Acta Scientific COMPUTER SCIENCES

Volume 4 Issue 12 December 2022

Research Article

Research Methodology on A Machine Learning Framework and Algorithms for Automatic Detection of Malware

M Atheequllah Khan[1], Imtiyaz Khan[2], Pankaj Kawadkar[3], M Upendra Kumar[4]* and D Shravani[5]

[1,2]*Assistant Professor, CS and AI Department, MJCET, OU, India*

[3]*Associate professor and HOD (CSE/IT/MCA), SSSUTMS Sehore, Madhya Pradesh, India*

[4]*Professor and Associate, Head CS and AI Department, MJCET, OU, India*

[5]*Associate Professor, ADCE Stanley College of Engineering and Technology for Women, OU, India*

***Corresponding Author:** M Upendra Kumar, Professor and Associate, Head CS and AI Dept, MJCET, OU, India.

Received: November 20, 2022

Published: November 29, 2022

Abstract

Cyberspace is ever expanding with inclusion of diversified networks and systems. With the emerging technologies such as Internet of Things (IoT) and distributed computing, there is seamless integration of heterogeneous applications with interoperability. This has brought unprecedented use cases and applications in various domains. Unfortunately, there is every growing threat to cyberspace due to different kinds of malicious programs termed as malware. Since adversaries are developing various kinds of malware, its detection has become a challenging task. Of late, machine learning (ML) techniques are widely used to solve problems in real world applications. Plenty of supervised learning methods came into existence. The objective of this paper is to explore and evaluate different ML models with empirical study. In this paper, we proposed a ML framework for analysing performance of different prediction models. An algorithm known as Machine Learning based Automatic Malware Detection (ML-AMD) is proposed. This algorithm is used to realize the framework with supervised learning. This empirical study has resulted in knowledge about ML models such as Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Multilayer Perceptron (MLP) and Gradient Boosting (GB). Random Forest model has exhibited highest accuracy with 97.96%. The research outcomes in this paper help in triggering further investigations towards automatic detection of malware.

Keywords: Malware Detection; Machine Learning; Decision Tree; Logistic Regression; Random Forest; Multilayer Perceptron and Gradient Boosting

Introduction

Malware is the malicious software that is created with bad intentions. It is often used to spread unwanted software that causes damage to systems. Adversaries are making business out of it and there are many incidents of it in the recent past. The term malware refers to software that damages devices, steals data, and causes chaos. There are many types of malware — viruses, Trojans, spyware, ransomware, and more. With the availability of malware instances and signatures, it became easier to identify known malware. Machine learning and Deep learning domains provide required AI enabled techniques that can be exploited for malware detection. With the recent advancements in ML and deep learning, it is possible to achieve near real time detection of malware. From the literature, it understood that ML and deep learning models are widely used for malware detection. Literature also provided certain feature selection approaches in order to improve quality of training data for supervised learning. However, there is need for improving feature selection to know the features that contribute to class label

prediction. From the literature it is also understood that ensemble approach could improve performance. Another insight is that deep learning models need to be used with advanced configurations for better performance. Therefore, the aim of the research is to propose a machine learning framework and algorithms for automatic detection of malware.

Literature Review

This section review literature on existing methods for detection of malware. Gibert., *et al*. [1] proposed a dep learning model for malware classification. It is made up of multiple models for efficient predictions. Li., *et al*. [2] proposed a malware detection model based on Domain Generation Algorithm (DGA). It is based on machine learning techniques. Pei., *et al*. [3] proposed a deep learning framework known as AMalNet based on CNN for malware detection. Karbab., *et al*. [4] focused on Android malware detection by defining an automated framework using deep learning methods. Karbab., *et al*. [5] proposed a data-driven malware detection approach using ML techniques. They used behaviour analysis reports for their empirical study. Wu [6] focused on a systematic study of malware detection methods based on deep learning. Jangam [7] explored deep learning, stacking and transfer learning methods for prediction purposes. Mahindru., *et al*. [8] proposed a methodology for automatic Android malware detection using ML techniques. Hosseinzadeh., *et al*. [9] proposed ML approaches that can be used for prediction of given disease. Chin., *et al*. [10] also focused on DGA based machine learning models for malware detection. Chen., *et al*. [11] used malware detection approach for Android malware using ML techniques. Masum., *et al*. proposed a deep learning model for Android malware detection. The model is known as Droid-NNet.

