurations of the Web API server. The *category 4* encompasses vulnerabilities typical of traditional web threats, such as injection attacks and incorrect logging configurations.

This study focuses on authorization-related vulnerabilities as outlined in *API 1:2019* and *API 5:2019*, both of which hold significant importance on this list. Additionally, we investigated the endpoints associated with *API 9:2019*. In this vulnerability, endpoints not defined in the API references are exposed because of a lack of proper authentication and authorization.

A. API 1:2019 Broken Object Level Authorization

API 1:2019 is an authorization-related vulnerability. Traditional web applications utilize a mechanism called *session* that stores users' authentication status for both web servers and clients. However, a Web API does not typically manage users' authentication status (i.e., stateless), and user management is often achieved by embedding an identifying flag at the endpoint. For example, the endpoint for accessing the information of a user with ID 1 is */user/1*. In such cases, if a Web API server fails to properly authorize access to user information, a malicious user can manipulate the user ID at the endpoint to access other users' information.

B. API 5:2019 Broken Function Level Authorization

API 5:2019 is also a vulnerability associated with the authorization process. Vulnerabilities occur when proper authorization is not set for each endpoint in the Web API server used by general and privileged users. Malicious users can attack a system by exploiting vulnerabilities in privileged accounts.

C. API 9:2019 Improper Assets Management

API 9:2019 is associated with improper data handling. Exposed debug endpoints and deprecated API versions can increase potential security risks. For example, if web systems have been updated to use a new API and the old endpoint remains operational, it could provide an accessible point for a malicious user.

IV. Vulnerability Assessment Tool

Several tools are available for detecting Web API vulnerabilities. However, these vulnerability assessment tools cannot detect the ten vulnerabilities described in the OWASP API Security Top 10. The reason for this failure is that they attempted to detect vulnerabilities by sending predetermined requests and could not specify the authentication information for each target Web API. In addition, various request body data (hereinafter referred to as ”parameters”) used for vulnerability assessment are set by a user who uses these tools, and if the user cannot set appropriate parameters, the assessment becomes difficult. The features of the three representative vulnerability assessment tools and the vulnerability items of the OWASP API TOP 10 that can be assessed are described below.

A. Automatic API Attack Tool

Automatic API Attack Tool [3] can flexibly generate vulnerability detection test cases according to the Web APIs to be vulnerability-tested by loading the Web API reference in JSON or YAML formats. For example, for an endpoint that uses a parameter of ”INT type” as ”id,” this tool will change the value of ”id” to a numeric value of another type, such as long or double, and send a vulnerability detection request. This tool can detect vulnerabilities in *API 1:2019*, *API 5:2019*, and *API 9:2019* of the OWASP API Security Top 10. However, Web API references in JSON or YAML formats describing the Web API information are required, and vulnerability assessments may require considerable work and time.

B. Vooki, Rest API Scanner

Vooki, Rest API Scanner [4] is a vulnerability assessment tool that sends requests to detect vulnerabilities in *API 3:2019*, *API 7:2019*, and *API 8:2019* of the OWASP API Security Top 10. However, all endpoints, headers, and parameters used in the Web API must be configured by the user of this tool. If the necessary information for vulnerability assessment cannot be set, vulnerabilities may not be detected, even if they exist.

C. OWASP ZAP

The *OWASP Zed Attack Proxy (OWASP ZAP)* [5] is one of the most representative tools used for vulnerability scanning. Primarily, OWASP ZAP is designed to scan vulnerabilities in web systems. However, it can read references written in the OpenAPI format and conduct vulnerability scans of Web APIs by incorporating an add-on or utilizing a command-line tool.

D. Problem of Traditional Tools

One of the major issues with these traditional vulnerability assessment tools is that the endpoints must be explicitly defined by developers. If any endpoint is omitted, this may lead to inaccurate vulnerability assessments. Therefore, developers must perform frequent maintenance and ensure that all endpoints are correctly and comprehensively defined.

V. Proposed Method

As we mentioned in Section IV, these traditional vulnerability assessment tools require developers to provide endpoint information to the vulnerability tools, which then scan the given endpoint. Since there are many endpoints and parameters in Web APIs, it is difficult for developers to configure them manually. In addition, the endpoints and parameters continue to change because web systems undergo continuous improvements after completion.

In this study, we propose a method for automatically conducting a Web API vulnerability assessment by utilizing references, requests, and responses for Web APIs to detect authorization-related Web API vulnerabilities. First, in the proposed method, if authentication information is required, developers (or vulnerability diagnosticians) set the authentication information to the HTTP request (e.g., HTTP header) according to the API references, and each validation check starts.

The proposed method was executed in three steps.

Step 1: Obtaining References

We collected information about Web API references to automatically generate request queries for vulnerability assessment.

Step 2: Validation Check of API References

We obtained the response information by constructing a request query using the Web API information obtained from **Step 1**. The constructed request query is sent to the endpoints to validate the Web APIs and the response (i.e., JSON data) is stored.

Step 3: Vulnerability Detection

To generate a query for vulnerability assessment, we define the string to be used as a *candidate key* and extract the values of the key, as well as from the JSON data stored in **Step 2**. Vulnerability assessment queries are generated from the *candidate key*, and the vulnerability detection process begins. If the HTTP status code corresponding to each vulnerability assessment query is 200, the proposed method indicates the existence of a vulnerability.

