International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 13, Issue 4, April 2024)

Automated Android Malware Detection Using Machine Learning Approach for Cybersecurity: A Review

Vashnvi Sharma[1], Dr.Damodar Tiwari[2], Dr.Shital Gupta[3], Twinkle Sharma[4]
[1]Research Scholar, Dept. of CSE, Bansal Institute of Science and Technology, Bhopal, India
[2]Professor, Dept. of CSE, Bansal Institute of Science and Technology, Bhopal, India
[3]Associate Professor, Dept. of CSE, Bansal Institute of Science and Technology, Bhopal, India
[4]Assistant Professor, Dept. of CSE, Bansal Institute of Science and Technology, Bhopal, India

***Abstract*— As the proliferation of Android devices continues, the threat landscape of malware targeting these platforms has intensified, necessitating advanced cybersecurity measures. This review systematically explores the landscape of automated Android malware detection with a specific focus on machine learning approaches. By surveying recent research, methodologies, datasets, and evaluation metrics, this review aims to provide a comprehensive understanding of the state-of-the-art techniques, challenges, and future directions in leveraging machine learning for enhancing cybersecurity on Android platforms.**

***Keywords*— Attack, Cyber, Artificial Intelligence, Android, Malware, Secuiry.**

I. INTRODUCTION

The advent of the digital age has transformed the way we communicate, work, and conduct our daily lives, with mobile devices playing a pivotal role in this technological evolution[1]. Among these devices, Android-based smartphones have emerged as ubiquitous companions, offering a myriad of functionalities to billions of users worldwide. However, this widespread adoption has also attracted the attention of malicious actors seeking to exploit vulnerabilities for various nefarious purposes, giving rise to the pressing need for robust cybersecurity measures. This introduction delves into the intricate landscape of automated Android malware detection, emphasizing the pivotal role of machine learning approaches in fortifying cybersecurity defenses[2].

Mobile devices, particularly smartphones, have become an integral part of modern society, facilitating seamless connectivity, information access, and personalized experiences. Android, as an open-source operating system, has gained dominance in the mobile market, powering a diverse range of devices [3]. The open nature of the Android ecosystem, while fostering innovation, has also opened the door to an escalating threat landscape. Malicious actors exploit vulnerabilities in Android's architecture, deploying sophisticated malware that poses significant risks to users' privacy, sensitive data, and overall digital security[4].

The evolution of Android malware is marked by its diversity and adaptability. Malware on Android devices encompasses a spectrum of threats, from traditional viruses and worms to more advanced forms such as ransomware, spyware, and trojans[5]. The dynamic nature of these threats necessitates equally dynamic and adaptive cybersecurity measures to counteract the ever-changing tactics employed by cybercriminals[6].

The motivation behind this research lies in the imperative need to address the escalating threat of Android malware and enhance the security posture of Android devices[7]. The pervasive use of smartphones for a plethora of activities, including financial transactions, communication, and data storage, underscores the criticality of safeguarding these devices from malicious intrusions. Automated malware detection stands as a frontline defense mechanism, capable of identifying and neutralizing threats before they can compromise user security[8].

Machine learning, with its capacity to analyze vast datasets, identify patterns, and adapt to evolving scenarios, has emerged as a transformative force in the realm of cybersecurity. The motivation to explore machine learning approaches for Android malware detection stems from the

desire to leverage advanced technologies to stay ahead of malicious actors[9]. By understanding the patterns and behaviors of malware, machine learning models can contribute to more effective and efficient detection, providing a proactive defense against the ever-evolving landscape of Android threats [10].

This review aims to comprehensively explore recent advancements in automated Android malware detection, focusing specifically on the integration of machine learning approaches. The scope encompasses a broad spectrum of methodologies, datasets, and evaluation metrics employed in contemporary research. By synthesizing information from existing literature, this review seeks to provide a comprehensive overview of the state-of-the-art techniques, identify challenges, and propose potential future directions in leveraging machine learning for Android cybersecurity.

Figure 1: Android malware

Conducting a search for harmful software on Android is an absolutely necessary step. Methods of detection that are based on permission pairs have a great deal of promise for use in practical detection. Different detecting methods are available in a wide variety [11]. On the other hand, conventional approaches are unable to concurrently fulfill requirements for practical application in terms of efficiency, intelligibility, and stability of detection performance [12]. This is because conventional methods are not suitable for real applications. Having these prerequisites in place is meant to guarantee that the approach may be used in an efficient manner. Even if the most current technique is based on distinctions between frequent pairings of harmless applications and malicious software, it is not solid enough to meet the criteria [13]. This is because the requirements are always changing. This is because new malware has a propensity to seek greater permissions in order to mimic harmless apps, which makes the use of the frequencies ineffective [14].