Xiao., *et al*. [13] defined a model based on deep learning behaviour graphs for malware detection. Usman., *et al*. [14] focused on building an intelligent system for malware detection and that is associated with digital forensics. Singh., *et al*. [15] used ML techniques to detect malware in executable files. Zhang., *et al*. [16] focused on feature exploration using deep learning towards classification of Android malware. Alzaylaee., *et al*. [17] proposed a deep learning framework known as DL-Droid for Android malware detection for real devices. Akarsh., *et al*. [18] used deep learning and visualized the detection of malware and the classification results. Dib., *et al*. [19] proposed multi-dimensional deep learning framework for malware classification in IoT environment. Kim., *et al*. [20] focused on extraction of features along with multi-modal deep learning in order to achieve Android malware detection performance. Pektaş., *et al*. [21] used opcode sequences and deep learning to detect Android malware. Gohari., *et al*. [22] used network traffic based deep learning for Android malware detection and classification.

Problem definition

Review of literature has revealed many significant gaps in the area of malware detection using machine learning. Three important gaps are considered for further investigation in this research.

- As investigated by Chandrasekhar and Sahin [23] feature selection methods are important for improving quality of training and leverage performance of detection models. Filter methods use ranking criterion that is suitable for data driven approaches. They discussed about both filter and wrapper methods. However, with respect to filter methods, there is specific research gap identified. It is understood from the review of the techniques that combining two or more filter methods has potential to improve the feature selection process and lead to enhancing prediction performance. Velswamy., *et al*. [24] also investigated on feature selection and its importance. They found that most of the existing approaches still suffer from stagnation as the search process is limited. It is advised to make a more robust approach by combining two or more approaches for improving feature selection process. As presented in [16,20], there is need for feature engineering for further improvement.
- Since ensemble approach exploits different efficient prediction models for improving prediction performance, it is to be considered for further research.
- As presented in [16,21] deep learning models provide performance benefits to malware detection systems. It is ascertained that there is need for improving deep learning models with pipelining in order to have better prediction performance.

Aim and Objectives

The aim of the research is to propose a machine learning framework and algorithms for automatic detection of malware. The research objectives are as follows.

- To propose a hybrid feature selection algorithm to leverage ML models for efficient detection of malware.
- To propose an ensemble algorithm that exploits multiple ML models and the hybrid feature selection algorithm to improve performance further.
- To propose an algorithm based on deep learning models for exploiting the advancements in ML for even better performance.

Methodology

The proposed methodology has both has both ML and deep learning approaches for malware detection. Malware detection in near real time provides benefits such as efficient handling of malware and improve cyber security.

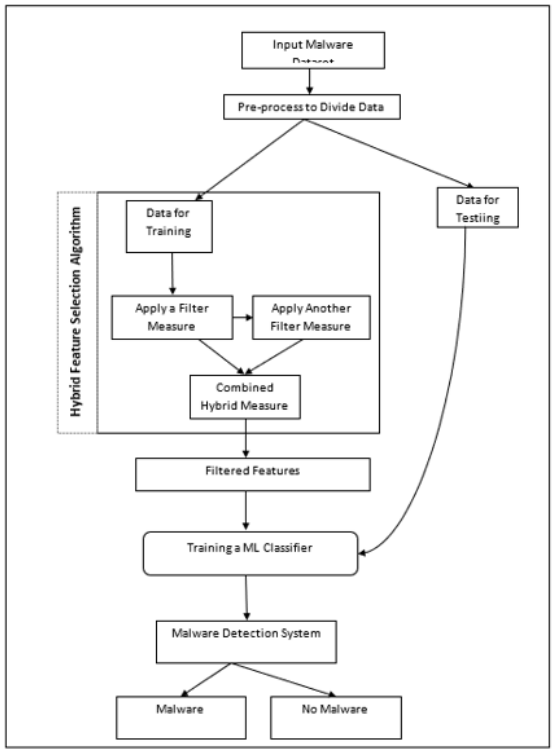

Figure 1: ML based framework for detection of malware.

As presented in figure 1, the proposal of a hybrid feature selection method ensures that the features selected are able to contribute in prediction of class labels. This supervised learning approach is widely used and suitable for the efficient malware prediction. The novelty in our approach is the proposal of a hybrid feature selection method that uses multiple filter based methods for feature engineering.

