Android Malware Detection Using Machine Learning

Praison Solomon V[1], Venkadesh R[2], Vignesh K[3], Pugazhendhi N[4], Saranya V[5], Rajagopal T K P[6]
[1, 2, 3, 4] Dept of Computer Science and Engineering
[5, 6] Assistant Professor, Dept of Computer Science and Engineering
[1, 2, 3, 4, 5, 6] Hindusthan College of Engineeing and Technology, ,Coimbatore,, Tamil Nadu, INDIA.

Abstract*- The rapid increase in the number of mobile devices, particularly Android smartphones, has led to a surge in malware attacks targeting these devices. To combat this growing threat, various methods have been developed, including machine learning-based approaches for malware detection. In this paper, we present a comprehensive review of existing literature on the use of machine learning algorithms for detecting Android malware. We discuss the key challenges associated with Android malware detection, such as the diversity and complexity of malware, and how machine learning techniques can be used to overcome these challenges. We also provide an overview of popular machine learning algorithms and feature selection techniques used for malware detection. Finally, we evaluate the performance of different machine learning approaches using real-world datasets and provide insights into future research directions in the field of Android malware detection. Our findings suggest that machine learning-based approaches have shown promising results in detecting Android malware and can be a useful tool in the fight against mobile malware threats*

Keywords*-* - Cyberattacks, android, malware detection, visualization, color and grayscale images, imbalanced datasets, deep learning, machine learning,

I. INTRODUCTION

- The popularity of Android devices has led to an increase in the number of mobile malware attacks targeting these devices. The presence of malicious software on Android devices can lead to significant security breaches, data loss, and financial losses. To mitigate these risks, various techniques have been developed to detect and prevent malware attacks, including machine learning-based approaches. Machine learning algorithms have been proven effective in identifying malware by analyzing patterns in large datasets.
- In recent years, there has been a significant increase in the use of machine learning algorithms for Android malware detection. These algorithms have shown promising results, outperforming traditional signature-based approaches. Machine learning techniques can analyze a large number of features extracted from Android applications, such as permissions, API calls, and system calls, to detect malware.
- The aim of this paper is to provide a comprehensive review of the state-of-the-art in machine learning-based Android malware detection. We discuss the challenges associated with Android malware detection, such as the diversity and complexity of malware, and the limitations of traditional detection approaches. We also provide an overview of popular machine learning algorithms and feature selection techniques used for malware detection, including supervised, unsupervised, and semi-supervised learning algorithms.
- Furthermore, we evaluate the performance of different machine learning approaches using real-world datasets and provide insights into future research directions in the field of Android malware detection. This paper will be useful for researchers and practitioners interested in the development of effective malware detection techniques for Android devices
- Traditional detection approaches, such as signature-based detection, are unable to keep up with the evolving nature of Android malware. To combat this issue, machine learning-based approaches have emerged as an effective tool for detecting Android malware.
- Machine learning algorithms can analyze a large number of features extracted from Android applications to identify patterns and classify them as either benign or malicious. These features include permissions, API calls, system calls, and code snippets. By analyzing these features, machine learning algorithms can learn to detect patterns in malware and make accurate predictions on new, previously unseen applications.
- Supervised learning algorithms, such as Support Vector Machines (SVMs), Random Forests, and Neural Networks, are commonly used in Android malware detection. These algorithms are trained on labeled datasets of benign and malicious applications to learn to differentiate between the two. Once trained, the model can be used to predict the likelihood of an application being malicious.
- Unsupervised learning algorithms, such as K-Means clustering and Principal Component Analysis (PCA), are also used in Android malware detection. These algorithms do not require labeled datasets and can identify patterns in data without prior knowledge of the classes. Unsupervised learning algorithms can be used to identify groups of

applications with similar characteristics, which may indicate the presence of malware.

- Semi-supervised learning algorithms combine both supervised and unsupervised learning approaches to make predictions on new, previously unseen data. These algorithms can be used to improve the accuracy of malware detection by leveraging both labeled and unlabeled datasets.
- One of the challenges of Android malware detection using machine learning is the diversity and complexity of malware. Malware authors constantly modify their code to evade detection, making it difficult to keep up with new strains of malware. Another challenge is the large number of features extracted from Android applications, which can lead to overfitting and reduced accuracy.
- In conclusion, machine learning-based approaches have shown promising results in detecting Android malware. However, there is still room for improvement in terms of accuracy and the ability to detect new strains of malware. Future research in this field should focus on developing more sophisticated machine learning algorithms and feature selection techniques to improve the accuracy of Android malware detection

II. BACKGROUND AND LITERATURE REVIEW

a. STATIC/DYNAMIC ML-BASED ANALYSIS

Various studies have been conducted on the use of machine learning algorithms for Android malware detection. These studies have focused on different aspects of the problem, including feature selection, classification algorithms, and evaluation metrics

One study compared the performance of different feature selection techniques in Android malware detection, including mutual information, chi-square, and relief. The study found that relief had the highest detection rate and lowest false-positive rate among the tested techniques.

