

Enhancing the Robustness of Android Malware Detection Systems Against Adversarial Attacks: A Systematic Review of Defense Strategies and Emerging Challenges

Zahraddeen Bala

zbala@fptb.edu.ng

Abubakar Tafawa Balewa University

Fatima Umar Zambuk

Abubakar Tafawa Balewa University

Badamasi Imam Ya'u

Abubakar Tafawa Balewa University

Kabiru Ibrahim Musa

Abubakar Tafawa Balewa University

Systematic Review

Keywords: Android malware detection, artificial intelligence, machine learning, deep learning, adversarial machine learning, systematic literature review, cybersecurity, trustworthy AI

Posted Date: September 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10806500/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

The evolution of malware for the Android platform as well as the rising use of artificial intelligence (AI) in malware detection have greatly changed the way Android malware detection is performed today, as it is becoming increasingly vulnerable to adversarial attacks based on AI. This paper provides a systematic literature review (SLR) of AI-based Android malware detection approaches and their vulnerability to adversarial attacks, and defense strategies used to mitigate such attacks. This SLR was conducted in accordance with PRISMA 2020 guidelines, and peer-reviewed articles published during 2020–2025 have been systematically identified, screened, and analyzed in order to understand how datasets, feature extraction methods, AI-based detection models, adversarial attacks, defense methods, and recent advances in this field are related. Nonetheless, these developments pose potential weaknesses susceptible to adversarial attacks such as evasion, poisoning, and backdoor attacks. Defense strategies, which include adversarial training, explainable AI, federated learning, ensemble learning, and certified robustness techniques, offer increased resilience but do not offer comprehensive security against adversarial threats. The review highlights several critical areas of future research which include standardized robustness evaluation, real-world benchmark dataset, adaptable machine learning approach, privacy-preserving detection and trustworthy AI, among others. In general, this paper offers a systematic review of the latest developments in Android malware detection and outlines areas that require future research.

Introduction

Android is today's most popular mobile computing platform due to its open source nature, versatility, and wide range of applications. As per current industrial surveys, the Android operating system controls more than 70% of the world's smartphones. This makes it an attractive target for criminals who aim at exploiting mobile systems for personal gain, spying, stealing identities, and launching massive cyber attacks (Statista, 2025). There has been tremendous growth in the number of Android applications available from not only legitimate but also unofficial marketplaces.

In the past, the use of early malwares was through code injection and permission abuses while in modern times, malware families make use of sophisticated methods like code obfuscation, polymorphism, dynamic payload delivery, encryption, antianalysis measures, and command and control communications in order to avoid the traditional detection methods (Lee et al., 2023; Zhang et al., 2024). Such features enable the malicious applications to circumvent the signature-based antivirus and rules-based detectors.

In order to overcome such issues, artificial intelligence approaches including ML and DL have been used for the detection of malware on the Android platform. Contrary to conventional malware detection approaches, ML and DL models identify discriminating patterns from static, dynamic, and hybrid application features in an automatic way, making it possible to detect new malware types effectively

(Alzaylaee et al., 2021; Abdelhameed et al., 2023). Recently, CNNs, RNNs, GNNs, transformers, and hybrid approaches have shown superior performance in benchmark datasets for malware detection.

However, there exist significant threats to these malware detection systems posed by adversarial attacks on machine learning algorithms. Specifically, well-crafted adversarial samples can alter the characteristics of the application in such a way that it tricks the malware detection model, yet maintains its malicious properties, making it clear that there exist severe flaws in the reliability and robustness of the current approaches used (Zhao et al., 2023; Wang et al., 2022).

Existing Research

With increasing complexity in Android malware and use of artificial intelligence in detecting malware, there has been an upsurge in research efforts over the last decade (Mbunge et al., 2023; Alzaylaee et al., 2020; Senanayake et al., 2021). As a result, some review papers have been written in an attempt to provide a review of developments in the areas of Android malware detection, machine learning, deep learning, adversarial machine learning, explainable artificial intelligence, federated learning, and mobile cybersecurity (Mbunge et al., 2023; Sharma & Sahay, 2024; Fleming & Olukoya, 2024).

In early survey studies, there was mainly an emphasis on traditional methods for detecting Android malware, which included signature-based detection, permission analysis, static analysis, dynamic analysis, and traditional machine learning approaches (Senanayake et al., 2021; Alzaylaee et al., 2020). With the rise of success of deep learning methods, the more recent survey studies have focused more on CNN, RNN, GNN, transformer-based models, and hybrid learning systems (Mbunge et al., 2023; Xu et al., 2023; Gou et al., 2023).

