JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Study of Android-Malware Prediction using Machine Learning

[1]Shrutika Singh, [2]Dr. Sadhna K. Mishra
[1]Research Scholar, [2]Professor & Head
Department of Computer Science & Engineering
Lakshmi Narain College of Technology, Bhopal, India

Abstract : The proliferation of mobile devices, particularly those running the Android operating system, has led to an unprecedented rise in the threat landscape of mobile malware. As malicious actors continue to exploit vulnerabilities and deploy sophisticated attack techniques, the need for proactive and effective defense mechanisms becomes imperative. This study delves into the realm of Android malware prediction using machine learning, aiming to develop robust models capable of identifying and mitigating potential threats in real-time. By leveraging the power of advanced machine learning algorithms, this research seeks to enhance the security posture of Android devices and contribute to the ongoing efforts in cybersecurity.

***Index Terms* – Android, Malware, Artificial Intelligence, Secuiry, Machine learning.**

I. INTRODUCTION

The widespread availability of mobile devices that run on the Android operating system has transformed communication and access to information, making them an essential component of our day-to-day life. On the other hand, this extensive acceptance has also drawn the attention of hostile actors that are looking to damage the integrity and privacy of users via the deployment of Android malware. This harmful software, which vary from relatively simple adware to more complex malware and ransomware, pose major threats to personally identifiable information as well as data belonging to organizations.

In order to keep up with the ever-changing nature of Android malware, traditional signature-based techniques of malware detection have been shown to be ineffective. As a consequence of this, there is an increasing need for ways that are both predictive and proactive in order to recognize and prevent possible dangers before they have the opportunity to do damage. As a result of its capacity to recognize patterns and irregularities within big datasets, machine learning has emerged as a potentially useful approach for improving the accuracy and efficiency of Android malware detection.

The purpose of this research is to investigate the possibility of incorporating machine learning strategies into the field of Android malware prediction. The specific objective is to create models that are capable of analyzing a wide range of characteristics and behaviours that are linked with dangerous apps. The project aims to train models that are capable of generalizing across various forms of Android malware by using past data and extracting important characteristics. This will result in the provision of a defensive mechanism that is both scalable and adaptive.

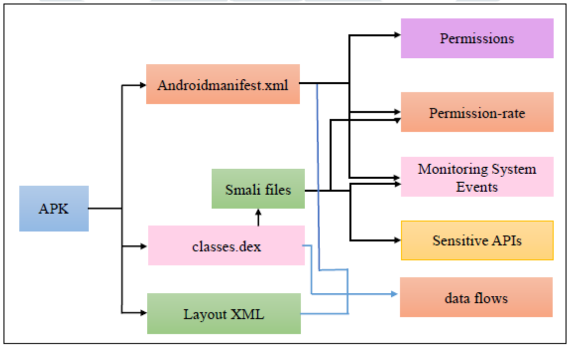

Figure 1: Android Malware

This study aims to accomplish a number of goals, including the identification of important characteristics that are indicative of dangerous behaviour, the selection and optimization of machine learning methods, and the assessment of the effectiveness of the suggested models against a variety of Android malware samples. The purpose of this research is to give useful insights to the area of mobile security by conducting an exhaustive analysis of real-world datasets. This will help to stimulate the creation of defensive mechanisms that are more robust and intelligent in order to combat the constantly shifting terrain of Android malware threats.

II. BACKGROUND

H. Alamro et al.,[1] Current technological advancement in computer systems has transformed the lives of humans from real to virtual environments. Malware is unnecessary software that is often utilized to launch cyber-attacks. Malware variants are still evolving by using advanced packing and obfuscation methods. These approaches make malware classification and detection more challenging. New techniques that are different from conventional systems should be utilized for effectively combating new malware variants. Machine learning (ML) methods are ineffective in identifying all complex and new malware variants. The deep learning (DL) method can be a promising solution to detect all malware variants. This paper presents an Automated Android Malware Detection using Optimal Ensemble Learning Approach for Cybersecurity (AAMD-OELAC) technique.

H. Zhu et al.,[2] The popularity of the Android platform in smartphones and other Internet-of-Things devices has resulted in the explosive of malware attacks against it. Malware presents a serious threat to the security of devices and the services they provided, e.g. stealing the privacy sensitive data stored in mobile devices. This work raises a stacking ensemble framework SEDMDroid to identify Android malware. Specifically, to ensure individual's diversity, it adopts random feature subspaces and bootstrapping samples techniques to generate subset, and runs Principal Component Analysis (PCA) on each subset. The accuracy is probed by keeping all the principal components and using the whole dataset to train each base learner Multi-Layer Perception (MLP). Then, Support Vector Machine (SVM) is employed as the fusion classifier to learn the implicit supplementary information from the output of the ensemble members and yield the final prediction result. The second one, a public big dataset, extracts the sensitive data flow information as the features, and the average accuracy is 94.92%. Promising experiment results reveal that the proposed method is an effective way to identify Android malware.