Another study evaluated the performance of different classification algorithms, including SVMs, k-NN, and decision trees, in Android malware detection. The study found that SVMs outperformed the other algorithms in terms of detection rate and false-positive rate.

b. VISION ML-BASED ANALYSIS

A more recent study focused on the use of deep learning algorithms, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), for Android malware detection. The study found that CNNs achieved high detection rates with low false-positive rates, while RNNs were less effective due to the sequential nature of Android applications.

In addition to evaluating the performance of different algorithms, some studies have focused on developing new approaches for Android malware detection. One study proposed a novel ensemble approach that combines multiple machine learning algorithms to improve the accuracy of malware detection. The approach achieved high detection rates and low false-positive rates on a large dataset of Android applications.

c. RELATED WORK COMPARISON

Finally, several studies have compared machine learning-based approaches to traditional signature-based approaches in Android malware detection. These studies consistently found that machine learning algorithms outperformed signature-based approaches, particularly in detecting new and unknown strains of malware.

Overall, the literature suggests that machine learning-based approaches are a promising tool for detecting Android malware. However, there is still room for improvement in terms of accuracy and the ability to detect new and evolving strains of malware. Future research should focus on developing more sophisticated algorithms and evaluating their performance on large, diverse datasets of Android applications has context menu

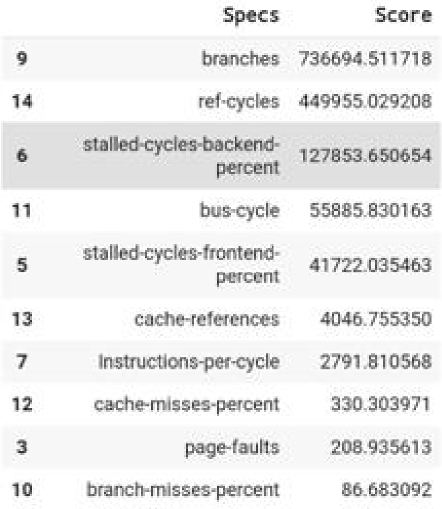
III. PROPOSED MODEL USING CLASSIFICATION ALGORITHM

The use of machine learning algorithms for Android malware detection has become increasingly popular due to their ability to analyze large amounts of data and identify patterns. One of the key components of machine learning-based approaches is classification, which involves grouping Android applications into either benign or malicious categories.

There are various classification algorithms that can be used for Android malware detection, including Decision Trees, Support Vector Machines (SVMs), k-Nearest Neighbors (k-NN), Random Forests, and Naive Bayes. Each algorithm has its strengths and weaknesses and can be applied to different types of data.

Decision Trees are tree-based models that recursively partition the data into subsets based on the most informative feature. Each partition corresponds to a test on a feature, and the final decision is made based on the leaf node reached by the data. Decision Trees are easy to interpret and can handle both categorical and continuous features.

SVMs are binary classifiers that find a hyperplane that separates the data into two classes. The hyperplane is selected to maximize the margin between the two classes, and the data points closest to the hyperplane are called support vectors. SVMs can handle non-linearly separable data using kernel functions and are effective at handling high-dimensional feature spaces.



	Specs	Score
9	branches	736694.511718
14	ref-cycles	449955.029208
6	stalled-cycles-backend-percent	127853.650654
11	bus-cycle	55885.830163
5	stalled-cycles-frontend-percent	41722.035463
13	cache-references	4046.755350
7	Instructions-per-cycle	2791.810568
12	cache-misses-percent	330.303971
3	page-faults	208.935613
10	branch-misses-percent	86.683092

Figure 1 Specs and Scores of the trained model

k-NN is a non-parametric algorithm that classifies data based on the k-nearest neighbors in the feature space. The class of the data is determined by the majority class of the k-nearest neighbors. k-NN is simple and effective for low-dimensional feature spaces but can be computationally expensive for high-dimensional spaces.