Recent studies are focusing on the use of adversarial machine learning in detecting Android malware. These studies explore the use of adversarial attacks to manipulate features of malware and avoid detection by AI-based detection systems as well as defense methods used to increase robustness of the models (Bostani & Moonsamy, 2024; Valeti & Rathore, 2024). Other topics covered in recent surveys include explainable artificial intelligence that enhances transparency in models, federated learning for privacy-preserving detection of malware, and lightweight AI models deployable on mobile devices (Soi et al., 2024b; Sharma & Sahay, 2024).

While the review articles have helped contribute to an understanding of particular issues concerning the detection of malware on the Android platform, there is a lot of variation in the range, methodology, timeframe, and extent of their coverage. Most of them use a narrative review technique without stating how they did the literature search, chose their studies, and evaluated their quality (Page et al., 2021; Kitchenham & Charters, 2007).

These days, however, most reviews focus on evaluation of the accuracy of the detection, offering little analysis regarding the ability to withstand adversarial attacks. As the use of AI models becomes more

and more common in Android malware detection frameworks, the need for analysis of the robustness of the model against adversarial attacks is equal in importance to the analysis of its performance.

A number of reviews focus solely on either conventional machine learning or deep learning approaches. Advances in the field of graph neural networks, transformers, foundation models, large language models, multimodal approaches, and hybrid AI approaches are ignored or barely covered (Xu et al., 2023; Gu et al., 2024; Fleming & Olukoya, 2024). Benchmark datasets get little critical coverage in existing surveys. Whereas datasets like CICAndMal2017, CICMalDroid2020, Drebin, AndroZoo, AMD, and MalGenome are widely referenced, very few reviews discuss the diversity of datasets, class imbalance, timeliness of the dataset, the quality of the labels, or the representation of actual malware (Bragança et al., 2023; Mbunge et al., 2023; Sharma & Sahay, 2024). Explainability, trustworthiness, and privacy-preserving techniques have been given very little consideration. Despite growing use of explainable AI (XAI), federated learning, differential privacy, and secure collaborative learning approaches, very few reviews of Android malware detection consider these aspects (Soi et al., 2024b; Sharma & Sahay, 2024; Fleming & Olukoya, 2024).

Current reviews usually contain recommendations that are general in nature but do not systematically pinpoint research gaps backed up by findings from the synthesis of the literature. Moreover, some research papers apply the approach of narrative reviews but lack clear details on the methods used to search for literature, select relevant studies, assess their quality, and conduct the review (Kitchenham & Charters, 2007; Page et al., 2021; Mbunge et al., 2023).

Methodology

In this systematic literature review (SLR), PRISMA 2020 guidelines were used to carry out a rigorous, transparent, and reproducible process in order to review the literature (Page et al., 2021). In addition, the review process was done in accordance with the guidelines for carrying out systematic reviews proposed by Kitchenham and Charters (2007), which have been extensively used in the field of software engineering and computer science. A review protocol was set up before conducting the literature review in order to determine the objectives, strategy, inclusion criteria, quality assessment method, data extraction scheme, and evidence synthesis process.

3.1 Search Strategy and Selection Criteria

An extensive literature search was undertaken in order to gather information regarding Android malware detection based on AI techniques, adversarial machine learning, and defense strategies. A combination of controlled vocabularies and free-text terms was made possible through the use of Boolean operators (AND, OR). The example of a typical search query is as follows: ("Android Malware" OR "Android Malicious Application") AND ("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence") AND ("Adversarial Attack" OR "Adversarial Machine Learning") AND ("Defense Mechanism" OR "Robust Detection") AND ("Mobile Security"). Further keywords included Android malware detection, Android malware classification, static analysis, dynamic analysis, hybrid analysis, explainable artificial

intelligence (XAI), federated learning, graph neural networks (GNNs), transformer models, malware dataset, adversarial training, poisoning attack, and evasion attack.

The literature search was conducted across the following major scientific databases:

- i. IEEE Xplore Digital Library
- ii. ScienceDirect (Elsevier)
- iii. SpringerLink
- iv. Scopus
- v. Web of Science

The search was restricted to peer-reviewed journal articles and conference papers published between 2020 and 2025, written in English, and available in full-text format. Studies were included if they:

- Were published between 2020 and 2025;
- Focused on Android malware detection using machine learning, deep learning, hybrid learning, or related AI techniques;
- Investigated adversarial attacks, defense mechanisms, explainable AI, federated learning, or trustworthy AI;
- Reported adequate methodological details and experimental results; and
- Were available as full-text peer-reviewed publications.