A. Alzubaidi et al.,[3] In recent years, the global pervasiveness of smartphones has prompted the development of millions of free and commercially available applications. These applications allow users to perform various activities, such as communicating, gaming, and completing financial and educational tasks. These commonly used devices often store sensitive private information and, consequently, have been increasingly targeted by harmful malicious software. This paper focuses on the concepts and risks associated with malware, and reviews current approaches and mechanisms used to detect malware with respect to their methodology, associated datasets, and evaluation metrics.

H. Kato et al.,[4] presented Android malware detection based on a Composition Ratio (CR) of permission pairs. We define the CR as a ratio of a permission pair to all pairs in an app. We focus on the fact that the CR tends to be small in malware because of unnecessary permissions. To obtain features without using the frequencies, we construct databases about the CR. For each app, we calculate similarity scores based on the databases. Finally, eight scores are fed into machine learning (ML) based classifiers as features. By doing this, stable performance can be achieved. Since our features are just eight-dimensional, the proposed scheme takes less training time and is compatible with other ML based schemes. Furthermore, our features can quantitatively offer clear information that helps human to understand detection results. Our scheme is suitable for practical use because all the requirements can be met.

C. Li et al et al.,[5] Machine learning (ML) has been widely used for malware detection on different operating systems, including Android. To keep up with malware's evolution, the detection models usually need to be retrained periodically (e.g., every month) based on the data collected in the wild. However, this leads to poisoning attacks, specifically backdoor attacks, which subvert the learning process and create evasion tunnels for manipulated malware samples. To date, we have not found any prior research that explored this critical problem in Android malware detectors. In this paper, we are motivated to study the backdoor attack against Android malware detectors. The backdoor is created and injected into the model stealthily without access to the training data and activated when an app with the trigger is presented.

L. Gong, Z. Li et al.,[6] Android overlay enables one app to draw over other apps by creating an extra View layer atop the host View, which nevertheless can be exploited by malicious apps (malware) to attack users. To combat this threat, prior countermeasures concentrate on restricting the capabilities of overlays at the OS level while sacrificing overlays usability; recently, the overlay mechanism has been substantially updated to prevent a variety of attacks, which however can still be evaded by considerable adversaries. To address these shortcomings, a more pragmatic approach is to enable early detection of overlay-based malware during the app market review process, so that all the capabilities of overlays can stay unchanged.

I. Almomani et al.et al.,[7] presented a new methodology for the detection of Ransomware that is depending on an evolutionary-based machine learning approach. The binary particle swarm optimization algorithm is utilized for tuning the hyperparameters of the classification algorithm, as well as performing feature selection. The support vector machines (SVM) algorithm is used alongside the synthetic minority oversampling technique (SMOTE) for classification. The utilized dataset is collected from various sources, which consists of 10,153 Android applications, where 500 of them are Ransomware. The performance of the proposed approach SMOTE- tBPSO-SVM achieved merits over traditional machine learning algorithms by having the highest scores in terms of sensitivity, specificity, and g-mean.

F. Mercaldo and A. Santone et al.,[8] Several techniques to overcome the weaknesses of the current signature based detection approaches adopted by free and commercial anti-malware were proposed by industrial and research communities. These techniques are mainly supervised machine learning based, requiring optimal class balance to generate good predictive models. In this paper, we propose a method to infer mobile application maliciousness by detecting the belonging family, exploiting formal equivalence checking.

L. N. Vu and S. Jung, "AdMat et al.,[9] The availability of big data and affordable hardware have enabled the applications of deep learning on different tasks. With respect to security, several attempts have been made to transfer deep learning’s application from the domain of image recognition or natural language processing into malware detection. In this study, we propose AdMat - a simple yet effective framework to characterize Android applications by treating them as images. The novelty of our study lies in the construction of an adjacency matrix for each application. These matrices act as “input images” to the Convolutional Neural Network model, allowing it to learn to differentiate benign and malicious apps, as well as malware families.