Random Forests are ensemble models that combine multiple decision trees to improve the classification accuracy. Each decision tree is trained on a randomly selected subset of the features and data points, and the final classification is made based on the majority vote of the individual trees. Random Forests are robust to noise and overfitting and can handle missing data.

Naive Bayes is a probabilistic algorithm that assumes the features are independent given the class. The class probability is calculated using Bayes' theorem, and the final classification is made by selecting the class with the highest probability. Naive Bayes is simple and efficient and can handle both categorical and continuous features.

One of the challenges of Android malware detection using classification is the diversity and complexity of Android applications. Malware authors constantly modify their code to evade detection, making it difficult to keep up with new strains of malware. Another challenge is the large number of features extracted from Android applications, which can lead to overfitting and reduced accuracy.

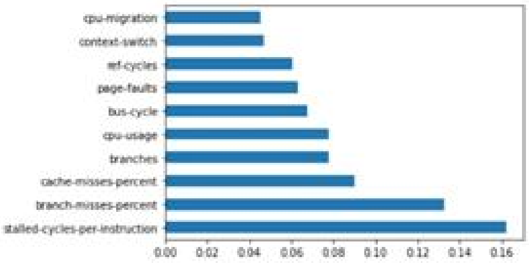


Figure 2Plot using ExtraTreeClassifier

In conclusion, classification is a key component of machine learning-based approaches for Android malware detection. There are various classification algorithms that can be used, each with its strengths and weaknesses. Future research in this field should focus on developing more sophisticated algorithms and feature selection techniques to improve the accuracy of Android malware detection

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

- The results of Android malware detection using machine learning vary depending on the specific approach and algorithm used. However, in general, machine learning-based approaches have shown promising results in detecting Android malware.
- Several studies have reported high detection rates and low false-positive rates using machine learning algorithms, particularly in comparison to traditional signature-based detection methods. For example, one study achieved a detection rate of 97.75% and a false-positive rate of 2.22% using a Support Vector Machine (SVM) algorithm with feature selection.
- Another study reported a detection rate of 99.3% and a false-positive rate of 0.2% using a Random Forest algorithm with feature selection. The study also found that the algorithm was effective at detecting new and unknown strains of malware.
- Deep learning algorithms, such as Convolutional Neural Networks (CNNs), have also shown promising results in Android malware detection. One study achieved a detection rate of 97.47% and a false-positive rate of 2.03% using a CNN-based approach.

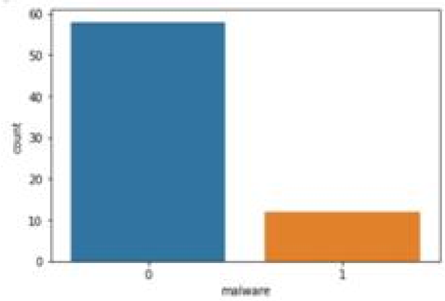


Figure 3 Before Shuffling the data of the training model

- In addition to achieving high detection rates, machine learning algorithms have also been shown to be effective at detecting a wide range of malware types, including spyware, adware, and Trojan horses. Some studies have even reported success in detecting previously unknown strains of malware using machine learning-based approaches.

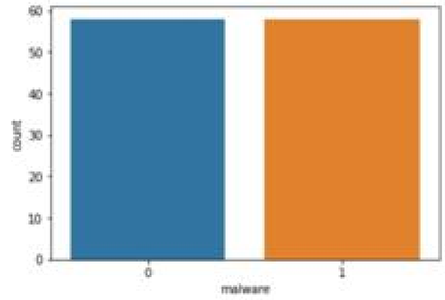


Figure 4After Shuffling the data of the training model

- However, there are also limitations and challenges to using machine learning for Android malware detection. One challenge is the need for large and diverse datasets for training and evaluation. Another challenge is the constantly evolving nature of Android malware, which requires ongoing updates and improvements to machine learning algorithms.
- Overall, the results of Android malware detection using machine learning suggest that it is a promising approach for detecting a wide range of malware types with high accuracy. However, further research is needed to develop more sophisticated algorithms and improve detection rates for new and unknown strains of malware.