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

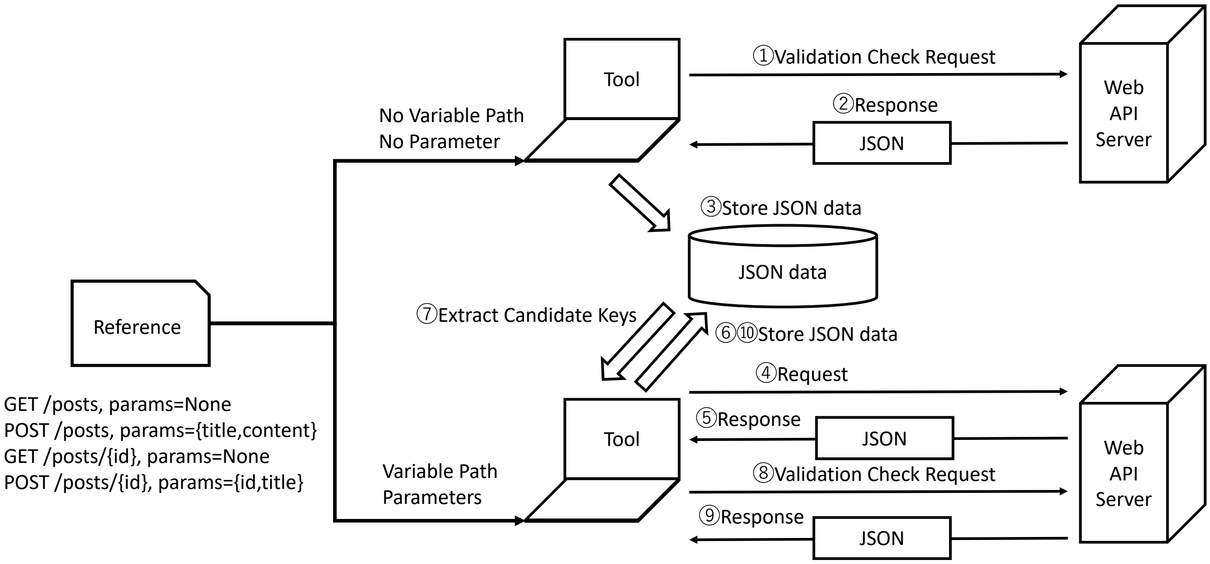


Fig. 1. API Validation Flow

TABLE III
Example of Web API References

No	HTTP Method	Endpoint	Request Parameters		Authentication
			Name	Type	
1	GET	/posts	None		Non-required
2	POST	/posts	title content	string string	Non-required
3	GET	/posts/{id}	None		Non-required
4	POST	/posts/{id}	id title	integer string	Non-required

A. Step 1: Obtaining References

Step 1 is to collect information about Web API references, such as the HTTP method, endpoints, parameters, and their types using **Method 1** and **Method 2**. Examples of the information about the Web API references are listed in Table III. In Table III, $\{id\}$ is variable which can be replaced. Hereafter, the */posts* is referred to as the *basic path* and $\{id\}$ is the *variable path*. These references were obtained using **Method 1** and **Method 2**.

Method 1: This method uses SwaggerHub [6], a resource-rich platform for Web API developers. SwaggerHub provides a wealth of references for Web APIs available in either JSON or YAML formats.

Method 2: This method employs the use of official websites. Automatic extraction of information from the HTML source becomes necessary because Web API references are provided in HTML format. In this study, we extracted information regarding Web API references from webpages written in HTML by adapting the methodology employed in a previous study [7].

If a reference does not exist in SwaggerHub (i.e., **Method 1**), the proposed method obtains the reference from the official websites of the Web APIs (i.e., **Method 2**).

B. Step 2: Validation Check of API References

Step 2 validates the Web APIs (i.e., endpoints and parameters) described in the API references. Request queries for validation are generated based on information about the API references and sent to the endpoints of the Web APIs. An overview of the validation process is presented in Figure 1. Table III shows an example of the information regarding API references.

The validation process for the Web APIs is as follows:

1) The validation check for endpoints without *variable path* and request parameters is first conducted. In Table III, the No.1 API, which has */posts* endpoint, is first selected, and the request query without request parameters is sent to the */posts* endpoint (① in Figure 1). After confirming the validation (i.e., HTTP Status Code is 200), response parameters (i.e., JSON data) that are returned from the endpoint are obtained and stored (② and ③ in Figure 1).
2) The validation check for endpoints including *variable path* without request parameters is conducted. In Table III, the No.3 API, which has */posts/*$\{id\}$ endpoint, is selected. The request query is sent to the endpoint */posts* which is an endpoint one level up from */posts/*$\{id\}$ (④ in Figure 1), Response parameters (i.e., JSON data) that are returned from the endpoint are obtained and stored (⑤ and ⑥ in Figure 1).
   Next, the proposed method searches for a parameter with the key *id* in the response parameters. When the keys *id* are found in the response parameters (⑦ in Figure 1), the *variable path* $\{id\}$ is replaced with the corresponding value for the first key *id*.
   For example, if the response parameters displayed in Figure 2 are obtained, the endpoint of the next request query is */posts/38*, and this query is sent to the */posts/38* endpoint (⑧ in Figure 1). If there is no endpoint one level up or the same key, the *variable path* $\{id\}$ is

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

```
[
    {”id”: 38, ”title”: ”ArticleName1”},
    {”id”: 1, ”title”: ”ArticleName2”}
]
```

Fig. 2. Example of response parameters from /posts endpoint in No. 3 API

```
{
    {”title”: ”ArticleName1”, ”content”: ”Content1”},
    {”title”: ”ArticleName2”, ”content”: ”Content2”}
}
```

Fig. 3. Example of response parameters from /posts endpoint in No. 2 API

replaced with the fixed value *1* (i.e., default value).

After confirming the validation (i.e., HTTP Status Code is 200), response parameters (i.e., JSON data) that are returned from the endpoint are obtained and stored (⑨ and ⑩ in Figure 1).

3) The validation check for endpoints, including request parameters without *variable path*, is conducted. In Table III, the No.2 API, which has */posts* endpoint and the request parameter with the keys *title* and *content* is selected, and the request query without request parameters is sent to the */posts* endpoint using HTTP GET Method (④ in Figure 1). Response parameters (i.e., JSON data) that are returned from the endpoint are obtained and stored (⑤ and ⑥ in Figure 1).

   When the key *title* and *content* are found in the response parameters (⑦ in Figure 1), the keys *title* and *content* of the request parameter are set to the corresponding value for the keys *title* and *content*.

   For example, if the response parameters displayed in Figure 3 are obtained, the endpoint and parameter are */posts* and {*”title”: ”ArticleName1”, ”Content1”: ‘”content1”*} respectively (⑦ in Figure 1). This query is sent to the */posts* endpoint (⑧ in Figure 1). If there is not the same key of the request parameters, the keys *title* and *content* of the request parameter are set to the fixed value *”a”* (i.e., default value).

   After confirming the validation (i.e., HTTP Status Code is 200), response parameters (i.e., JSON data) that are returned from the endpoint are obtained and stored (⑨ and ⑩ in Figure 1).

4) The validation check for endpoints including *variable path* and request parameters is conducted. In Table III, the No.4 API, which has */posts/*{*id*} endpoint and the request parameter with the keys *id* and *title*, is selected. Similar to the above *process 2)*, when the keys *id* is found in the response parameters, the *variable path* {*id*} is replaced with the corresponding value for the first key *id* (④ – ⑦ in Figure 1). Additionally, similar to the above *process 3)*, when the key *title* is found in the reponse parameters, the key *title* of the request parameter is set to the corresponding value for the key *title* (④ – ⑦ in Figure 1).

   For example, if the response parameters displayed in Figure 2 are obtained, the endpoint and parameter are */posts/38* and {*”id”: 38, ”title”: ”ArticleName1”*} respectively (⑦ in Figure 1). This query is sent to the */posts/38* endpoint (⑧ in Figure 1). If there is no endpoint one level up or the same key in the request parameters, the *variable path* {*id*} is replaced with the fixed value *1* (i.e., */posts/1*), and the keys *id* and *title* of the request parameter are set to the fixed values *1* and *”a”* (i.e., {*”id”: 1, ”title”: ”a”*}). After confirming the validation (i.e., HTTP Status Code is 200), response parameters (i.e., JSON data) that are returned from the endpoint are obtained and stored (⑨ and ⑩ in Figure 1).

