PE Header-Based Multinomial Malware Classification Using ML for Computing Systems

Md Raihan Subhan[§], Rubina Akter[†], Dong-Seong Kim[§], and Taesoo Jun[§]
[§]*Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi 39177, Korea*
[†]*ICT Convergence Research Center, Kumoh National Institute of Technology, Gumi 39177, South Korea*
(raihan*, rubinaakter2836, dskim, taesoo.jun)@kumoh.ac.kr

Abstract—**Malware is an envelope of various malicious software or programs which hackers use to harm and attack computer systems. Malware detection is the procedure of identifying the computer vulnerability and tracing the malicious files to discover the malware and, thus, secure the computing system. However, detecting and classifying the malware manually is almost impossible and time-consuming. This paper introduces a Portable Executable (PE) header-based approach for detecting malicious programs to handle the issue. Various Machine Learning (ML) algorithms are applied and compared to the performance to detect malware types.**

Index Terms—**Computing systems, malware detection, machine learning**

I. Introduction

The use of computers is increasing daily because of their fantastic service and performance. People are using computers to do their work [1]. But malware, short-term for malicious software, is a blanket term for viruses, worms, trojans, and other harmful computer programs hackers use to wreak destruction and access sensitive information. It is designed to cause extensive damage to data and systems or gain unauthorized access to a network. Malware is typically delivered as a link or file over email and requires the user to click on the link or open the file to execute the malware. So, malware detection is crucial to protect computers from these harmful programs. Various detection and classification techniques have been used to detect and classify malware. These techniques are responsible for determining the types of files and parsing them. Parsing various formats through an antivirus scanner is a highly complex task. On the other hand, machine learning methods for malware detection have proven effective against new malware. Machine learning is an artificial intelligence (AI) application that allows systems to learn and improve from experience without being explicitly programmed automatically [2]. It uses statistical methods to enable machines to improve with experience. A new and faster approach to differentiate between malware and legitimate *.exe* files by looking at the properties of the PE headers was introduced by [3]. The Portable Executable format is a file format for executables, object code, DLLs, and others used in 32-bit and 64-bit versions of Windows operating systems. The PE format is a data structure that encapsulates the information necessary for the Windows OS loader to manage the wrapped executable code. A PE file is a file that begins with a header containing information about the executable or the library (for example, a signature indicating that it is an executable, sizes, dates...), followed by the actual content (the executable or the library) that the system can read. The extensions that recognize that file format are .cpl, .dll, .drv, .efi, .exe, .ocx, .scr, and .sys [4]. This paper works on detecting malware and classifying the PE file header generated by the PE Parser header, which helps to protect our computing system.

II. Methodology of PE-Header-based malware detection

As technology is growing, the number of malware is also increasing. Malware is now designed with mutation characteristics that cause an enormous growth in malware variation. With these growths in new malware, traditional signature-based malware detection is proven ineffective against the wide variety of malware. In that scenario, Machine learning methods [5] for malware detection have a high false-positive rate for detecting malware. Therefore, we have applied several ML algorithms to aggregate the features and differentiate the characteristics of malware types. This paper used an open-source dataset for malware detection. The dataset consists of five types of malware: Locker, Mediyes, WinIbsec, Zbot, and Zeroaccess. Authors in [3] developed PE- Header-Parser using Python and the PEfile library. The author extracted the features from the file header, optional header, and section header, comparing the differences between malware and benign files. This study also developed a parser that can detect malware based on five features: the initialized data size, unknown section name, DLL characteristics, the central image version, and Checksum.

In this paper, we regenerate the parser using feature extraction algorithms and extract the key features. After the retrieve the key features, we stored the dataset in Attribute-Relation File Format (.arff). For the availability of any feature, it shows the value **1**, which represents the available feature, and for non-availability, it shows **0**. The selection of crucial feature status is to help apply various machine learning algorithms using IKA tools. The details workflow of this study is shown in Fig. 1.

III. Dataset description and Result analysis

The collected open-source dataset consists of five types of malware, such as Locker, Mediyes, WinIbsec, Zbot, and Zeroaccess, including a total of 8970 malware samples. The