II. LITERATURE SURVEY

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] presents a stacking outfit structure SEDMDroid to recognize Android malware. In particular, to guarantee person's variety, it takes on arbitrary element subspaces and bootstrapping tests strategies to create subset, and runs Head Part Examination (PCA) on every subset. The precision is tested by keeping all the primary parts and utilizing the entire dataset to prepare each base student Multi-facet Discernment (MLP). Then, at that point, Backing Vector Machine (SVM) is utilized as the combination classifier to gain the verifiable beneficial data from the result of the gathering individuals and yield the last expectation result.

H. Kato et al.,[3] To acquire highlights without utilizing the frequencies, it is develop information bases regarding the CR. For each application, it is ascertain closeness scores in view of the data sets. At last, eight scores are taken care of into AI (ML) based classifiers as elements. By doing this, steady exhibition can be accomplished. Since our highlights are only eight-layered, the present plot takes less preparation time and is viable with other ML based plans. Moreover, our highlights can quantitatively offer clear data that assists human with understanding discovery results. Our plan is reasonable for useful use since every one of the prerequisites can be met. By utilizing genuine datasets, our outcomes show that our plan can identify malware with up to 97.3% exactness. Moreover, contrasted and a current plan, our plan can diminish the component aspects by around close to 100% with keeping up with equivalent precision on late datasets.

C. Li et al et al.,[4] AI (ML) has been generally utilized for malware location on various working frameworks, including Android. To stay aware of malware's advancement, the recognition models for the most part should be retrained intermittently (e.g., consistently) in light of the information gathered in nature. Be that as it may, this prompts harming assaults, explicitly indirect access assaults, which undermine the learning system and make avoidance burrows for controlled malware tests. Until now, it is have not found any earlier examination that investigated this basic issue in Android malware indicators.

L. Gong et al.,[5] To address these deficiencies, a more logical methodology is to empower early identification of overlay-based malware during the application market audit process, with the goal that every one of the abilities of overlays can remain unaltered. For this reason, in this work it is first lead an enormous scope near investigation of overlay attributes in harmless and vindictive applications, and afterward execute the Overlay Checker framework to naturally recognize overlay-based malware for one of the universes biggest Android application stores. Specifically, it is have put forth deliberate attempts in include designing, UI investigation, copying engineering, and run-time climate, hence keeping up with high identification exactness (97% accuracy and 97% review) and short per-application filter time (1.7 minutes) with just two ware servers, under a serious responsibility of 10K recently submitted applications each day.

I. Almomani et al.et al.,[6] lately, Ransomware has been a basic danger that assaults cell phones. Ransomware is a sort of malware that impedes the portable's framework and forestalls the client of the contaminated gadget from getting to their information until a payment is paid. Around the world, Ransomware assaults have prompted genuine misfortunes for people and partners. Be that as it may, the sensational increment of Ransomware families makes to the method involved with recognizing them more testing because of their consistently developed attributes.

F. Mercaldo et al.,[7] A few methods to defeat the shortcomings of the momentum signature based recognition approaches took on by free and business against malware were present by modern and exploration networks. These procedures are mostly regulated AI based, requiring ideal class equilibrium to produce great prescient models. In this work, it is propose a strategy to derive versatile application perniciousness by distinguishing the having a place family, taking advantage of formal proportionality checking. it is acquaint a bunch of heuristics with lessen the quantity of portable application examinations and it is characterize a measurement mirroring the application vindictiveness. True tests on 35 Android malware families (going from 2010 to 2018) affirm the viability of the present technique in portable malware recognition and family distinguishing proof.

L. N. Vu et al.,[8] The accessibility of enormous information and reasonable equipment have empowered the utilizations of profound learning on various errands. Regarding security, a few endeavors have been made to move profound gaining's application from the space of picture acknowledgment or regular language handling into malware location. In this review, it is propose AdMat - a straightforward yet compelling structure to describe Android applications by regarding them as pictures. The curiosity of our review lies in the development of a nearness grid for every application. These lattices go about as "input pictures" to the Convolutional Brain Organization model, permitting it to figure out how to separate harmless and noxious applications, as well as malware families.