L. Gong et al et al.,[10], machine learning (ML) techniques have been widely explored for automated, robust malware detection, but till now we have not seen an ML-based malware detection solution applied at market scales. To systematically understand the real-world challenges, we conduct a collaborative study with T-Market, a popular Android app market that offers us large-scale ground-truth data. Our study illustrates that the key to successfully developing such systems is multifold, including feature selection and encoding, feature engineering and exposure, app analysis speed and efficacy, developer and user engagement, as well as ML model evolution. Failure in any of the above aspects could lead to the “wooden barrel effect” of the whole system. This article presents our judicious design choices and first-hand deployment experiences in building a practical ML-powered malware detection system.

III. CHALLENGES

The study of Android malware prediction using machine learning is not without its challenges. Addressing these challenges is crucial for the development and deployment of effective and reliable malware detection systems. Here are some key challenges associated with this research area:

1. **Dataset Imbalance:**

    - **Issue:** Datasets often suffer from class imbalance, where the number of benign applications significantly outweighs that of malicious ones. This imbalance can lead machine learning models to favor accuracy over effectiveness in identifying malware.

    - **Challenge:** Designing strategies to balance datasets, such as oversampling minority classes or using advanced techniques like synthetic data generation, is essential to ensure models learn and generalize well across both benign and malicious instances.

2. **Feature Extraction and Selection:**

    - **Issue:** Identifying relevant features that effectively differentiate between benign and malicious behavior is challenging. Moreover, selecting the most informative features from a vast pool can be complex.

    - **Challenge:** Developing robust feature extraction mechanisms and employing feature selection algorithms are crucial for enhancing the efficiency and accuracy of machine learning models. This involves domain expertise to ensure the inclusion of meaningful features and the exclusion of irrelevant ones.

3. **Dynamic Malware Behavior:**

    - **Issue:** Malware is adaptive and can exhibit dynamic behavior, making it challenging to create models that can accurately predict future malicious activities based on historical data.

    - **Challenge:** Incorporating temporal and dynamic features into machine learning models is essential. Techniques like sequence modeling and recurrent neural networks (RNNs) can be employed to capture the evolving nature of Android malware.

4. **Evasion Techniques:**

    - **Issue:** Malware developers continually evolve evasion techniques to circumvent detection mechanisms, including those based on machine learning.

    - **Challenge:** Developing models that are resilient to adversarial attacks and can adapt to new evasion techniques is critical. Regular model updates and the incorporation of anomaly detection methods can contribute to staying ahead of emerging threats.

5. **Resource Constraints on Mobile Devices:**

    - **Issue:** Mobile devices often have limited computational resources, including processing power and memory, which can restrict the implementation of resource-intensive machine learning models.

    - **Challenge:** Designing lightweight models that can run efficiently on mobile devices without compromising detection accuracy is essential. This involves optimizing algorithms and leveraging techniques such as model quantization and compression.

6. **Privacy Concerns:**

    - **Issue:** Gathering and analyzing sensitive data for training machine learning models may raise privacy concerns among users.

- **Challenge:** Implementing privacy-preserving techniques, such as federated learning or differential privacy, to ensure that user data is protected while still contributing to the collective improvement of malware detection models.

IV. CONCLUSION

This paper shows the study of Android malware prediction using machine learning represents a crucial frontier in the ongoing battle to secure mobile devices against evolving threats. This research endeavors to harness the power of advanced computational techniques to proactively identify and mitigate the risks posed by malicious applications on the Android platform. Through an interdisciplinary approach that combines expertise in machine learning, cybersecurity, and mobile computing, the proposed strategy aims to overcome the myriad challenges inherent in this field. In future we will implement efficient machine learning classification algorithm to prediction of the android malware.

