Review on Machine Learning Based Malware Detection

[1] Lubna Javaid, [2] Sudesh Kumar
[1]Student, [2]Assistant Professor
[1]SoCSE, SMVDU, Katra,
[2]SoCSE, SMVDU, Katra,
[1] 21mms004@smvdu.ac.in, [2] Sudesh.bhadu@smvdu.ac.in

ABSTRACT

Malware detection using machine learning has gained significant attention in recent years due to the increasing number of malware attacks. With the increasing use of mobile devices, the need for effective malware detection techniques has become even more critical. Machine learning has emerged as a promising approach for detecting malware, as it can learn to identify patterns in large datasets and classify them as either benign or malicious. Previous research in this area has mainly focused on the detection of Android malware using static and dynamic analysis techniques. This review paper examines the efficiency of machine learning for malware identification, with a focus on the latest research in the field. The paper presents an analysis of the various machine learning algorithms used for identification of malware, their strengths and limitations, and the evaluation metrics used for measuring the performance of these methods. Overall, this review paper provides insights into the novelty in machine learning-based malware identification and highlights the need for further research in this field to build more potent and effective techniques for detecting unknown or zero-day attacks.
Keywords: Malware Detection, Machine Learning, Benign, Malicious Files

INTRODUCTION

Malware has become an increasingly prevalent threat to computer security, with new and sophisticated attacks being developed on a regular basis. Traditional signature-based detection methods are often inadequate in identifying new and unknown types of malware, making it necessary to develop more advanced approaches to detect these threats. Machine learning has emerged as a promising technique for detecting malware, leveraging the ability of algorithms to learn from large datasets and identify patterns in the data. The emergence of machine learning techniques for identification of malicious software has the potential to significantly improve our ability to detect and respond to malware threats. By leveraging the power of machine learning algorithms, we can develop more robust and effective methods for detecting known and unknown types of malware, ultimately helping to enhance the security and privacy of computer systems. Previous research in this area has focused primarily on detecting Android malware using static and dynamic analysis techniques [1], [2]. However, the limitations of these techniques in detecting unknown or zero-day attacks have driven the development of more advanced machine learning-based approaches.

This review paper aims to provide a inclusive overview of the current research in the area of malware detection based on machine learning, with a particular emphasis on the challenges and opportunities presented by the use of this approach. We will discuss various machine learning algorithms utilised for identifying malware, including supervised, unsupervised, and reinforcement learning. We will also examine the challenges involved in building effective models, interpreting results, and evaluating performance. we hope to provide a enhanced support of the present-day art in machine learning-based malware detection and identify future research directions.

RELATED WORK

Malware detection using machine learning is an important research area in computer security. In this literature survey, we will review the current state-of-the-art techniques and approaches for detecting malware using machine learning.

Velasco-Mata *et al.*, 2021 [3] presented a model which included selection of key features using Information Gain and Gini importance to enhance the functioning of traffic classification of Botnet. Three models were evaluated with two datasets, and Decision Trees with a five-feature set demonstrated the best performance with 0.78 microseconds classification time and 85% F1-score.

Martins *et al.*, 2020 [4] presented a survey to inspect intrusion revealing scenarios with the use of adversarial machine learning. Several attacks were found to be effective in detecting adversarial examples, but their application in intrusion scenarios requires further testing. Adversarial defenses have been less explored but have shown effectiveness in resisting adversarial attacks.

Odat and Yaseen, 2023 [5] projected a model which for purely for detecting malware of Android utilising machine learning which uses concurrence of static attributes. The model was tested using a new dataset of co-existed features at different levels, extracted from the Malgenome, Drebin and MalDroid2020 datasets. The proposed model outperformed the contemporary model, achieving a high accuracy of up to 98% using several conventional machine learning algorithms.

Shaukat *et al.*, 2020 [6] provided an inclusive summary of the downside in using machine learning techniques to protect information space from outbreaks, including detecting spam/ham, intrusion detection, and identifying malware. While ML has been increasingly used in cyber security applications, malicious adversaries can exploit the vulnerabilities of ML systems, making it challenging to ensure their trustworthiness.