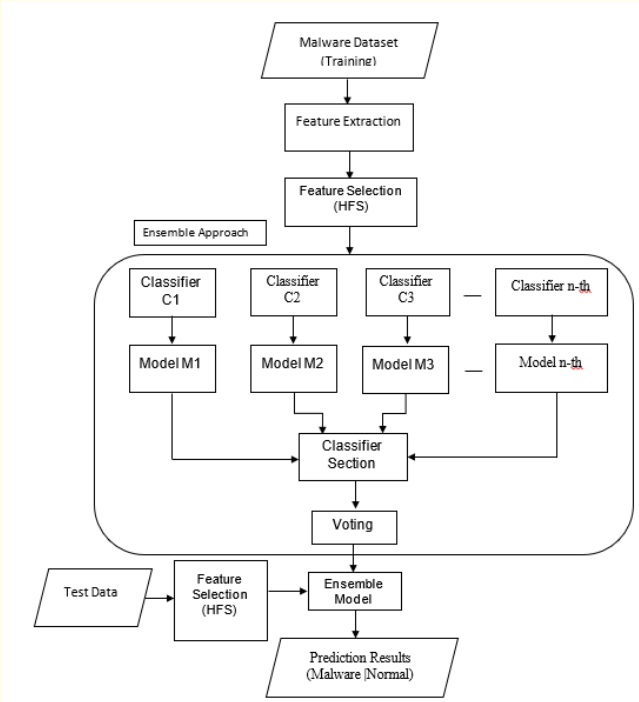

Figure 2: Ensemble model for malware detection.

As presented in figure 2, ensemble model is illustrated. Since ensemble approach exploits different efficient prediction models for improving prediction performance, it is to be considered for further research.

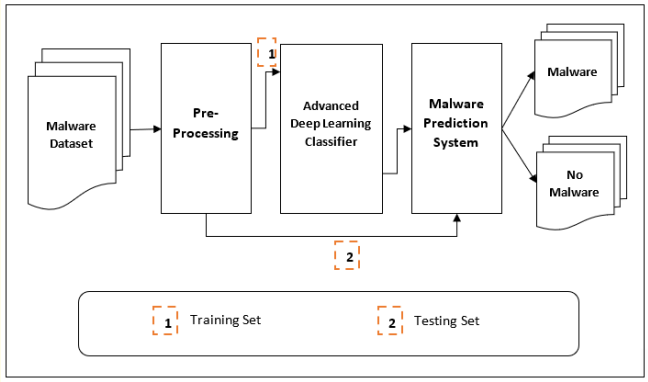

Figure 3: Deep learning based approach to detect malware.

As presented in figure 3, advanced deep learning algorithm is defined to analyse given test samples and predict malware. This methodology is further improved with a hybrid deep learning models in order to have better performance.

Dataset details

Drebin dataset is used for the research.

Dataset URL: https://www.impactcybertrust.org/dataset_view?idDataset=1372

Evaluation procedure

Based on confusion matrix, the evaluation of the proposed algorithm is compared with the state of the art. Table 1 shows different metrics used in the evaluation process.

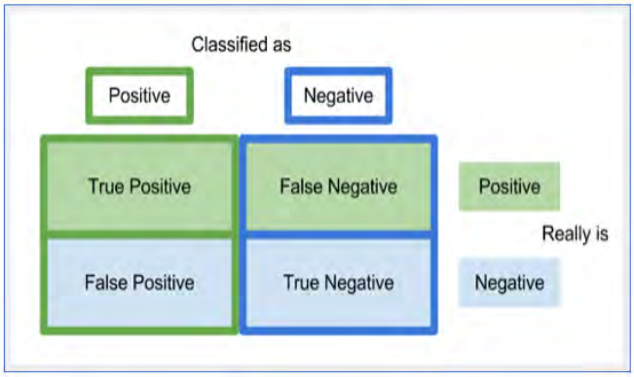

Figure 4: Confusion matrix.

Based on the confusion matrix presented in figure 4, the confusion matrix shows the measures like true positive (TP), false positive (FP), false negative (FN) and true negative (TN). These are determined by comparing result of ML algorithm when compared with the ground truth.

Metric	Formula	Value range	Best Value
Precision (p)		[0; 1]	1
Recall (r)		[0; 1]	1
Accuracy		[0; 1]	1
F1-Score		[0; 1]	1

Table 1: Performance metrics used for evaluation.

Precision refers to positive predictive value while the recall refers to true positive rate. F1-score is the harmonic mean of both precision and recall which is used to have a measure without showing imbalance while accuracy measure may show imbalance.

Conclusion and Future Work

In this paper, we proposed a ML framework for analysing performance of different prediction models. This empirical study has resulted in knowledge about ML models such as Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Multilayer Perceptron (MLP) and Gradient Boosting (GB). An algorithm known as Machine Learning based Automatic Malware Detection (ML-AMD) is proposed. This algorithm is used to realize the framework with supervised learning. It exploits a pipeline of aforementioned ML models to evaluate their performance in malware detection. Out experimental results revealed the utility of various ML models. Performance of different models are evaluated in terms of precision, recall, F1-Score and accuracy. Random Forest model has exhibited highest accuracy with 97.96%. The research outcomes in this paper help in triggering further investigations towards automatic detection of malware. In future, we explore deep learning models for malware detection as they have the capacity to have in-depth learning of features from data leading to improved prediction performance.