The above four processes (*process 1) - process 4)*) are conducted in order from the *process 1)*.

After completing the validation check for all endpoints, the vulnerability detection process detailed in **Step 3** begins.

C. Step 3: Vulnerability Detection

The objective of **Step 3** is to detect authorization-related Web API vulnerabilities. This step generates two types of vulnerability assessment queries.

- The first query aims to detect the existence of endpoints unintended for exposure by the developers. This query is basically for API 1:2019 in the OWASP API Security Top 10.
- The second query aims to detect vulnerabilities within known endpoints by replacing the *variable path* and changing parameters using the JSON data stored in **Step 2**. This query is basically for API 5:2019 in the OWASP API Security Top 10.

An overview of the vulnerability assessment query generation process is shown in Figure 4. To generate vulnerability assessment queries, we must first extract strings to replace the *variable path* and change the parameters from the JSON data stored in **Step 2**.

Candidate keys comprise the following two types of words:

- Words that are used as the key names and values in the JSON data.
- Words that are included in the endpoints provided by the API references.

When strings that are used as the key names and values in the JSON data are provided as text, we separate the texts into words using morphological analysis. For example, the words ”id,” ”1,” ”title,” and ”ArticleA” are extracted from the JSON data in ④ of Figure 4.

1) Queries for Detection of Endpoints Including Unintended Exposure: To detect endpoints, including unintended exposure, we need to estimate unintended endpoints for developers. The *candidate keys* are further classified into two types based on morpheme analysis. One is all noun words except numerals, and the other is numerals (e.g., 0, 1, ...). For example, in ⑤ in Figure 4, the noun words ”id,” ”title,” and ”Article” are extracted from the words ”id,” ”1,” ”title,”

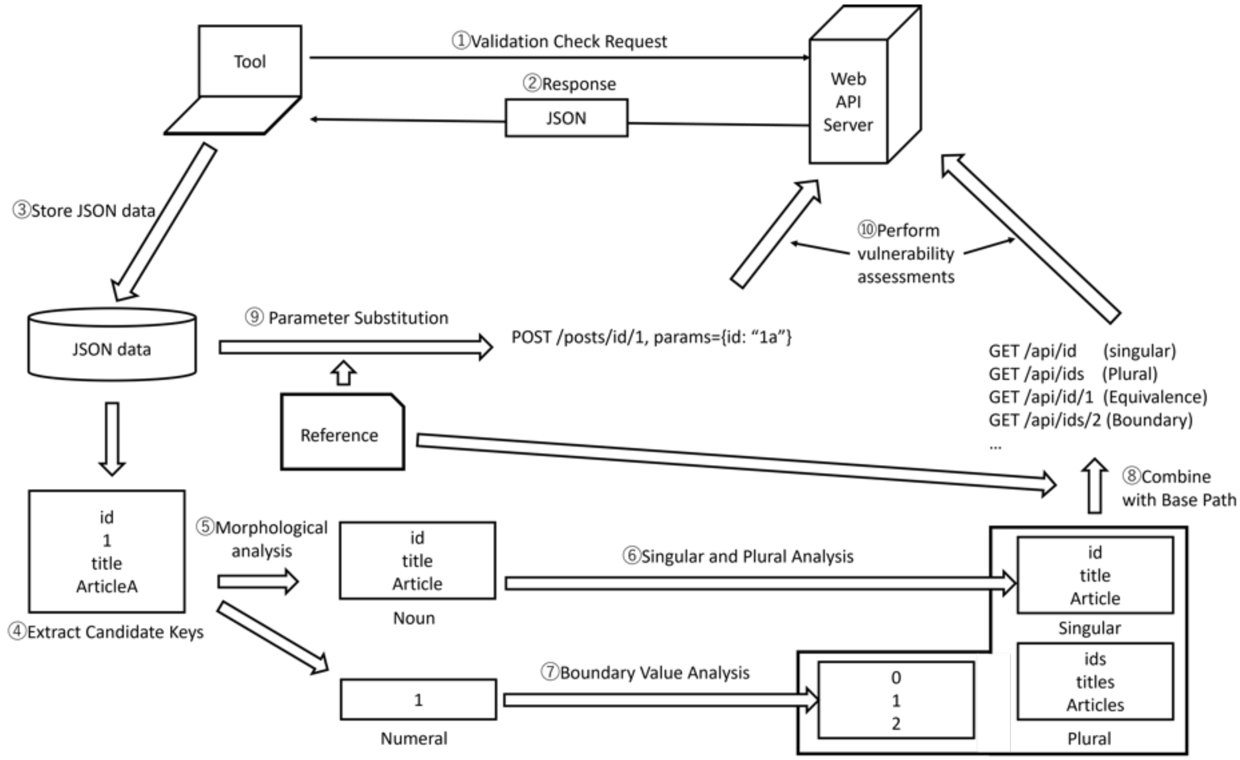


Fig. 4. Vulnerability Assessment Query Generation Flow

and "ArticleA." Additionally, the numeral *1* is extracted from the words "id," "1," "title," and "ArticleA."

For nouns, the proposed method adapts both singular and plural forms of words to detect endpoints, including unintended exposures. If only a plural form of the word is extracted, the proposed method transforms it into a singular form, and vice versa. The words "id," "ids," "title," "titles," "Article," and "Articles" are transformed from the words "id," "1," "title," and "Article" in ⑥ in Figure 4.

In the case of a numeral, given a set of numerals, the maximum value, the maximum value ± 1, the minimum value, and the minimum value ± 1 are adapted according to the boundary value analysis to detect endpoints, including unintended exposure. For example, numerals *0*, *1*, and *2* are generated from numeral *1* in ⑦ of Figure 4.

Finally, by the above-mentioned processes, we can obtain the numerals *0*, *1* and *2* and the noun words "id," "ids," "title," "titles," "Article," and "Articles" are generated in ⑥ and ⑦ of Figure 4.

The vulnerability assessment queries were constructed from these numerals and nouns. In the proposed method, the four combinations of vulnerability assessment queries (⑧ in Figure 4) are as follows:

- <API entry point>/<plural noun word>
- <API entry point>/<singular noun word>
- <API entry point>/<plural noun word>/<number>
- <API entry point>/<singular noun word>/<number>

The *API entry point* represents the base path obtained from API references. For example, in Table IV, */api* is defined as *API entry point*. The *API entry point* is the */api*, which is one level away from the */api/*{*id*} if the *variable path* is included.