Vinayakumar *et al.*, 2019 [7] proposed a unique procedure for deep learning and Machine learning architectures for achieving an unbiased and efficient intrusion detection, outperforming classical MLAs and also highlighted the challenges of detecting unknown malware in real-time and how machine learning algorithms (MLAs) can be used for malware analysis.

Wan *et al.*, 2020 [8] proposed a method to classify and identify IoT malicious programs by utilizing machine learning approach while surveying fundamental favoring data kept in byte sequences. The proposed method for malware family classification achieved 98.47% of accuracy over a large

balanced data which included 111000 malicious files and 111000 Benign files.

Yang *et al.*, 2019 [9] proposed two unique approaches to address the malware detection issue: malware classification using ensemble models and malware family clustering using t-SNE algorithm for visualization. Real-life malware samples were used to test the proposed model, and got significant improvement over existing methods as a result. These methods offer higher accuracy and which can be expended for family clustering as well as malicious code classification.

Mahindru and Sangal., 2021 [10] focused on developing a model for malware detection using permission and API calls as features. The study conducted experiments on 500,000 distinct Android apps belonging to 30 different categories and implemented ten different feature selection approaches and five unsupervised machine learning algorithms to develop distinct models. The experiments showed that using machine learning algorithm (farthest first) along with feature selection shows promising results by achieving 98.8% of detection rate.

Machine learning is a promising approach to detecting and classifying malware. Supervised learning, unsupervised learning, and semi-supervised learning have all been used to develop effective malware detection systems. However, there are still challenges to overcome, such as the ability of malware to evade detection and the need for more diverse and realistic datasets for evaluating machine learning algorithms.

Author	**Dataset**	**Performance Evaluation**	**Limitations**
[3] Velasco-Mata *et al.*, 2021	Created their own dataset EQB-CTU13 and QB-CTU13 based on dataset CTU-13	F1 score - 85%	When data packets are sent in large numbers it becomes difficult to detect in real-time.
[4]Martins *et al.*, 2020	_____	_____	Datasets used are outdated hence standarised datasets should be used.
[5]Odat and Yaseen, 2023	Drebin, CIC_MALDROID 2020 and Malgenome	Accuracy - 98%	Dynamic features are not used.
[6]Shaukat *et al.*, 2020	_____	_____	Benchmark and representative datasets are unavailable.

Table (1) – Summary Table of Literature Survey

[7] Vinayakumar *et al.*, 2019	Ember dataset, Malimg dataset and self-collected dataset.	Accuracy - 98.8%	Deep learning technique's robustness is not discussed which can lead to misclassification.
[8] Wan *et al.*, 2020	Generated their own dataset.	Accuracy - 98.47%	Generalisation performance needs to be improved.
[9] Yang *et al.*, 2019	Generated their own dataset of 30000 samples.	Accuracy : 98.3%	Classification performance was not up to the mark.
[10]Mahindru and Sangal., 2021	Generated their own dataset by collecting 500000 android apps.	Accuracy : 98.8%	———

GENERAL APPROACH FOR MALWARE DETECTION USING MACHINE LEARNING

In this section, general approach for malware detection based on machine learning (ML) methods will be discussed.

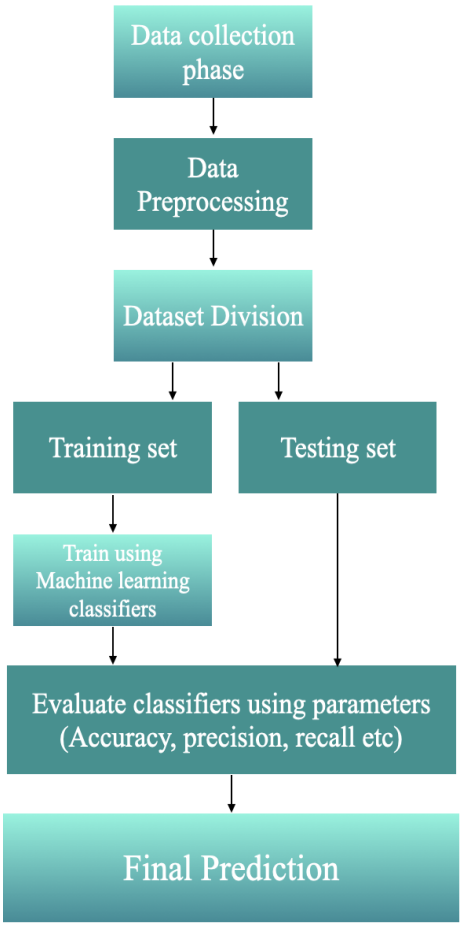

Fig.1 : Generic approach for machine learning(ML) based malware detection

The steps involved in machine learning can be defined as follows:

