Android malware detection: An in-depth investigation of the impact of the use of imbalance datasets on the efficiency of machine learning models.

ZAKARIA SAWADOGO *†,JEAN-MARIE DEMBELE†, GERVAIS MENDY *,SAMUEL OUYA *.
*LITA(Laboratory of Computer Science, Telecommunications and Applications), Cheikh Anta Diop University of Dakar, Senegal
†LANI(Laboratory of Numerical Analysis and Computer Science), Gaston Berger University,Senegal
{sawadogo.zakaria,jean-marie.dembele}@ugb.edu.sn, {gervais.mendy,samuel.ouya}@ucad.edu.sn

***Abstract*—Machine learning techniques have become an essential part of research into the detection and classification of malicious applications. There are several approaches or algorithms that learn from existing data and predict classes. Machine learning principles recommend a balance of classes in the training dataset, but the reality on the ground is quite different. The majority of datasets used for malicious application detection are unbalanced. Class imbalance degrades classifier performance, so it is a common problem in classification tasks. This observation is much more significant in the field of Android malware detection and classification. There is little work to our knowledge on the effects of unbalanced datasets in the field of Android malware detection. Our contribution focuses on the impact of unbalanced datasets on the performance of different algorithms and the relevance of using evaluation metrics in Android malware detection. And the state of the databases from which researchers typically draw datasets. We show that for malicious application detection, some classification algorithms are not suitable for unbalanced datasets. We also prove that some of the most widely used performance evaluation metrics in the literature (Accuracy, Precision, Recall) are not very well suited to unbalanced datasets. On the other hand, the metrics (Balanced Accuracy, Geometric mean) are more suitable. These results were obtained by evaluating the performances of eleven classification algorithms as well as the adequacy of the different evaluation metrics (Accuracy, Recall, Precision, F1_score, Balanced accuracy, Matthews corrcoef, Geometric mean, Fowlkes_mallows). Also not all databases are accessible by researchers and many of these databases are not updated.**

***Index Terms*—imbalanced dataset, Android malware detection, Malware classification, Artificial intelligence, Machine learning.**

Zakaria SAWADOGO received his Master II degree in software engineering and information systems from Joseph Ki-Zerbo University in Burkina Faso. He is currently a researcher in cybersecurity including artificial intelligence and is affiliated with the Laboratoire d'Informatique, de Télécommunications et Applications (LITA) of the University Cheikh Anta Diop of Dakar and the Laboratoire d'Analyse Numérique et d'Informatique (LANI) of the University Gaston Berger of Saint Louis. His research interests include mobile security, system security and artificial intelligence. He is a member of the IEEE.

Pr. Gervais MENDY is a researcher-scientist at the ESP polytechnic school at UCAD university where he was head of the IT department from 2012 to 2016. Holder of a PhD in Computer Science from Paris-Sud XI University. His research interests are in Computer Combinatory, Social Network Analysis, Internet of Things (IoT), Artificial Intelligence and IT Security.He is a member of the LITA Laboratory of UCAD.

Pr. Jean-Marie DEMEBELE is Assimilated Professor in Computer Science at the UGB, ex Director of the UFR SAT and member of the UMI UMMISCO Senegal. He has to his credit, several scientific publications in international peer-reviewed journals. His research interests include artificial intelligence, agent-based modeling, dynamical systems, evolutionary algorithms, genetic regulatory networks, machine learning

Pr. Samuel Ouya is currently the director of the LITA laboratory at the UCAD. He was from 2013 to May 2017 the first Director of Infrastructure and Information System of the first virtual university of Senegal (UVS). Holder of a thesis in Applied Mathematics from the Gaston Berger University of Saint-Louis Senegal and a Telecommunications Thesis from the UCAD university in Dakar-Senegal, he is interested in he is interested in Applications of innovative telecom services to virtual organizations.