TABLE IV
Example of Web API References (Non-parameter)

No	HTTP Method	Endpoint	Request Parameters		Authentication
			Name	Type	
1	GET	/api	None		Non-required
2	GET	/api/{id}	None		Non-required

Finally, if the numerals *1* and *2*, and the noun words "id" and "ids" are given, the endpoints of the vulnerability assessment queries are generated as follows:

- /api/id
- /api/ids
- /api/id/1
- /api/ids/1
- /api/id/2
- /api/ids/2

The vulnerability assessment queries which has these endpoints are illustrated in No.1 – No.6 of Table VI.

2) Queries for Detection of Vulnerabilities within Known Endpoints Using Replacement and Changing: To detect vulnerabilities in terms of parameters within known endpoints, we must estimate the resources that are not managed at the configuration or code levels. The proposed method first uses the values for the keys of the request parameters using JSON data, which are stored in **Step 2**, and the API references. Next, the proposed method changes the type of parameters and generates new values for the keys of the request parameters for vulnerability assessment using the JSON data and API

Copyright ⓒ 2024 Global IT Research Institute (GiRI)

TABLE V
Example of Web API References (Parameters)

No	HTTP Method	Endpoint	Request Parameters		Authentication
			Name	Type	
1	POST	/posts	id	integer	Non-required
2	POST	/posts	status	string	Non-required

```
{{”id”: 5}}
{{”status”: ”publish”}}
```

Fig. 5. Example of JSON data

references (⑨ of Figure 4).

We explain the generation of parameters for vulnerabilities using the API references listed in Table V and the JSON data illustrated in Figure 5.

If a key has a numeric type in the API references, three values for the key are generated: first, the same value as that of the numeric type is set; second, the fixed value ”a” as that of the string type is set; third, the proposed method changes the numeric type to the string type keeping the numeral, and concatenates the numeral of the string type and the fixed value ”a” of the string type. For example, if the key ”id” is 5 (i.e., {”id”:5}), the proposed method first uses {”id”:5}. Second, the proposed method generates {”id”:”a”} where ”a” is used as the fixed value of string type in this study. Third, the proposed method generates {”id”:”5a”} after concatenating ”5” and ”a” of the string type.

Finally, the request parameters {”id”:5}, {”id”:”a”} and {”id”:”5a”} are generated from {”id”:5}. The vulnerability assessment queries which has these parameters are illustrated in No.7 – No.9 of Table VI.

If a key has a string type in the API references, the following three values for the key are generated. First, the same value of string type is set. Second, when the value of the string type is a numeral, the numeral is converted to a string type; otherwise, the value ”1” is used as the fixed value of the string type in this study. Third, the proposed method concatenates the original value of the string type and the fixed value ”1” of the string type. For example, if the key ”status” is ”publish” (i.e., {”status”:”publish”}), the proposed method first uses {”status”: ”publish”}. Second, the proposal method replaces ”publish” with the fixed value ”1” (i.e., {”status”:”1”}). Third, the proposal method replaces ”publish” with ”publish1” (i.e., {”status”:”publish1”}).

Finally, the request parameters {”status”:”publish”}, {”status”:”1”} and {”status”:”publish1”} are generated from {”status”:”publish”}. The vulnerability assessment queries which has these parameters are illustrated in No.10 – No.12 of Table VI.

3) Vulnerability Detection: The above-mentioned vulnerability assessment queries (Table VI) were generated, and the vulnerability detection process was initiated.

TABLE VI
Example of the vulnerability assessment queries

No	HTTP Method	Endpoint	Request Parameters
1	GET	/api/id	None
2	GET	/api/ids	None
3	GET	/api/id/1	None
4	GET	/api/ids/1	None
5	GET	/api/id/2	None
6	GET	/api/ids/2	None
7	POST	/posts	{”id” : 5}
8	POST	/posts	{”id” : ”a”}
9	POST	/posts	{”id” : ”5a”}
10	POST	/posts	{”status” : ”publish”}
11	POST	/posts	{”status” : ”1”}
12	POST	/posts	{”status” : ”publish1”}

The HTTP status code of the response is used to determine the vulnerability assessment results. If the HTTP status code corresponding to each vulnerability assessment query is 200, the proposed method indicates the existence of a vulnerability. Additionally, if the HTTP status code is 501, the proposed method indicates the existence of a vulnerability because the API server attempts to perform certain processes. When the HTTP status codes are not equal to 200 and 501, the proposed method indicates the absence of vulnerabilities.

VI. Experiment Environment

We conducted a preliminary experiment using actual environment to evaluate whether the proposed method can detect the OWASP API Security Top 10 vulnerabilities. We used the well-known CMSs as the experimental environments. In our selection criteria, we focus on whether the CMSs include known vulnerabilities and whether they have higher market shares [8]. Consequently, we selected three well-known CMSs and a vulnerable training environment.

- vAPI [9]
- WordPress [10]
- Ghost CMS [11]
- Joomla [12]

A. vAPI

vAPI was developed to test the vulnerabilities in the OWASP API Security Top 10. It replicates the vulnerabilities described in the OWASP API Security Top 10, serving as an application that enables users to experience attacks.

B. WordPress

WordPress is a leading CMS with a market share of 63.7% as of January 2023, according to a report by W3Techs [13]. It is utilized by global organizations, such as Microsoft, Mozilla, and Apache. However, it has been targeted by many cyber attackers owing to its widespread use. WordPress versions 4.7.0 and 4.7.1 contain a vulnerability that allows article tampering by bypassing authentication [14]. This risk was categorized as *API 1:2019* in the OWASP API Security Top 10.

Copyright © 2024 Global IT Research Institute (GiRI)

C. Ghost CMS

Ghost CMS has a relatively modest market share of 0.08% compared with the other CMSs. However, it has steadily gained traction as an emerging CMS [15]. Additionally, it has been adopted by well-known websites such as Cloudflare and Duolingo. Ghost CMS versions from 4.0.0 to 4.9.4 contain vulnerabilities that allow an undisclosed endpoint to access information that typically requires administrative privileges [16]. This vulnerability was categorized as *API 5:2019* in the OWASP API Security Top 10.

D. Joomla

Joomla has a small market share of 2.7% as of January 2023, according to the report of W3Techs [17]. Despite its small market share, it is renowned for its strong emphasis on security and is frequently used on government websites.

