PROCEEDINGS of the 2020 INTERNATIONAL CONFERENCE on INFORMATION TECHNOLOGY IN EDUCATION AND DEVELOPMENT (ITED)

Comparative Analysis of Machine Learning Classifiers for Detecting Malware in Portable Executables

by

Faden David Nanven and Morufu Olalere

Abstract

Over the years, malware vendors have evolved from using non intelligent malwares which are easily identifiable to intelligent malwares by employing polymorphism and metamorphism in malware behavior paving the way for evasive malware techniques ranging from environmental awareness, confusing automated tools, timing based evasion and obfuscated internal data. Modern malware detection techniques use machine learning algorithms mostly classifiers to detect malware signatures and malware behavior. Some of the machines learning algorithms are not effective in detecting malware behavior whereas some are. This research is a comparative analysis of commonly used machine learning classifiers ranging from Decision Tree, Random Forest and Bayesian Network. The training dataset comprises of 138,047 Portable Executable (PE) header file record samples which was divided into: 41,323 clean files containing exe and dll file samples and 96,724 malware file samples. This research evaluates the above-mentioned machine learning classifiers in relation to their performance using Machine Learning performance Metrics. Based on th experimental results of selected classifiers, the overall best performance was achieved by Gradient Boosting Classifier with accuracy of 98.5% and a Recall of 0.98 followed by both Random Forest and Decision Tree classifiers respectively. Index Terms— Malware, Classifiers, Obfuscation, Polymorphism, Metamorphism.

https://www.academiainformationtechnology.org/event/archive_uploads/archive_proceedings/ITED%202020.pdf