L. Gong et al et al.,[9] Regardless of being essential to the present portable biological system, application markets have in the interim turned into a characteristic, helpful malware conveyance channel as they in fact "loan believability" to noxious applications. In the beyond couple of years, AI (ML) strategies has been broadly investigated for computerized, hearty malware location, however till now it is have not seen a ML-based malware discovery arrangement applied at market scales. To deliberately comprehend this present reality challenges, it is lead a cooperative report with T-Market, a well known Android application market that offers us huge scope ground-truth information.

W. Yuan et al.,[10] To defeat this test, it is plan a lightweight on-gadget Android malware locator, in view of the as of late present expansive learning strategy. Our locator principally involves a single shot calculation for model preparation. Consequently it very well may be completely or steadily prepared straightforwardly on cell phones. All things considered, our finder beats the shallow learning-based models, including support vector machine (SVM) and AdaBoost, and approaches the profound learning-based models multi-facet perceptron (MLP) and convolutional brain organization (CNN). Additionally, our indicator is more powerful to ill-disposed models than the current identifiers, and its heartiness can be additionally worked on through on-gadget model retraining. At last, its benefits are affirmed by broad examinations, and its reasonableness is exhibited through runtime assessment on cell phones.

K. Liu et al.,[11] Android applications are growing quickly across the versatile biological system, yet Android malware is likewise arising in an interminable stream. Numerous scientists have concentrated on the issue of Android malware identification and have advanced speculations and techniques according to alternate points of view. Existing exploration proposes that AI is a powerful and promising method for distinguishing Android malware. Regardless, there exist audits that have reviewed various issues connected with Android malware location in light of AI.

D. Li et al.,[12] Malware stays a major danger to network protection, calling for AI based malware recognition. While promising, such identifiers are known to be defenseless against avoidance assaults. Group advancing normally works with countermeasures, while assailants can use this procedure to further develop assault adequacy also. This rouses us to examine which sort of heartiness the gathering protection or adequacy the outfit assault can accomplish, especially when they battle with one another. it is accordingly propose another assault approach, named combination of assaults, by delivering assailants equipped for various generative techniques and different control sets, to bother a malware model without destroying its malevolent usefulness.

Q. Han, et al.,[13] presented FARM beats standard baselines when no attacks occur. Though we cannot guess all possible attacks that an adversary might use, we propose three realistic attacks on FARM and show that FARM is very robust to these attacks in all classification problems. Additionally, FARM has automatically identified two malware samples which were not previously classified as rooting malware by any of the 61 anti-viruses on VirusTotal. These samples were reported to Google's Android Security Team who subsequently confirmed our findings.

J. Ribeiro et al.,[14] Previous research efforts on developing an Intrusion Detection and Prevention Systems (IDPS) for Android mobile devices rely mostly on centralized data collection and processing on a cloud server. However, this trend is characterized by two major limitations. First, it requires a continuous connection between monitored devices and the server, which might be infeasible, due to mobile network's outage or partial coverage. Second, it increases the risk of sensitive information leakage and the violation of user's privacy. To help alleviate these problems, in this paper, we develop a novel Host-based IDPS for Android (HIDROID), which runs completely on a mobile device, with a minimal computation burden. It collects data in run-time, by periodically sampling features reflecting the utilization of scarce resources on a mobile device (e.g. CPU, memory, battery, bandwidth, etc.).

X. Wang et al.,[15] presented, to thoroughly investigate Android kernel behaviors, we first present a kernel feature based framework, CrowdNet, for cloud computing platforms. CrowdNet includes an automatic data provider that collects footprints of kernel features and a parallel malware predictor that validates Android malicious behaviors. Then we calculate and select hidden centers by a heuristic approach for 12,750 Android applications to reduce the number of iterations and time complexity. Our experimental results show that CrowdNet protects large-scale data validation and speeds up the learning of kernel behaviors twofold. Further, identifying malicious attacks with CrowdNet improves the classification efficiency compared to traditional neural network and other machine learning techniques.

Y. Zhang et al et al.,[16] learned representation is then fed into our outlier-aware clustering to partition the weakly-labeled malware into known and unknown families. The malware whose malicious behaviours are close to those of the existing families on the network, are further classified using a three-layer Deep Neural Network (DNN). The unknown malware are clustered using a standard density-based clustering algorithm. We have evaluated our approach using 5,416 ground-truth malware from Drebin and 9,000 malware from VirusShare (uploaded between Mar. 2017 and Feb. 2018), consisting of 3324 weakly-labeled malware. The evaluation shows that Andre effectively clusters weakly-labeled malware which cannot be clustered by the state-of-the-art approaches, while achieving comparable accuracy with those approaches for clustering ground-truth samples.