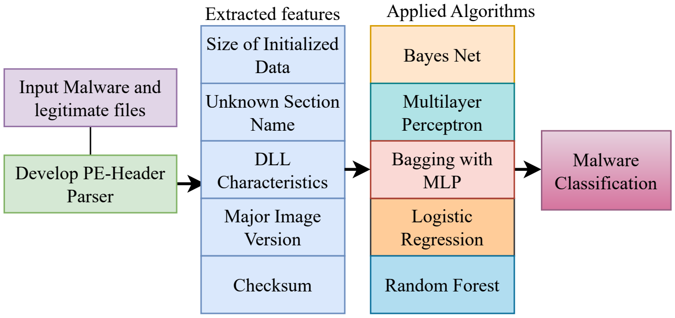

Fig. 1. Workflow for PE-Header-based malware detection Redraw the figure

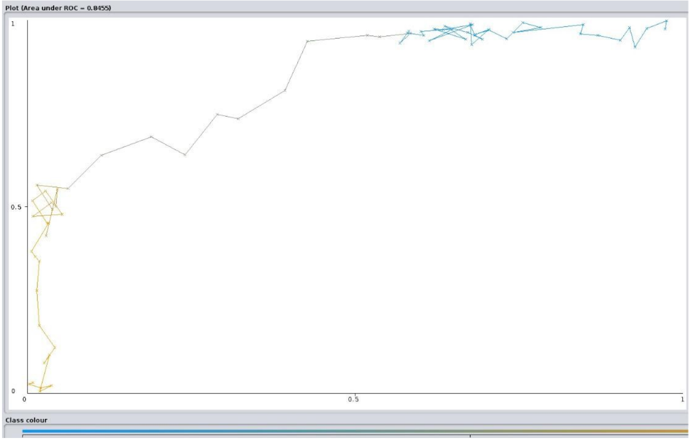

Fig. 2. The ROC curve after applying Random Forest (RF).

fundamental five-feature distribution of the dataset is shown in Table I.

After collecting the dataset, we develop the PE-header-parser to apply the ML algorithms for malware detection. During training, we divided the dataset for train and testing, such as 5382 samples for training, 1346 samples for the test, and 2691 samples for cross-validation holds. Importantly, we have applied BayesNet, Logistic Regression, Bagging with Multilayer Perceptron, Multilayer Perceptron, and Random Forest machine learning algorithm for malware classification. The classification results of the ML mentioned above algorithms are shown in Table I. BayesNet achieved 67.43% classification accuracy with the malware dataset, where the Recall and ROC area is 0.674 and 08.48. Applying Multilayer Perceptron as a parameter, the correctly classified instances, Recall, and ROC area are 66.59%, 0.666, and 0.850 accordingly. When we applied Logistic Regression as a parameter, we received 67.43% correctly classified instances Recall value 0.974, and the ROC area value 0.847, respectively. The ROC curve analysis of the RF is shown in Fig. 2. In addition, we also applied the Multilayer Perceptron model and received 68.09% classification accuracy, 0.681 Recall value, and 0.852 ROC area value. The analysis of the Random Forest algorithm provides 68.17% correctly classified instances, 0.682 Recall, 0.853 ROC area. This analysis shows that the Random Forest algorithm provides superior classification accuracy compared to the others algorithm.

TABLE I
PERFORMANCE COMPARISON OF MALWARE DETECTION USING ML ALGORITHMS

ML Algorithm	Classification accuracy, %	Recall	ROC Area
BayesNet	67.43	0.674	0.848
MP	66.59	0.666	0.850
LR	67.43	0.674	0.847
MLP	68.09	0.681	0.852
RF	68.17	0.682	0.853

IV. CONCLUSION

This paper focuses on machine learning-based malware classification techniques. The study collects an open-source dataset that includes five types of malware: Locker, Mediyes, WinIbsec, Zbot, and Zeroaccess. After collecting the dataset, we generated the PE header and decorated the features. Significantly, five malware features, such as Size of Initialized Data, Unknown Section Name, DLL Characteristics, Major Image Version, and Checksum, hold 3.2%, 0.0%, 85.77%, 93.70%, and 28.18% instances, respectfully. As a consequence, we applied BayesNet, Multilayer Perceptron, Adaboost, Logistic Regression, and Random Forest ML algorithms, where the Random Forest algorithm shows better results compared to the others model.

ACKNOWLEDGMENT

This work was supported by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science & Technology (MEST) (2018R1A6A1A03024003), MSIT (Ministry of Science & ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2023-2020-0-01612) supervised by the IITP (Institute for Information and communications Technology Planning and Evaluation), and the project for Industry-University-Research Institute platform cooperation R&D funded Korea Ministry of SMEs and Startups in 2022, (S3311338).

REFERENCES

[1] A. Zainudin, R. Akter, D.-S. Kim, and J.-M. Lee, "FedDDoS: An Efficient Federated Learning-based DDoS Attacks Classification in SDN-Enabled IIoT Networks," in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2022, pp. 1279–1283.

[2] R. Akter, M. Golam, V.-S. Doan, J.-M. Lee, and D.-S. Kim, "IoMT-Net: Blockchain integrated unauthorized UAV localization using lightweight convolution neural network for internet of military things," *IEEE Internet of Things Journal*, 2022.

[3] Y. Liao, "Pe-header-based malware study and detection," *Retrieved from the University of Georgia: http://www. cs. uga. edu/˜ liao/PE_Final_Report. pdf*, 2012.

[4] N. Balram, G. Hsieh, and C. McFall, "Static malware analysis using machine learning algorithms on apt1 dataset with string and pe header features," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019, pp. 90–95.

[5] A. Zainudin, L. A. C. Ahakonye, R. Akter, D.-S. Kim, and J.-M. Lee, "An efficient hybrid-DNN for DDoS detection and classification in software-defined IIoT networks," *IEEE Internet of Things Journal*, 2022.