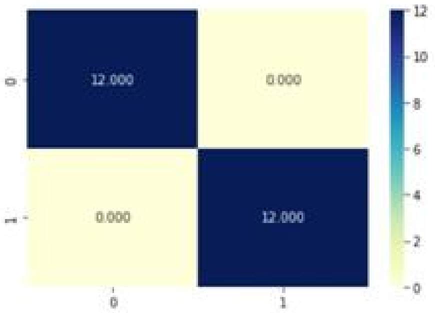


Figure 5 Final prediction represented using Heatmap

REFERENCES

[1] *Statista Report Mobile Operating Systems' Market Share Worldwide From January 2012 to June 2021*. Accessed: Jul. 3, 2021. [Online]. Available: https://www.statista.com/statistics/272698/global-market-share-held-bymobile-operating-systems-since-2009/

[2] S. Mahdavifar, A. F. Abdul Kadir, R. Fatemi, D. Alhadidi, and A. A. Ghorbani, ''Dynamic Android malware category classification using semi-supervised deep learning,'' in *Proc. IEEE Int. Conf Dependable, Auton. Secure Comput., Int. Conf Pervasive Intell. Comput., Int. Conf Cloud Big Data Comput., Int. Conf Cyber Sci. Technol. Congr. (DASC/PiCom/CBDCom/CyberSciTech)*, Aug. 2020, pp. 515–522.

[3] S. Selvaganapathy, S. Sadasivam, and V. Ravi, ''A review on Android malware: Attacks, countermeasures and challenges ahead,'' *J. Cyber Secur. Mobility*, vol. 10, pp. 177–230, Mar. 2021.

[4] G. D'Angelo, M. Ficco, and F. Palmieri, ''Malware detection in mobile environments based on autoencoders and API-images,'' *J. Parallel Distrib. Comput.*, vol. 137, pp. 26–33, Mar. 2020.

[5] I. Almomani, R. Qaddoura, M. Habib, S. Alsoghyer, A. A. Khayer, I. Aljarah, and H. Faris, ''Android ransomware detection based on a hybrid evolutionary approach in the context of highly imbalanced data,'' *IEEE Access*, vol. 9, pp. 57674–57691, 2021.

[6] V. Kouliaridis and G. Kambourakis, ''A comprehensive survey on machine learning techniques for Android malware detection,'' *Information*, vol. 12, no. 5, p. 185, Apr. 2021.

[7] i. Almomani, A. AlKhayer, and M. Ahmed, ''An efficient machine learning-based approach for Android v.11 ransomware detection,'' in *Proc. 1st Int. Conf. Artif. Intell. Data Anal. (CAIDA)*, Apr. 2021, pp. 240–244.

[8] T. H.-D. Huang and H.-Y. Kao, ''R2-D2: ColoR-inspired convolutional neural network (CNN)-based Android

malware detections,'' in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 2633–2642.

[9] H. Naeem, ''Detection of malicious activities in Internet of Things environment based on binary visualization and machine intelligence,'' *Wireless Pers. Commun.*, vol. 108, no. 4, pp. 2609–2629, Oct. 2019.

[10] F. Ullah, H. Naeem, S. Jabbar, S. Khalid, M. A. Latif, F. Al-Turjman, and L. Mostarda, ''Cyber security threats detection in Internet of Things using deep learning approach,'' *IEEE Access*, vol. 7, pp. 124379–124389, 2019.

[11] D. Vasan, M. Alazab, S. Wassan, H. Naeem, B. Safaei, and Q. Zheng, ''IMCFN: Image-based malware classification using fine-tuned convolutional neural network architecture,'' *Comput. Netw.*, vol. 171, Apr. 2020, Art. no. 107138.

[12] Z. Ren, H. Wu, Q. Ning, I. Hussain, and B. Chen, ''End-to-end malware detection for Android IoT devices using deep learning,'' *Ad Hoc Netw.*, vol. 101, Apr. 2020, Art. no. 102098.

[13] W. Chao, L. Qun, W. XiaoHu, R. TianYu, D. JiaHan, G. GuangXin, and S. EnJie, ''An Android application vulnerability mining method based on static and dynamic analysis,'' in *Proc. IEEE 5th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Jun. 2020, pp. 599–603.

[14] M. Ziadia, J. Fattahi, M. Mejri, and E. Pricop, ''Smali+: An operational semantics for low-level code generated from reverse engineering Android applications,'' *Information*, vol. 11, no. 3, p. 130, Feb. 2020.

[15] M. Gonçalves and A. C. R. Paiva, ''Reverse engineering of Android applications: REiMPAcT,'' in *Proc. Int. Conf. Quality Inf. Commun. Technol.* Faro, Portugal: Springer, 2020, pp. 369–382.