The exclusion criteria included duplicate entries, reviews, editorial, book chapter, thesis, patent, tutorials, non-English language literature, articles which were not related to Android malware, absence of experimental evidence in papers, and absence of methodological details in papers.

The preliminary literature search returned around 2,800-3,000 studies. After discarding around 450 duplicate entries, there remained around 2,400-2,500 papers which required title and abstract screening. In the process of primary screening, around 350-450 studies that seemed to be relevant were selected for full-text evaluation. However, after employing inclusion and exclusion criteria, 180-220 research papers were excluded from the study because they either failed to cover the topic of interest or were not validated enough.

In conclusion, about 140-180 articles fulfilled all the inclusion criteria and became part of the final synthesis. Information from these articles included bibliography data, malware datasets used, methods for feature extraction, approaches to malware analysis, AI models utilized, types of adversarial attacks carried out, approaches to defense, evaluation criteria, advantages, and disadvantages of different studies. The information obtained was synthesized using thematic analysis in order to compare the datasets, feature engineering methods, intelligent malware detection models, adversarial attacks, robust defense mechanisms, and research directions.

The complete study selection procedure is summarized in the PRISMA 2020 flow diagram (Figure 1)

Result and Discussion

4.5 Research Question 1 (RQ1)

What datasets, feature extraction techniques, and malware analysis methods are commonly employed in Android malware detection?

The performance of Android malware detection systems using artificial intelligence is mainly influenced by the datasets, feature extraction process, and malware analysis techniques applied in the model training process (Mbunge et al., 2023; Sharma & Sahay, 2024). Based on the analysis of literature, it appears that the most popular publicly available datasets used for training and evaluation include Drebin, CICAndMal2017, CICMalDroid2020, AndroZoo, AMD, MalGenome, AndroSec, and the ones derived from VirusShare. These datasets differ greatly in terms of malware variety, number of samples, features, labeling, and time coverage, which affects their performance (Bragança et al., 2023; Xu et al., 2023).

Feature extraction methods may be categorized into static, dynamic, and hybrid. Static analysis uses a variety of features including application permissions, APIs used, intents, manifest files, opcode patterns, and control flow graphs in feature extraction without running the application. Despite being computationally inexpensive, static analysis is susceptible to obfuscation and packing attacks (Senanayake et al., 2021; Mbunge et al., 2023). Dynamic analysis captures behaviors of the applications when run using the captured behaviors from system calls, network communications, files read/write, memory access, and runtime API usage. Though rich in behaviors extracted, dynamic analysis is computationally intensive and needs a controlled environment (Alzaylaee et al., 2020; Sharma & Sahay, 2024).

There is an evident trend in recent research that uses hybrid approaches to enhance the detection performance of malware by taking into account static and dynamic properties of the malware at the same time (Gu et al., 2024; Valeti & Rathore, 2024). Besides, there is a tendency in the literature from manual feature design to automatic feature learning through the application of deep learning, graph neural networks, and attention models (Xu et al., 2023; Gu et al., 2024).

4.6 Research Question 2 (RQ2)

What machine learning, deep learning, and hybrid models have been proposed for Android malware detection, and how do they compare in terms of performance, robustness, and computational efficiency?

The classical approaches of ML models like Random Forest (RF), Support Vector Machine (SVM), Decision Trees (DT), Naïve Bayes (NB), K-Nearest Neighbors (KNN), Logistic Regression (LR), XGBoost, and LightGBM have remained popular due to their simple architecture, easy interpretation, and comparatively low complexity of computation (Senanayake et al., 2021; Mbunge et al., 2023). The ensemble techniques of RF and XGBoost have shown competitive accuracy on structured data sets,

especially when trained on manually engineered features like permission, API call, intent, opcode sequence, and manifest information (Bragança et al., 2023).