S. Aonzo et al.,[17] introduces BAdDroIds, a mobile application leveraging machine learning for detecting malware on resource constrained devices. BAdDroIds executes in background and transparently analyzes the applications as soon as they are installed, i.e., before infecting the device. BAdDroIds relies on static analysis techniques and features provided by the Android OS to build up sound and complete models of Android apps in terms of permissions and API invocations. It uses ad-hoc supervised classification techniques to allow resource-efficient malware detection. By exploiting the intrinsic nature of data, it has been possible to implement a state-of-the-art data-driven model which provides deep insights on the detection problem and can be efficiently executed on the device itself as it requires a very limited computational effort.

H. Zhang et al.,[18] presented a novel Android malware detection method based on the method-level correlation relationship of application's abstracted API calls. First, we split each Android application's source code into separate function methods and just keep the abstracted API calls of them to form a set of abstracted API calls transactions. And then, we calculate the confidence of association rules between the abstracted API calls, which forms behavioral semantics to describe an application.

R. Kumar et al.,[19] presented a novel framework that combines the advantages of both machine learning techniques and blockchain technology to improve the malware detection for Android IoT devices. The proposed technique is implemented using a sequential approach, which includes clustering, classification, and blockchain. Machine learning automatically extracts the malware information using clustering and classification technique and store the information into the blockchain. Thereby, all malware information stored in the blockchain history can be communicated through the network, and therefore any latest malware can be detected effectively.

T. Lei et al.,[20] presented to use event group to describe apps' behaviors in event level, which can capture higher level of semantics than in API level. In event group, we adopt function clusters to represent behaviors in each event so that behaviors hidden in events can still be captured as time goes on, which enables EveDroid to detect new malware in the event level. The function clusters can generalize API calls into vectors based on their API composition to capture new API calls, which makes EveDroid scalable to malware evolving. Moreover, a neural network is specifically designed to aggregate the multiple events and automatically mine the semantic relationship among them. We train the system and evaluate its F1-measure on a dataset of 14 956 benign and 28 848 malicious Android apps released in different years. The experimental results show that EveDroid outperforms other malware detection systems.

III. CHALLENGES

The landscape of automated Android malware detection is riddled with challenges that necessitate constant innovation and adaptation to stay ahead of evolving threats. This section explores key challenges encountered in the field, ranging from issues inherent to the Android ecosystem to complexities associated with the dynamic nature of malware.

Android Ecosystem Diversity

One of the primary challenges in automated Android malware detection is the vast diversity within the Android ecosystem itself. The multitude of device manufacturers, varying hardware specifications, and customized versions of the Android operating system contribute to a highly fragmented environment. As a result, designing detection mechanisms that are universally effective across all devices becomes a complex task. Machine learning models trained on one subset of devices may not generalize well to others, leading to potential blind spots in malware detection.

Polymorphic and Evolving Malware

Android malware exhibits a high degree of polymorphism, wherein the malicious code continually mutates to evade detection by traditional signature-based methods. Polymorphic malware presents a formidable challenge for machine learning models, as they need to adapt to rapidly changing patterns. The dynamic nature of malware, with new strains emerging regularly, requires constant model retraining and updates to ensure efficacy against the latest threats.

Class Imbalance

Class imbalance, a common challenge in machine learning, is particularly pronounced in Android malware detection. The vast majority of Android applications are benign, leading to imbalanced datasets where the number of malware samples is significantly smaller than benign ones. This imbalance can bias the learning process, causing models to be overly conservative in classifying samples, leading to increased false negatives.

Adversarial Attacks

Adversarial attacks pose a significant threat to the robustness of machine learning models in Android malware detection. Malicious actors actively seek to manipulate input data to deceive models into misclassifying malware as benign or vice versa. Crafting adversarial samples that are subtly modified yet effective at evading detection becomes a sophisticated game of cat and mouse, challenging the resilience of machine learning models.

Resource Constraints on Mobile Devices

The limited computational resources on mobile devices pose practical challenges for deploying resource-intensive machine learning models. Real-time detection with minimal impact on device performance is crucial for user experience. Striking a balance between model complexity and

computational efficiency becomes paramount, requiring the development of lightweight yet accurate models suitable for deployment on resource-constrained Android devices.

