Android malware detection: Investigating the impact of imbalanced data-sets on the performance of machine learning models.

Zakaria SAWADOGO *†, Gervais MENDY *,Jean Marie DEMBELE†, Samuel OUYA *.
*LITA(Laboratoire d'informatique, de télécommunications et Applications), Université Cheikh Anta Diop de Dakar, Sénégal
†LANI (Laboratoire d'Analyse Numérique et Informatique), Université Gaston Berger, Sénégal

***Abstract*—Artificial intelligence has revolutionized many areas of research, including research on malicious application detection and classification. Nowadays, there are many approaches that learn from existing data and predict the classes of new data. Machine learning principles recommend a balance of classes in the training dataset, but the reality in the field is quite different. The majority of datasets used for malicious application detection are imbalanced. Class imbalance degrades classifier performance, so it is a common problem in classification tasks. This observation is much more significant in the area of Android malware detection and classification. There are few works to our knowledge on the effects of imbalanced datasets in the field of Android malware detection. Our contribution focuses on the impact of imbalanced datasets on the performance of different algorithms and the suitability of using evaluation metrics in Android malware detection. We show that for malicious application detection, some classification algorithms (KNN, AdaBoost, SVM, Naive Bayes, LogisticRegression) are not suitable for unbalanced datasets . We also proved that some of the most used performance evaluation measures (Accuracy, Precision, Recall) are not very well adapted to unbalanced datasets. On the other hand, the metrics (Balanced_accuracy, Geometric mean) are more adapted. These results were obtained by evaluating the performances of eleven classification algorithms (KNN, ExtraTrees Classifier, Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier, GradientBoostingClassifier, Hist Gradient Boosting, SVM, Naive Bayes, Logistic Regression, Ridge Classifier) and also the adequacy of the different evaluation metrics (Accuracy, Recall, Precision, F1_score, Balanced accuracy, Matthews corrcoef, Geometric mean, Fowlkes_mallows).**

***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.