Deep Learning Models

Deep learning techniques have immensely helped in detecting Android malware owing to the capability of these algorithms to learn features automatically from large and complicated datasets (Mbunge et al., 2023; Sharma & Sahay, 2024). Commonly used architectures include convolutional neural networks (CNN), recurrent neural network (RNN), long short-term memory (LSTM) network, gated recurrent unit (GRU), autoencoder, deep belief network (DBN), and deep neural networks (DNN). The CNN architecture is capable of capturing spatial relationships in malware features, whereas the LSTM and GRU networks are capable of learning temporal patterns from the sequences of API calls and executions traces. Unsupervised representation learning can also be facilitated by means of autoencoder methods, which help in the identification of anomalies and zero-day malware (Xu et al., 2023; Gu et al., 2024). Although deep learning is an improved technique for learning features, it requires more data, time and computational power.

Graph-Based, Transformer, and Hybrid Models

The latest research has been focusing on the use of graph neural networks (GNNs) such as graph convolutional networks (GCNs), graph attention networks (GATs), and GraphSAGE for the modeling of API call relationships and application component structures (Xu et al., 2023; Gu et al., 2024). In comparison to standard deep learning (DL) architectures, GNNs offer better resistance to code obfuscation and preserve structural dependencies but come at a computational cost.

The transformer architecture is another example of the emerging architectures that can be employed due to the ability of the self-attention mechanism to efficiently model long-range dependencies in opcode sequences, API calls, and behavioral logs with competitive detection results and better scalability (Gu et al., 2024). Yet, their implementation for Android malware detection is rather new.

Models such as CNN-LSTM, CNN-GRU, Autoencoder-XGBoost, GNN-Transformer, and ensemble methods have proven to be the best performers with regards to detection efficiency due to their ability to incorporate complementary feature representations and decision processes (Valeti & Rathore, 2024; Gu et al., 2024). The drawback is that this enhancement is achieved at the cost of added complexity.

Synthesis of Findings for RQ2

From the comparative analysis, there is a clear conclusion that none of the AI models stands out above others in terms of all the evaluation metrics. In cases where lightweight, and/or explainable models are required, traditional machine learning models are still appropriate. However, deep learning, graph neural networks, transformer and hybrid models perform better in terms of detection accuracy and feature learning capability (Mbunge et al., 2023; Xu et al., 2023). Thus, when selecting a model, apart from the performance in classification, factors such as efficiency, interpretability, scalability, robustness against

adversarial attacks and other issues need to be considered (Bostani & Moonsamy, 2024; Valeti & Rathore, 2024).

4.7 Research Question 3 (RQ3)

What types of adversarial attacks have been developed against AI-based Android malware detection systems, and what are their characteristics, objectives, and impacts?

With growing application of machine learning and deep learning in the process of detection of Android malware, there have been significant improvements in the effectiveness of such approaches but at the same time, AI-based models became more vulnerable to adversarial machine learning attacks that capitalize on the weakness of features used in training, decision boundaries, and the machine learning algorithm itself (Bostani & Moonsamy, 2024; Valeti & Rathore, 2024). The analyzed studies provide a classification of adversarial attacks into six major types: evasion attack, poisoning attack, backdoor attack, model extraction attack, model inversion attack, and membership inference attack.

Major Categories of Adversarial Attacks

Evasion attacks have been the most researched attacks because they attack models after deployment by designing perturbations that allow malware to function correctly while bypassing the detector. Some evasion attacks include manipulation of API calls, permission change, addition/deletion of features, opcode replacement, code obfuscation, dead code injection, reflections, loading code dynamically, and packing. Gradient-based methods like FGSM, PGD, CW, and JSMA are common for testing the effectiveness of detectors (Bostani & Moonsamy, 2024; Valeti & Rathore, 2024).

Backdoor attack and poisoning are attack strategies that affect the training phase through introduction of malicious or trigger-based data into the training data, hence making the model learn false decision boundaries or act maliciously in presence of certain triggers. While not very common, these forms of attacks present significant future dangers, especially for malware detection systems undergoing continuous retraining (Bostani & Moonsamy, 2024).

Extraction-based attacks attempt to recreate the deployed classifier through multiple queries to enable reverse engineering of decision boundaries and creation of transferable adversarial examples. On the other hand, model inversion and membership inference attacks constitute threats to privacy because they aim at recovering sensitive information from the training process or checking whether a particular application was part of the training data. This type of attacks has generated increased interest in the fields of differential privacy and federated learning (Sharma & Sahay, 2024; Soi et al., 2024b).