Lack of Explain ability in Models

The inherently complex nature of some machine learning models, particularly deep learning architectures, often results in a lack of interpretability or explain ability. Understanding why a model makes a specific decision is crucial for building trust in automated detection systems. Addressing this challenge involves developing models that not only provide accurate predictions but also offer insights into the features contributing to those predictions.

Privacy Concerns

As machine learning models become increasingly sophisticated, there is a growing concern regarding user privacy. Some advanced detection mechanisms may require access to extensive user data, raising questions about the ethical use of such information. Striking a balance between effective detection and safeguarding user privacy is an ongoing challenge in the development of automated Android malware detection systems.

IV. CONCLUSION

The review of automated Android malware detection using machine learning approaches highlights significant progress in addressing the escalating cybersecurity threats posed by malicious software targeting Android devices. Machine learning techniques have emerged as powerful tools for automating the detection process, offering the potential to identify and mitigate malware at scale. Through the examination of various ML algorithms, feature extraction methods, datasets, and evaluation metrics, it becomes evident that researchers have made strides in developing sophisticated detection systems. However, several challenges persist, including dataset imbalance, feature selection, and the adaptability of models to new malware variants.

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 2022, doi: 10.1109/TNSE.2020.2996379.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. W. Yuan, Y. Jiang, H. Li and M. Cai, "A Lightweight On-Device Detection Method for Android Malware," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 9, pp. 5600-5611, Sept. 2021, doi: 10.1109/TSMC.2019.2958382.
11. K. Liu, S. Xu, G. Xu, M. Zhang, D. Sun and H. Liu, "A Review of Android Malware Detection Approaches Based on Machine Learning," in IEEE Access, vol. 8, pp.

124579-124607, 2020, doi: 10.1109/ACCESS.2020.3006143.

12. D. Li and Q. Li, "Adversarial Deep Ensemble: Evasion Attacks and Defenses for Malware Detection," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 3886-3900, 2020, doi: 10.1109/TIFS.2020.3003571.
13. Q. Han, V. S. Subrahmanian and Y. Xiong, "Android Malware Detection via (Somewhat) Robust Irreversible Feature Transformations," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 3511-3525, 2020, doi: 10.1109/TIFS.2020.2975932.
14. J. Ribeiro, F. B. Saghezchi, G. Mantas, J. Rodriguez and R. A. Abd-Alhameed, "HIDROID: Prototyping a Behavioral Host-Based Intrusion Detection and Prevention System for Android," in IEEE Access, vol. 8, pp. 23154-23168, 2020, doi: 10.1109/ACCESS.2020.2969626.
15. X. Wang, C. Li and D. Song, "CrowdNet: Identifying Large-Scale Malicious Attacks Over Android Kernel Structures," in IEEE Access, vol. 8, pp. 15823-15837, 2020, doi: 10.1109/ACCESS.2020.2965954.
16. Y. Zhang et al., "Familial Clustering for Weakly-Labeled Android Malware Using Hybrid Representation Learning," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 3401-3414, 2020, doi: 10.1109/TIFS.2019.2947861.
17. S. Aonzo, A. Merlo, M. Migliardi, L. Oneto and F. Palmieri, "Low-Resource Footprint, Data-Driven Malware Detection on Android," in IEEE Transactions on Sustainable Computing, vol. 5, no. 2, pp. 213-222, 1 April-June 2020, doi: 10.1109/TSUSC.2017.2774184.
18. H. Zhang, S. Luo, Y. Zhang and L. Pan, "An Efficient Android Malware Detection System Based on Method-Level Behavioral Semantic Analysis," in IEEE Access, vol. 7, pp. 69246-69256, 2019, doi: 10.1109/ACCESS.2019.2919796.
19. R. Kumar, X. Zhang, W. Wang, R. U. Khan, J. Kumar and A. Sharif, "A Multimodal Malware Detection Technique for Android IoT Devices Using Various Features," in IEEE Access, vol. 7, pp. 64411-64430, 2019, doi: 10.1109/ACCESS.2019.2916886.
20. T. Lei, Z. Qin, Z. Wang, Q. Li and D. Ye, "EveDroid: Event-Aware Android Malware Detection Against Model Degrading for IoT Devices," in IEEE Internet of Things Journal, vol. 6, no. 4, pp. 6668-6680, Aug. 2019, doi: 10.1109/JIOT.2019.2909745.