REFERENCES

1. H. Alamro, W. Mtouaa, S. Aljameel, A. S. Salama, M. A. Hamza and A. Y. Othman, "Automated Android Malware Detection Using Optimal Ensemble Learning Approach for Cybersecurity," in IEEE Access, vol. 11, pp. 72509-72517, 2023, doi: 10.1109/ACCESS.2023.3294263.
2. H. Zhu, Y. Li, R. Li, J. Li, Z. You and H. Song, "SEDMDroid: An Enhanced Stacking Ensemble Framework for Android Malware Detection," in IEEE Transactions on Network Science and Engineering, vol. 8, no. 2, pp. 984-994, 1 April-June 2021, doi: 10.1109/TNSE.2020.2996379.
3. A. Alzubaidi, "Recent Advances in Android Mobile Malware Detection: A Systematic Literature Review," in IEEE Access, vol. 9, pp. 146318-146349, 2021, doi: 10.1109/ACCESS.2021.3123187.
4. H. Kato, T. Sasaki and I. Sasase, "Android Malware Detection Based on Composition Ratio of Permission Pairs," in IEEE Access, vol. 9, pp. 130006-130019, 2021, doi: 10.1109/ACCESS.2021.3113711.
5. C. Li et al., "Backdoor Attack on Machine Learning Based Android Malware Detectors," in IEEE Transactions on Dependable and Secure Computing, doi: 10.1109/TDSC.2021.3094824.
6. L. Gong, Z. Li, H. Wang, H. Lin, X. Ma and Y. Liu, "Overlay-based Android Malware Detection at Market Scales: Systematically Adapting to the New Technological Landscape," in IEEE Transactions on Mobile Computing, doi: 10.1109/TMC.2021.3079433.
7. I. Almomani et al., "Android Ransomware Detection Based on a Hybrid Evolutionary Approach in the Context of Highly Imbalanced Data," in IEEE Access, vol. 9, pp. 57674-57691, 2021, doi: 10.1109/ACCESS.2021.3071450.
8. F. Mercaldo and A. Santone, "Formal Equivalence Checking for Mobile Malware Detection and Family Classification," in IEEE Transactions on Software Engineering, doi: 10.1109/TSE.2021.3067061.
9. L. N. Vu and S. Jung, "AdMat: A CNN-on-Matrix Approach to Android Malware Detection and Classification," in IEEE Access, vol. 9, pp. 39680-39694, 2021, doi: 10.1109/ACCESS.2021.3063748.
10. L. Gong et al., "Systematically Landing Machine Learning onto Market-Scale Mobile Malware Detection," in IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 7, pp. 1615-1628, 1 July 2021, doi: 10.1109/TPDS.2020.3046092.
11. A. Pandey, A. Chaturvedi, M. Gupta, Praveen Kumar Mannepalli, S. Kumar and G. Chhabra, "An Automated Face Mask Detection System using Deep CNN on AWS Cloud Infrastructure," 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2023, pp. 648-655, doi: 10.1109/ICESC57686.2023.10193710.
12. Praveen Kumar Mannepalli, A. Khan, P. Chugh, S. Patka and R. Ponmalar, "Classification of Pepper Bell into Healthy and Bacterial Spot Using Deep Learning," 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI), Greater Noida, India, 2023, pp. 727-732, doi: 10.1109/ICCSAI59793.2023.10421243.
13. D. K. Rathore and Praveen Kumar Mannepalli, "A Review of Machine Learning Techniques and Applications for Health Care," 2021 International Conference on Advances in Technology, Management & Education (ICATME), Bhopal, India, 2021, pp. 4-8, doi: 10.1109/ICATME50232.2021.9732761.
14. R. Singh and Praveen Kumar Mannepalli, "Cloud Malicious Threat Detection Using Convolution Filter and EBPNN," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-6, doi: 10.1109/ISCON52037.2021.9702492.
15. R. Singh and Praveen Kumar Mannepalli, "Invasive Weed optimization Algorithm Based Trained Neural Network for Cloud Malicious Threat Detection," 2021 IEEE International Conference on Technology, Research, and Innovation for Betterment of Society (TRIBES), Raipur, India, 2021, pp. 1-5, doi: 10.1109/TRIBES52498.2021.9751674.
16. Tiwari, S.K., Praveen Kumar Mannepalli, (2022). Wolf Algorithm Based Routing and Adamic Adar Trust for Secured IOT Network. In: Kumar, A., Fister Jr., I., Gupta, P.K., Debayle, J., Zhang, Z.J., Usman, M. (eds) Artificial Intelligence and Data Science. ICAIDS 2021. Communications in Computer and Information Science, vol 1673. Springer, Cham. https://doi.org/10.1007/978-3-031-21385-4_42
17. Rashmi Singh; Praveen Kumar Mannepalli "Cloud malicious threat detection by features from intelligent water drop set and EBPN" International journal of advanced research in engineering and technology (ijaret) > volume 11, issue 12, december 2020, DOI:10.34218/IJARET.11.12.2020.086

18. Rashmi Singh; Praveen Kumar Mannepalli “Cloud malicious threat detection by features from intelligent water drop set and EBPN” International journal of advanced research in engineering and technology (ijaret) volume 11, issue 12, december 2020, DOI:10.34218/IJARET.11.12.2020.086
19. R. Sisodiya, Praveen Kumar Mannepalli, "Invasive Weed Optimization Based Sentiment Mining of Digital Review Content," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-6, doi: 10.1109/ISCON52037.2021.9702424.
20. K. Anil, Praveen Kumar Mannepalli, "Achieving Effective Secrecy based on Blockchain and Data Sharing in Cloud Computing," 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2021, pp. 709-716, doi: 10.1109/CSNT51715.2021.9509663.