A. **Data Collection:** The first step is to collect the data that is required to train the model. The data can be obtained from various sources like public repositories, web scraping, and through surveys. In malware detection most of the researchers create their own dataset.
B. **Data Pre-processing:** After collecting the data, next step is to convert the format of dataset for machine learning algorithms to use it. This includes tasks like cleaning, handling missing values, getting rid of duplicate data, and transforming categorical data to numerical data.
C. **Dataset Division:** Once the data has been pre-processed, it is divided in three splits: with validation set for examining the model while training the data, training set to get model trained, and testing set for evaluating, after the model is trained
D. **Model Selection:** The next step is to select a suitable model that can be trained on the data. There are many models available in machine learning such as logistic regression, Random Forest, linear regression, and neural networks. Choosing model is based on problem type which is being solved, size of the dataset, and the desired accuracy.

E. Model Evaluation: Finally, To check model's functionality, evaluation over validation set is done. It includes computing various parameters like precision, AUC, accuracy, and F1-score. If the model is not performing well, it may be necessary to adjust the model parameters or try a different model.

DISCUSSION

The literature survey on malware detection using machine learning revealed that there has been a significant amount of study under this area in recent years. Various methods have been proposed in detecting malware, including supervised and unsupervised learning methods. One common approach is to use supervised learning, like Decision Tree, Support Vector Machine (SVM), etc. for classifiers to get trained on large datasets of known malware and benign samples. These classifiers have shown high accuracy rates in detecting malware, often outperforming traditional signature-based approaches. However, the effectiveness of these classifiers can be limited by the quality of the training dataset and the ability of malware to evade detection. Unsupervised learning methods, such as clustering and anomaly detection, have also been explored for malware detection. These methods are useful when there is limited or no prior knowledge about the malware being analysed. However, the high false positive rate and the inability to identify specific malware families are significant challenges for these approaches. Several studies have also focused on improving the performance of machine learning-based malware detection by combining multiple detection methods, including both signature-based and behaviour-based techniques[1], [2]. This approach has shown promising results, as it can overcome the limitations of individual methods and provide more accurate and comprehensive malware detection. Figure 2 shows the most commonly used machine learning classifiers[11]