VII. Experiment Results

We conducted a preliminary experiment using well-known CMSs, such as WordPress, Ghost CMS, and Joomla. We also used a deliberately vulnerable testing environment called vAPI. For vAPI, the proposed method could detect vulnerabilities (*API 1:2019* and *API 5:2019* of the OWASP API Security Top 10) because vAPI was developed to test vulnerabilities in the OWASP API Security Top 10. WordPress has a known vulnerability (*API 1:2019* of the OWASP API Security Top 10), and Ghost CMS has a known vulnerability (*API 5:2019*). Therefore, in this study, we first evaluate whether these known vulnerabilities can be detected. Additionally, the proposed method attempts to detect vulnerabilities that are not known (i.e., vulnerabilities where CVE ID are not assigned). CVE (Common Vulnerabilities and Exposures) is a list of publicly disclosed vulnerabilities and each vulnerability is assigned a CVE ID number. Joomla does not have a known vulnerability described in the OWASP API Security Top 10, and the proposed method attempts to detect vulnerabilities that are not known (i.e., vulnerabilities where CVE ID are not assigned).

A. Vulnerability Detection for vAPI

First, we evaluated whether the proposed method could detect *API 1:2019* and *API 5:2019* from the OWASP API Security Top 10. Given that vAPI provides a webpage with API references for each endpoint, we used this information to create endpoint-specific settings.

1) Detection for API 1:2019: API 1:2019 of the vAPI contains a vulnerability in the authorization process. This vulnerability allows an authenticated user to access information regarding other users. First, in the proposed method, developers (or vulnerability diagnosticians) set the authentication information to the HTTP request (e.g., HTTP header) according to the API references, and **Step 1** begins. In **Step 1**, we can obtain the API references listed in Table VII. The number of APIs in *API 1:2019* of the vAPI was 3. In **Step 2**, the No.1 API which has the */vapi/api1/user/*{*api1_id*} endpoint, is first selected. {*api1_id*} is not included in the response parameter from */vapi/api1/user*, which is one level up from */vapi/api1/user/*{*api1_id*}. Therefore, the *variable path* {*api1_id*} is replaced by the default value of *1*, and a request with authorization information is sent to the */vapi/api1/user/1* endpoint. Similarly, the No.2 API and 3 API are selected, and a request with authorization information was sent to the endpoints. After completing **Step 2**, the number of *candidate keys* is shown in Table VIII is obtained from the JSON data. Table IX presents part of the *candidate keys*. In **Step 3**, the request with the authorization information shown in Figure 6 is sent to the */vapi/api1/user/1* endpoint, which is constructed using *the candidate keys* of the *user* and *1*, and the response shown in Figure 7 is obtained. Because the vAPI issues a flag when an attack is successful and an authenticated user can access information about another user (i.e., id = 1), the proposed method indicates the existence of the *API 1:2019* vulnerability. In this step, 2714 vulnerability assessment queries were sent to the vAPI server.

TABLE VII
Web API References (*API 1:2019* of vAPI)

No	HTTP Method	Endpoint	Request Parameters Name	 Type	Authentication
1	GET	/vapi/api1/user/{api1_id}	None		Required
2	POST	/vapi/api1/user	username name course password	string string string string	Required
3	PUT	/vapi/api1/user/{api1_id}	username name course password	string string string string	Required

TABLE VIII
Number of Candidate Keys for *API 1:2019* of vAPI

Item	Value
The Number of String Candidate Keys	18
The Number of Numerical Candidate Keys	5

TABLE IX
Part of Candidate Keys for *API 1:2019* of vAPI

String Candidate Key	Numerical Candidate Key
site	0
id	1
entry	9
user	237235
error	2324
username	

```
GET http://[Server IP Address]/vapi/api1/user/1
Content−Type: application/json
Authorization−Token: (** mask **)
```

Fig. 6. Request for *API 1:2019* of vAPI

```
{
  ”id”: 1,
  ”username”: ”michaels”,
  ”name”: ”Michael Scott”,
  ”course”: ”flag{api1_(** mask **)}”
}
```

Fig. 7. Response for *API 1:2019* of vAPI

```
GET /vapi/api5/users
Content−Type: application/json
Authorization−Token: (∗∗ mask ∗∗)
```

Fig. 8. Request for *API 5:2019* of vAPI

```
{
    ”id”: 1,
    ”username”: ”admin”,
    ”name”: ”Admin User”,
    ”address”: ”flag{api5_(∗∗ mask ∗∗)}”,
    ”mobileno”: ”8080808080”
}
```

Fig. 9. Response for *API 5:2019* of vAPI

TABLE X
Web API References (*API 5:2019* of vAPI)

No	HTTP Method	Endpoint	Request Parameters		Authentication
			Name	Type	
1	GET	/vapi/api5/user/{api5_id}	None		Required
2	POST	/vapi/api5/user	username password name address mobileno	string string string string string	Required

TABLE XI
Number of Candidate Keys for *API 5:2019* of vAPI

Item	Value
The Number of String Candidate Keys	6
The Number of Numerical Candidate Keys	0

TABLE XII
Part of Candidate Keys for *API 5:2019* of vAPI

String Candidate Key
u
false
s
cause
user

2) Detection for API 5:2019: API 5:2019 implementation of the vAPI also exposes a vulnerability in the authorization process. This vulnerability permits general users to access confidential endpoints at which administrator privileges are required. If this endpoint is accurately deduced, a list of users with administrative privileges can be procured. First, in the proposed method, developers (or vulnerability diagnosticians) set the authentication information to the HTTP request (e.g., HTTP header) according to the API references, and **Step 1** begins. In **Step 1**, we can obtain the API references listed in Table X. The number of APIs in *API 5:2019* of the vAPI was 2. In **Step 2**, the No.1 API which has the */vapi/api5/user/*{*api5_id*} endpoint, is first selected. {*api5_id*} is not included in the response parameter from */vapi/api5/user*, which is one level up from */vapi/api5/user/*{*api5_id*}. Therefore, the *variable path* {*api5_id*} is replaced by the default value of *1*, and a request with authorization information is sent to the */vapi/api5/user/1* endpoint. Similarly, the No.2 API are selected, and a request with authorization information was sent to the endpoints. After completing **Step 2**, the number of *candidate keys* shown in Table XI is obtained from the JSON data. Table XII presents part of the *candidate keys*. In **Step 3**, the request with the authorization information shown in Figure 8 is sent to the */vapi/api5/users* endpoint, which is constructed by *the candidate keys* of *users*, which is a plural of the singular form *user*, as shown in Figure 8. Because the vAPI issues a flag when an attack is successful, and an general user can access the confidential endpoint for which administrator privileges are needed, the proposed method indicates the existence of the *API 5:2019* vulnerability. In this step, 15586 vulnerability assessment queries were sent to the vAPI server. The number of queries was relatively large because the number of generated queries rapidly increased as the number of parameters increased.

B. Vulnerability Detection for WordPress

The 4.7.0 version of WordPress has the privilege vulnerability identified as CVE-2017-1001000. This vulnerability prevails in the authorization process of the */posts/*{*id*} endpoint. If a request parameter including an invalid content identifier (i.e., a string type value composed of a numeral and string) is sent to the */posts/*{*id*} endpoint, it becomes possible to overwrite the content. This vulnerability was categorized as *API 1:2019* and *API 2:2019* in the OWAPS API Security Top 10. We evaluated whether this vulnerability and other unknown vulnerabilities could be detected by implementing the proposed method.