Bibliography

1. Gibert Daniel., *et al*. "HYDRA: A multimodal deep learning framework for malware classification". *Computers and Security* 95 (2020): 1-47.

2. Li Yi., *et al*. "A Machine Learning Framework for Domain Generation Algorithm (DGA)-Based Malware Detection". *IEEE Access* (2019): 1-18.

3. Pei Xinjun., *et al*. "AMalNet: A deep learning framework based on graph convolutional networks for malware detection". *Computers and Security* 93 (2020): 1-21.

4. Karbab ElMouatez Billah., *et al*. "MalDozer: Automatic framework for android malware detection using deep learning". *Digital Investigation* 24 (2018): S48-S59.

5. Karbab ElMouatez Billah and Debbabi Mourad. "MalDy: Portable, data-driven malware detection using natural language processing and machine learning techniques on behavioral analysis reports". *Digital Investigation* 28 (2019): S77-S87.

6. Huanyu Wu. "A Systematical Study for Deep Learning Based Android Malware Detection". Proceedings of the 2020 9th International Conference on Software and Computer Applications (2020): 1-6.

7. Ebenezer Jangam., *et al*. "Automatic detection of COVID-19 from chest CT scan and chest X-Rays images using deep learning, transfer learning and stacking". *Applied Intelligence* (2021): 1-17.

8. Mahindru Arvind and Sangal AL. "MLDroidâ□□framework for Android malware detection using machine learning techniques". *Neural Computing and Applications* (2020): 1-58.

9. Sara Hosseinzadeh Kassania., *et al*. "Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning Based Approach". *Biocybernetics and Biomedical Engineering* (2021): 1-13.

10. Chin T., *et al*. "A Machine Learning Framework for Studying Domain Generation Algorithm (DGA)-Based Malware". *Security and Privacy in Communication Networks* (2018): 433-448.

11. Chen Xiao., *et al*. "Android HIV: A Study of Repackaging Malware for Evading Machine-Learning Detection". *IEEE Transactions on Information Forensics and Security* (2019): 1-15.

12. Masum Mohammad and Shahriar Hossain. "IEEE 2019 IEEE International Conference on Big Data (Big Data) - Los Angeles, CA, USA (2019.12.9-2019.12.12)". 2019 IEEE International Conference on Big Data (Big Data) - Droid-NNet: Deep Learning Neural Network for Android Malware Detection (2019): 5789-5793.

13. Xiao Fei., *et al*. "Malware Detection Based on Deep Learning of Behavior Graphs". *Mathematical Problems in Engineering* (2019): 1-10.

14. Nighat Usman., *et al*. "Intelligent Dynamic Malware Detection using Machine Learning in IP Reputation for Forensics Data Analytics". *Future Generation Computer Systems* (2021): 1-18.

15. Singh Jagsir., *et al*. "A survey on machine learning-based malware detection in executable files". *Journal of Systems Architecture* (2020): 1-24.

16. Nan Zhang., *et al*. "Deep learning feature exploration for Android malware detection". Ap*plied Soft Computing* (2021): 1-7.

17. Alzaylaee Mohammed K., *et al*. "DL-Droid: Deep Learning Based Android Malware Detection Using Real Devices". *Computers and Security* (2019): 1-28.

18. S Akarsh., *et al*. "IEEE 2019 5th International Conference on Advanced Computing and Communication Systems (ICACCS) - Coimbatore, India (2019.3.15-2019.3.16)". 2019 5th International Conference on Advanced Computing and Communication Systems (ICACCS) - Deep Learning Framework and Visualization for Malware Classification (2019): 1059-1063.

19. Mirabelle Dib., *et al*. "A Multi-Dimensional Deep Learning Framework for IoT Malware Classification and Family Attribution". *IEEE Transactions on Network and Service Management* (2021): 1-12.

20. Kim Tae Guen., *et al*. "A Multimodal Deep Learning Method for Android Malware Detection using Various Features". *IEEE Transactions on Information Forensics and Security* (2018): 1-16.

21. Pektaş Abdurrahman and Acarman Tankut. "Deep Learning To Detect Android Malware via Opcode Sequences". *Neurocomputing* (2019): 1-21.

22. Mahshid Gohari., *et al*. "Android Malware Detection and Classification Based on Network Traffic Using Deep Learning". 2021 7th International Conference on Web Research (ICWR) (2021): 1-7.

23. Chandrashekar G and Sahin F. "A survey on feature selection methods". *Computers & Electrical Engineering* 40.1 (2014): 16-28.

24. Velswamy Karunakaran., *et al*. "Exploring a Filter and Wrapper Feature Selection Techniques in Machine Learning" (2021).