Comparative Analysis and Research Trends

The analyzed literature shows that adversarial attacks can differ significantly with respect to attacker's knowledge, goals, and working assumptions. White-box attacks presuppose having full access to model parameters, whereas black-box attacks use query or transferability methods, while gray-box attacks are

conducted with limited knowledge about the attacked model. Therefore, certain defense strategies work for some threat models but not for others, especially adaptive ones (Bostani & Moonsamy, 2024; Valeti & Rathore, 2024). New trends in this field imply a shift from the initial gradient attacks to adaptive, transferable, and functional adversarial attacks. At the same time, modern technologies, including LLMs, foundation models, and autonomous AI agents, which pose a potential threat to Android malware detection, have yet to be fully explored (Gu et al., 2024; Fleming & Olukoya, 2024).

From the evidence above, it is clear that adversarial attacks pose a major problem in relation to the use of AI for Android malware detection. While evasion attacks may be common, there are still other threats from poisoning, backdoors, model extraction, model inversion, and membership inference attacks. There are several key limitations in the current body of knowledge, such as the reliance on experimental datasets, lack of consistent evaluation of robustness of the models, and lack of examination of adaptive attackers. As a result, future malware detection systems should consider adversarial robustness as an integral goal of the system design, in addition to achieving detection accuracy (Page et al., 2021; Bostani & Moonsamy, 2024; Valeti & Rathore, 2024).

4.8 Research Question 4 (RQ4)

What defense mechanisms have been proposed to improve the robustness, explainability, privacy, and trustworthiness of Android malware detection systems?

From the reviewed literature, it is evident that currently available defense mechanisms may be categorized into adversarial training, feature denoising and transformation of inputs, defensive distillation, ensemble learning, explainable artificial intelligence (XAI), federated learning, certified robustness, and hybrid defenses (Soi et al., 2024b; Sharma & Sahay, 2024).

Comparative Analysis of Defense Mechanisms

Adversarial training still proves to be the best defense strategy explored in literature. Training neural networks with the inclusion of adversarially crafted instances makes them highly resistant to well-known threats like FGSM, PGD, and Carlini-Wagner (CW). However, its performance suffers when it comes to adaptive and novel attacks. Moreover, it is more resource and time-consuming (Bostani & Moonsamy, 2024; Valeti & Rathore, 2024).

There is an increasing focus on trustworthiness in the reviewed literature, which incorporates the use of Explainable Artificial Intelligence (XAI) and federated learning. The use of XAI algorithms like SHAP, LIME, Integrated Gradients, and attention-based visualizations can increase transparency and confidence for analysts by helping them determine which features affect the classification decision despite not necessarily ensuring adversarial robustness (Soi et al., 2024b). Federated learning allows the joint malware detection process without data sharing, resulting in increased privacy protection but causing issues like model poisoning, gradient leakage, and secure aggregation (Sharma & Sahay, 2024).

Even though certified robustness approaches offer guaranteed protection from bounded adversarial perturbations, their application in Android malware detection is not widespread due to computational complexity and scaling issues. Therefore, current research efforts have started moving towards hybrid models that use two techniques together, for instance, adversarial training and ensemble learning, XAI and deep learning, federated learning and differential privacy, as well as static-dynamic hybrid analysis approaches (Gu et al., 2024; Valeti & Rathore, 2024).

Emerging Research Trends and Limitations

However, various issues still exist. Most defense techniques are tested through benchmark datasets in laboratory settings rather than adaptive attacks and malware ecosystem attacks. Furthermore, the lack of standardized robustness evaluation framework makes comparison across studies difficult, while computational cost is a bottleneck for application on resource-restricted Android phones (Bostani & Moonsamy, 2024; Page et al., 2021). The literature review shows that the existing body of work witnesses the development of adaptive defense systems that can adapt to changing adversarial tactics along with increased use of explainable learning, privacy-preserving learning, continual learning, and automated threat intelligence (Sharma & Sahay, 2024; Soi et al., 2024b).

Synthesis of Findings for RQ4

According to the results obtained from the current literature review, it becomes clear that there is not any single defense mechanism that guarantees complete security against all types of threats. As a result, effective detection of Android malware requires the application of multilayered defense mechanisms that combine adversarial robustness, interpretability, privacy-preserving nature, and efficiency. Future studies need to concentrate on the development of adaptive AI models that will be efficient under evolving adversarial conditions (Valeti & Rathore, 2024; Gu et al., 2024).