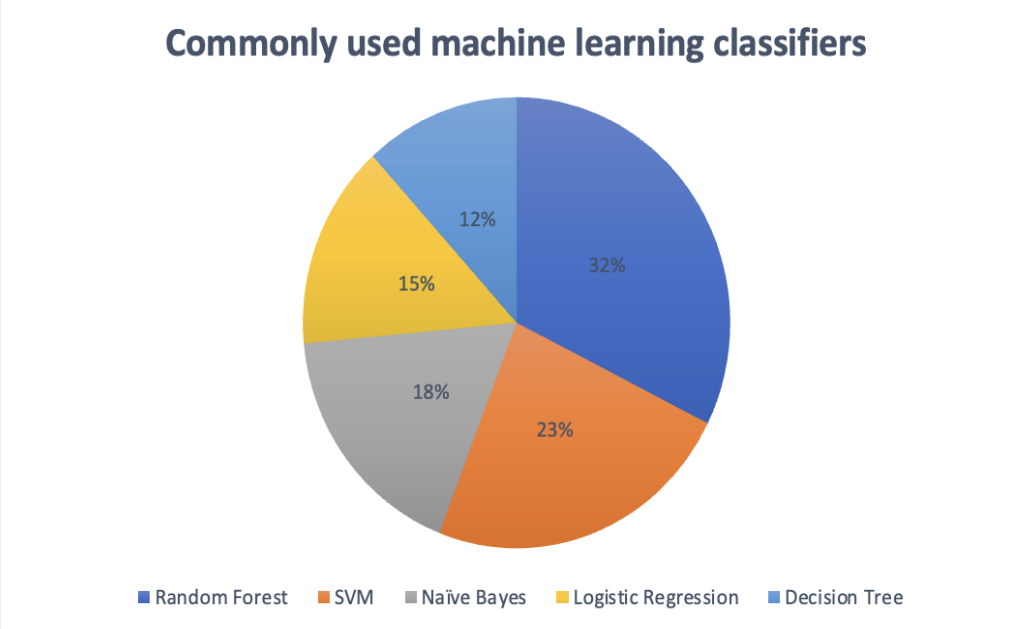

Fig. 2 : Most commonly used machine learning classifiers

CONCLUSION

In order to ensure confidentiality of computer systems and networks, it makes identifying malware a critical task the security of computer systems and networks. However, Literature survey shows that machine learning techniques can be a powerful tool for detecting malware. However, selecting the most relevant features for detecting malware is a challenging task. This is where feature selection techniques come into play, which can efficiently reduce the feature space and increases the accuracy rate of malware detection. Future work regarding this area should focus on developing new feature selection techniques that can handle the complexity of modern malware. The researchers should explore the use of deep learning based approaches for malware detection, which have provided the magnificent results for other domains such as natural language processing. Additionally, the creation of new datasets that include diverse types of malware can help to train more effective machine learning models. Overall, detecting malicious programs with machine learning is rapidly evolving, yet there is still much to be explored and discovered.

REFERENCES

[1]D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon, and K. Rieck, 'DREBIN: Effective and Explainable Detection of Android Malware in Your Pocket', 2014. [Online]. Available: http://dx.doi.org/doi-info-to-be-provided-later

[2]J. Li *et al.*, 'Networked human motion capture system based on quaternion navigation', in *BodyNets International Conference on Body Area Networks*, 2017. doi: 10.1145/0000000.0000000.

[3]J. Velasco-Mata, V. Gonzalez-Castro, E. F. Fernandez, and E. Alegre, 'Efficient Detection of Botnet Traffic by Features Selection and Decision Trees', *IEEE Access*, vol. 9, pp. 120567–120579, 2021, doi: 10.1109/ACCESS.2021.3108222.

[4]N. Martins, J. M. Cruz, T. Cruz, and P. Henriques Abreu, 'Adversarial Machine Learning Applied to Intrusion and Malware Scenarios: A Systematic Review', *IEEE Access*, vol. 8. Institute of Electrical and Electronics Engineers Inc., pp. 35403–35419, 2020. doi: 10.1109/ACCESS.2020.2974752.

[5]E. Odat and Q. M. Yaseen, 'A Novel Machine Learning Approach for Android Malware Detection Based on the Co-Existence of Features', *IEEE Access*, vol. 11, pp. 15471–15484, 2023, doi: 10.1109/ACCESS.2023.3244656.

[6]K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, and M. Xu, 'A Survey on Machine Learning Techniques for Cyber Security in the Last Decade', *IEEE Access*, vol. 8, pp. 222310–222354, 2020, doi: 10.1109/ACCESS.2020.3041951.

[7]R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, and S. Venkatraman, 'Robust Intelligent Malware Detection Using Deep Learning', *IEEE Access*, vol. 7, pp. 46717–46738, 2019, doi: 10.1109/ACCESS.2019.2906934.

[8]T.-L. Wan *et al.*, 'Efficient Detection and Classification of Internet-of-Things Malware Based on Byte Sequences from Executable Files', *IEEE Open Journal of the Computer Society*, vol. 1, pp. 262–275, Oct. 2020, doi: 10.1109/ojcs.2020.3033974.

[9]H. Yang, S. Li, X. Wu, H. Lu, and W. Han, 'A Novel Solutions for Malicious Code Detection and Family Clustering Based on Machine Learning', *IEEE Access*, vol. 7, pp. 148853–148860, 2019, doi: 10.1109/ACCESS.2019.2946482.

[10]A. Mahindru and A. L. Sangal, 'SemiDroid: a behavioral malware detector based on unsupervised machine learning techniques using feature selection approaches', *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 5, pp. 1369–1411, May 2021, doi: 10.1007/s13042-020-01238-9.

[11]V. Kouliaridis and G. Kambourakis, 'A comprehensive survey on machine learning techniques for android malware detection', *Information (Switzerland)*, vol. 12, no. 5, 2021, doi: 10.3390/info12050185.