First, in the proposed method, developers (or vulnerability diagnosticians) set the authentication information to the HTTP request (e.g., HTTP header) according to the API references, and **Step 1** begins. In **Step 1**, we obtain the API references listed in Table XIII. The number of APIs that required authorization was 47. Part of these APIs are shown in Table XIII. In **Step 2**, the No.1 API which has the */wp-json/wp/v2/posts/*{*post_id*} endpoint, is selected. {*post_id*} is not included in the response parameter from */wp-json/wp/v2/posts*, which is one level up from */wp-json/wp/v2/posts/*{*post_id*}. Therefore, the *variable path* {*post_id*} is replaced by the default value of *1*, and a request with authorization information is sent to the */wp-json/wp/v2/posts/1* endpoint. Similarly, other APIs are selected, and a request with authorization information is sent to the endpoints. After completing **Step 2**, the number of *candidate keys* shown in Table XIV is obtained from the JSON data. Table XV presents an example of the *candidate keys*. In **Step 3**, the request with authorization information shown in Figure 10 with the parameter {*”id”:”1a”,”title”:”a”*} is sent to the */wp-json/wp/v2/posts/1* endpoint and the response shown in Figure 11 is obtained. The method for generating the parameter {*”id”:”1a”,”title”:”a”*} is described in Section V-C2. As illustrated in Figure 11, the value of the ”title” are tampered to ”a.” This result is considered to be the vulnerability identified as CVE-2017-1001000. Additionally, other unknown vulnerabilities were not detected using the

TABLE XIII
Part of Web API References (WordPress)

No	HTTP Method	Endpoint	Request Parameters		Authentication
			Name	Type	
1	GET	/wp-json/wp/v2/posts/{post_id}	None		Required
2	POST	/wp-json/wp/v2/posts	id title content status	integer string string string	Required
3	GET	/wp-json/wp/v2/posts/{parent}/revisions/{id}	None		Required
..	..	..	..		..

TABLE XIV
Number of Candidate Keys for WordPress

Item	Value
The Number of String Candidate Keys	88
The Number of Numerical Candidate Keys	29

TABLE XV
Part of Candidate Keys for WordPress

String Candidate Key	Numerical Candidate Key
site	0
id	1
http	27079
message	27080
world	27081
setting	

```
POST /wp−json/wp/v2/posts/1
Content−Type: application/json

{”id”: ”1a”, ”title”: ”a”}
```

Fig. 10. Request for WordPress

```
{
    ”id”: 1,
    ”date”: ”2023−07−07T20:11:50”,
    ”modified”: ”2023−07−18T22:27:16”,
    ”slug”: ”hello−world”,
    ”type”: ”post”,
    ”title”: {
        ”raw”: ”a”,
        ”rendered”: ”a”
    },
}
```

Fig. 11. Response for WordPress

proposed method. In this step, 5288 vulnerability assessment queries were sent to the WordPress server.

C. Vulnerability Detection for Ghost CMS

Ghost CMS versions from 4.0.0 to 4.9.4 contain a vulnerability pertinent to inadequate authorization management. This susceptibility permits the extraction of an administrator API key, including a response to a specific API, thereby enabling privilege escalation for any authenticated user. This vulnerability was categorized under *API 5:2019* of the OWAPS API Security Top 10.

First, in the proposed method, developers (or vulnerability diagnosticians) set the authentication information to the HTTP request (e.g., HTTP header) according to the API references, and **Step 1** begins. In **Step 1**, we can obtain the API references listed in Table XVI. The number of APIs that required authorization was 29. Part of these APIs are shown in Table XIII. In this experiment, we used version 2 of the API of Ghost CMS, and the *variable path* {*version*} was replaced with *v2*. In **Step 2**, the No.1 API, which contains the */ghost/api/v2/admin/posts* endpoint, is selected, and a request with authorization information is sent to the */ghost/api/v2/admin/posts* endpoint. Similarly, other APIs are selected, and a request with authorization information is sent to the endpoints. After completing **Step 2**, the number of *candidate keys* shown in Table XVII is obtained from the JSON data. Table XVIII presents part of the *candidate keys*. In **Step 3**, a request with authorization information, as shown in Figure 12 is sent to the */ghost/api/v2/admin/users* endpoint, and the response shown in Figure 13 is obtained. Consequently, 11 endpoints were identified as vulnerabilities. The 11 endpoints are listed in Table XIX. The list of users was procured using the administrator endpoint. Among the 11 endpoints, the */admin/site/* endpoint was defined as accessible by general users in the API references. Therefore, 10 endpoints were identified as vulnerabilities. In particular, the */admin/session/* and */admin/integrations/* endpoints were not described in the API references. The result was considered a vulnerability, identified as CVE-2021-39192. In this step, 2108 vulnerability assessment queries were sent to the Ghost CMS server.

D. Vulnerability Detection for Joomla

In the Joomla Web API, no obvious vulnerabilities associated with the OWASP API Security Top 10 were reported. However, the proposed method attempted to detect vulnerabilities that are not known (i.e., vulnerabilities where CVE ID are not assigned).

First, in the proposed method, developers (or vulnerability diagnosticians) set the authentication information to the HTTP request (e.g., HTTP header) according to the API references, and **Step 1** begins. Similar to the detections for vAPI, WordPress, and Ghost CMS, **Steps 1, 2, and 3** were processed. After completing **Step 2**, the number of *candidate keys* shown in Table XX were obtained from the JSON data. Table XXI presents part of the *candidate keys*. In **Step 3**, a request with the authorization information shown in Figure 14 is sent to the */api/index.php/v1/extensions*

TABLE XVI
Part of Web API References (Ghost CMS)

No	HTTP Method	Endpoint	Request Parameters		Authentication
			Name	Type	
1	GET	/ghost/api/{version}/admin/posts	None		Required
2	GET	/ghost/api/{version}/admin/posts/{id}	None		Required
3	POST	/ghost/api/{version}/admin/posts	title	string	Required
4	GET	/ghost/api/{version}/admin/settings	None		Required
5	GET	/ghost/api/{version}/admin/posts/{parent}/revisions/{id}	None		Required

```
GET /ghost/api/v2/admin/users/
Content−Type: application/json
Authorization: Ghost (** mask **)
```

Fig. 12. Request for Ghost CMS

```
{
  ”users”: [ {
        ”id”: ”1”,
        ”name”: ”Yuki Ishida”,
        ”slug”: ”yuki”,
        ”email”: ”(** mask **)”,
        ”profile_image”: null,
        ”cover_image”: null,
        ”bio”: null,
        ”website”: null,
        ”location”: null,
        ”facebook”: null,
        ”twitter”: null,
        ”accessibility”: null,
        ”status”: ”active”,
        ”meta_title”: null,
        ”meta_description”: null,
        ”tour”: null,
        ”last_seen”: ”2023−07−22T10:53:11.000Z”,
        ”created_at”: ”2023−05−06T08:14:32.000Z”,
        ”updated_at”: ”2023−07−22T10:53:11.000Z”,
        ”url”: ”http://[Server IP Address]/404/”
  }, ... ],
}
```

Fig. 13. Response for Ghost CMS

```
POST /api/index.php/v1/extensions
Content−Type: application/json
Authorization: Bearer (** mask **)
```

Fig. 14. Request for Joomla

endpoint, and the response shown in Figure 15 was obtained.

