The development of 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 realworld 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, AIbased 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.**

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

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.