Consequently, 3 unpublished endpoints were identified as vulnerabilities because the endpoints were not described in the API references. The 3 unpublished endpoints are listed in Table XXII. This vulnerability was categorized under *API 9:2019* of the OWAPS API Security Top 10. In particular, the */api/index.php/v1/extensions* endpoint illustrated in Figure 15 shows a catalog of extended functions installed within Joomla. This vulnerability implies the potential risk of exploiting the system configuration information if privilege escalation is perpetrated owing to other vulnerabilities.

TABLE XVII
Number of Candidate Keys for Ghost CMS

Item	Value
The Number of String Candidate Keys	974
The Number of Numerical Candidate Keys	887

TABLE XVIII
Part of Candidate Keys for Ghost CMS

String Candidate Key	Numerical Candidate Key
sell	0
pages	1
tags	40633578
upload	40633579
site	40633580
session	

TABLE XIX
Vulnerability Assessment Results for Ghost CMS

Endpoint	HTTP StatusCode
/admin/pages/	200
/admin/tags/	200
/admin/posts/	200
/admin/roles/	200
/admin/users/	200
/admin/settings/	200
/admin/site/	200
/admin/users/1/	200
/admin/themes/	200
/admin/session/	200
/admin/integrations/	200

E. System Impact of Vulnerability Assessment using the Proposed Method

We evaluated the impact of each CMS in terms of system load. Table XXIII lists the number of endpoints recorded in the API reference for each CMS, the number of *candidate keys* extracted from JSON data, and the number of vulnerability assessment queries.

Compared with the number of queries between WordPress and Ghost CMS, although the number of endpoints in WordPress is greater than that in Ghost CMS, the number of *candidate keys* in WordPress is lower than that in Ghost CMS. One reason for this is that there are more types of strings in keys and values in WordPress is more than that in Ghost CMS. However, although the number of *candidate keys* in WordPress was less than that in Ghost CMS, the number of vulnerability queries in WordPress was greater than that in Ghost CMS. One of the reasons for this is that the number of keys and values in WordPress is greater than that in Ghost CMS. Specifically, although the number of endpoints and *candidate keys* in the

Copyright © 2024 Global IT Research Institute (GiRI)

```
{
  ”links”: {
    ”self”: ”http://[Server IP Address]/api/index.php/v1/
        extensions”,
    ”next”: ”http://[Server IP Address]/api/index.php/v1/
        extensions?page%5Boffset%5D=20&page%5
        Blimit%5D=20”,
    ”last”: ”http://[Server IP Address]/api/index.php/v1/
        extensions?page%5Boffset%5D=220&page%5
        Blimit%5D=20”
  },
  ”data”: [{
      ”type”: ”manage”,
      ”id”: ”91”,
      ”attributes”: {
        ”name”: ”Authentication − Joomla”,
        ”type”: ”plugin”,
        ”folder”: ”authentication”,
        ”client_id”: 0,
        ”status”: 2,
        ”version”: ”3.0.0”,
        ”id”: 91
      }
  }, ... ]
}
```

Fig. 15. Response for Joomla

TABLE XX
Number of Candidate Keys for Joomla

Item	**Value**
The Number of String Candidate Keys	134
The Number of Numerical Candidate Keys	57

TABLE XXI
Part of Candidate Keys for Joomla

String Candidate Key	**Numerical Candidate Key**
site	0
featured	1
next	25079
page	25080
privacy	25081
category	

TABLE XXII
Vulnerability Assessment Results for Joomla

Endpoint	HTTP StatusCode
/api/index.php/v1/redirects	200
/api/index.php/v1/contacts	200
/api/index.php/v1/extensions	200

vAPI(API 5) was the lowest, the number of vulnerability queries was the highest. Although this number is considered tolerable during the development phase, there will be a need to devise an approach that does not generate obvious invalid vulnerability requests, such as reductions in *candidate keys*.

VIII. Conclusion

TABLE XXIII
Vulnerability Assessment Results in terms of System Load

	vAPI1	vAPI5	WordPress	Ghost	Joomla
The number of endpoints	3	2	47	29	47
The number of candidate keys	23	6	116	1861	191
The number of queries	2714	15586	5288	2108	657
The number of http status 200	5	15554	184	11	12
The number of http status 500	2704	22	0	0	46
The number of invalid requests	5	10	5104	2097	599

In this study, we proposed a vulnerability assessment method for addressing *API 1:2019*, *API 5:2019*, and *API 9:2019* vulnerabilities in the OWASP API Security Top 10. The proposed method conducts a dynamic vulnerability assessment using requests to validate the Web APIs according to API references and vulnerability requests that take advantage of their corresponding responses.

In this experiment, we confirmed the following:

- *API 1:2019* and *API 5:2019* vulnerabilities in the OWASP API Security Top 10 can be exactly detected using the vAPI.
- In WordPress, the known vulnerabilities of *API 1:2019* and *API 5:2019* (i.e., CVE-2017-1001000) can be exactly detected.
- In Ghost CMS, the known vulnerabilities of *API 5:2019* (i.e., CVE-2021-39192) can be exactly detected. Additionally, 10 endpoints are identified as vulnerabilities. Specifically, the */admin/session/* and */admin/integrations/* endpoints are not described in the API reference.
- In Joomla, the unknown vulnerabilities of *API 9:2019* can be detected. Additionally, 3 endpoints are identified as unpublished endpoints.

In addition, we evaluated the impact on each CMS in terms of system load. The number of assessment queries of the proposed method depends on the type of strings in the keys and values and the number of keys. Therefore, there is a need to develop an approach that does not generate obvious invalid vulnerability requests, such as a reduction in the number of *candidate keys*.

With the application of the proposed method, we anticipate an effective vulnerability assessment of Web APIs during the developmental phase and prior to production.

References

[1] Akamai, “State of the internet / report — api: The attack surface that connects us all — akamai,” https://www.akamai.com/site/en/documents/state-of-the-internet/soti-security-api-the-attack-surface-that-connects-us-all.pdf, (Accessed on 07/10/2023).

[2] Open Worldwide Application Security Project, “Owasp top 10 api security risks – 2019 - owasp api security top 10,” https://owasp.org/API-Security/editions/2019/en/0x11-t10/, (Accessed on 07/10/2023).

[3] imperva, “imperva/automatic-api-attack-tool: Imperva’s customizable api attack tool takes an api specification as an input, generates and runs attacks that are based on it as an output.” https://github.com/imperva/automatic-api-attack-tool, (Accessed on 07/13/2023).

[4] VegaBird Technologies, “Vooki - web application and api vulnerability scanner — vooki infosec,” https://www.vegabird.com/vooki/, (Accessed on 07/13/2023).

Copyright © 2024 Global IT Research Institute (GiRI)

[5] Open Worldwide Application Security Project, “OWASP ZAP,” https://www.zaproxy.org/, (Accessed on 07/25/2023).

[6] SwaggerHUB, “SwaggerHUB,” https://app.swaggerhub.com/search, (Accessed on 07/25/2023).

[7] Takai Masanari and Sakaguchi Tetsuo, “Automatic generation of program libraries for accessing web apis,,” *IPSJ Information Fundamentals and Access Technologies (IFAT)*, vol. 2012-IFAT-108, no. 1, pp. 1–8, 09 2012, (in Japanese).

[8] Q-Success: World Wide Web Technology Surveys, “Market share yearly trends for content management systems, july 2023,” https://w3techs.com/technologies/overview/content_management, (Accessed on 07/15/2023).

[9] roottusk, “roottusk/vapi: vapi is vulnerable adversely programmed interface which is self-hostable api that mimics owasp api top 10 scenarios through exercises.” https://github.com/roottusk/vapi, (Accessed on 07/16/2023).

[10] WordPress Foundation, “Blog tool, publishing platform, and cms – wordpress.org,” https://wordpress.org/, (Accessed on 07/16/2023).

[11] Ghost Foundation, “Ghost: The creator economy platform,” https://ghost.org/, (Accessed on 07/16/2023).

[12] Open Source Matters, Inc., “Joomla content management system (cms) - try it! it’s free!” https://www.joomla.org/, (Accessed on 07/16/2023).

[13] Q-Success: World Wide Web Technology Surveys, “Usage statistics and market share of wordpress, july 2023,” https://w3techs.com/technologies/details/cm-wordpress, (Accessed on 07/15/2023).

[14] National Institute of Standards and Technology, “Nvd - cve-2017-1001000,” https://nvd.nist.gov/vuln/detail/CVE-2017-1001000, (Accessed on 07/17/2023).

[15] Q-Success: World Wide Web Technology Surveys, “Usage statistics and market share of ghost, july 2023,” https://w3techs.com/technologies/details/cm-ghost, (Accessed on 07/15/2023).

[16] National Institute of Standards and Technology, “Nvd - cve-2021-39192,” https://nvd.nist.gov/vuln/detail/CVE-2021-39192, (Accessed on 07/17/2023).

[17] Q-Success: World Wide Web Technology Surveys, “Usage statistics and market share of joomla, july 2023,” https://w3techs.com/technologies/details/cm-joomla, (Accessed on 07/15/2023).

Yuki Ishida was born in Japan 1991, received the B.E. and M.S. degrees in Informatics from Tokyo University of Information Sciences, Japan, in 2014 and 2016, respectively. Upon graduation in 2016, he joined Digital Arts, Inc., where he contributed to the development of security products. In 2019, he transitioned to SecureBrain Corporation, also in Japan, with a primary focus on research and development in the field of cybersecurity. Since 2022, he has been concurrently enrolled in the doctoral program at the Graduate School of Informatics at Tokyo University of Information Sciences in Japan. His research interests encompass cybersecurity and network quality control. He holds memberships in IEEE, IEICE, and IPSJ.

Masaki Hanada was born in Japan 1973, received the B.E. degree in resources engineering from Waseda University in 1996, the M.S. degree in information science from Japan Advanced Institute of Science and Technology (JAIST) in 1999, and the M.S. and D.S. degrees in global information and telecommunication studies from Waseda University in 2003 and 2007, respectively. He worked at Waseda University and Tokyo University of Science. After joining Tokyo University of Information Sciences as an Assistant Professor in 2011, he has been a Professor in the Department of Information Systems, Tokyo University of Information Sciences, since 2019. His research interests include network QoS control, network resource control and management, and network security. He is a member of the IEEE, IEICE and IPSJ.

Atsushi Waseda was born in Japan 1977, received the B.E. degree in communication engineering from the University of Electro-Communications in 2000. He received his M.S. and Ph.D. in information science from Japan Advanced Institute of Science and Technology (JAIST) in 2002, and 2007, respectively. He worked at the National Institute of Information and Communications Technology and KDDI research inc. After joining Tokyo University of Information Sciences as an Assistant Professor in 2019. His research interests include information security, quantum security and privacy protection. He is a member of the IEICE and IPSJ.

Moo Wan Kim was born in Korea 1951, received B.E., M.E. and Ph.D degree in electronic engineering from Osaka University, Osaka, Japan in 1974, 1977 and 1980, respectively. He joined Fujitsu Lab. in 1980 and had been engaged in research and development on multimedia communication systems, Intelligent Network, ATM switching system and operating system. In 1998 he joined Motorola Japan and had been engaged in research and development on CDMA2000 system. In 2000 he joined Lucent Japan and had been engaged in research and development on W-CDMA system, IMS and Parlay. In 2005 he joined Tokyo University of Information Sciences and had been engaged in research on Ubiquitous Network. In 2022 he joined TA Tech. as a lecture.

Copyright © 2024 Global IT Research Institute (GiRI)

Articles in this publication may be cited in other publications. In order to facilitate access to the original publication source, the following form for the citation is suggested:

Name of Author(s), "Title of Paper," in The 26th International Conference on Advanced Communications Technology, Technical Proceedings, 2024, page numbers

IEEE Catalog Number: CFP24561-ART
ISSN 1738-9445
ISBN 979-11-88428-12-0

Copyright© 2024, Global IT Research Institute

Supported By

IEEE Communications Society, Gangwon Convention & Visitors Bureau, National Information Society Agency, Electronic and Telecommunications Research Institute, Korea Institute of Communication Sciences, IEEK Communications Society, Korean Institute of Information Scientists and Engineers, Open Standards and Internet Association, Korean Institute of Information Security and Cryptology, Information Technology Institute of Vietnam National University

Global IT Research Institute, All Rights Reserved. Contact: office@icact.org Tel:+82-70-4146-4991
1713 Obelisk, 216 Seohyunno, Bundang-gu, Sungnam Kyunggi-do, Republic of Korea 13591