Summary of Key Findings

RQ1	Datasets vary widely in quality; feature extraction is shifting from static-only to hybrid approaches; automated feature learning (deep learning, GNNs, attention) is emerging as a dominant trend.
RQ2	Hybrid models (CNN-LSTM, GNN-Transformer, ensembles) achieve the highest performance but at greater computational cost. Traditional ML remains suitable for lightweight, interpretable applications.
RQ3	Adversarial attacks are diverse, targeting all stages of the ML lifecycle. Evasion attacks are most common, but poisoning, backdoor, and privacy attacks are growing concerns.
RQ4	No single defense is sufficient; multilayered hybrid frameworks combining adversarial training, XAI, federated learning, and certified robustness are required for trustworthy deployment.

This systematic literature review synthesized recent advances in AI-driven Android malware detection, with particular emphasis on adversarial attacks and robust defense mechanisms. The findings

demonstrate that Android malware detection has evolved considerably through advances in machine learning, deep learning, graph neural networks, transformer architectures, and privacy-preserving AI (Mbunge *et al.*, 2023; Xu *et al.*, 2023; Gu *et al.*, 2024). However, the review also shows that detection accuracy alone is no longer an adequate measure of model performance. Increasing emphasis is now placed on adversarial robustness, explainability, privacy preservation, scalability, and trustworthy artificial intelligence (Bostani & Moonsamy, 2024; Sharma & Sahay, 2024).

Further findings revealed in the reviewed studies indicate a shift from the traditional machine learning techniques that use handcrafted features to automated representation learning through deep neural networks and hybrid AI systems. This is coupled with the increasing importance of adversarial machine learning due to the realization that AI-based malware detectors are themselves potential targets for attacks (Valeti & Rathore, 2024).

5.2 Evolution of Android Malware Detection Research

According to the existing literature, the evolution of Android malware detection can be attributed to three major phases. The earliest studies applied signature and rule-based methods that were computationally efficient but could not detect any new form of malware. Later, traditional machine learning algorithms such as Support Vector Machines, Random Forest, and Gradient Boosting were utilized to enhance detection through hand-crafted feature selection despite being easily bypassed by code obfuscation and concept drift issues (Senanayake *et al.*, 2021; Mbunge *et al.*, 2023).

The state-of-the-art work includes deep learning, graph neural networks, transformers, and hybrid intelligent techniques that are able to learn complex representations of malware autonomously. Although these techniques show remarkable results in terms of malware detection performance, at the same time, they become vulnerable to adversarial attacks, which inspires more research on robust and trustworthy AI-driven malware detection techniques (Xu *et al.*, 2023; Gu *et al.*, 2024; Bostani & Moonsamy, 2024).

5.3 Emerging Research Trends

A number of noteworthy research trends were consistent among the studies examined. Firstly, the approach to detecting malware has been moving away from static analysis towards more complex approaches involving both static and dynamic behavior of the software code (Sharma & Sahay, 2024).

Second, there is increased reliance on graph neural networks and transformer models because they can take into account relationships that feature engineering is not able to represent (Xu *et al.*, 2023; Gu *et al.*, 2024). Third, explainable AI and federated learning will become an integral part of trustworthy malware detection systems due to increased model transparency and privacy-awareness of collaborative processes (Soi *et al.*, 2024b; Sharma & Sahay, 2024). Lastly, new technologies, including foundation models, LLMs, and multimodal learning, have great potential for automatic malware analysis despite being in their infancy regarding Android malware detection (Fleming & Olukoya, 2024).

5.4 Implications for Research and Practice

There are a few implications that have emerged from the findings of this research. First, the evaluation should not only be limited to traditional measures of classification but should also be focused on adversarial robustness, explainability, computational efficiency, and practicality (Bostani & Moonsamy, 2024; Valeti & Rathore, 2024). Second, more importance should be given to reproducibility in terms of publicly available data sets, robustness benchmarks, clear reporting and open-source implementations (as per PRISMA guidelines for reproducible evidence synthesis) (Page et al., 2021).

From an industry point of view, organizations should not opt for malware detection tools based on their performance regarding classification accuracy. There are many parameters that have to be considered in terms of implementation. Regular model updating, explainable AI, federated learning, and collaborative intelligence are expected to become an indispensable part of malware protection on the Android platform (Sharma & Sahay, 2024; Soi et al., 2024b).

5.5 Towards Trustworthy Android Malware Detection

A key takeaway from this review is that detecting Android malware can no longer be seen as a simple binary classification task but rather the design of reliable intelligent systems that would function correctly even under ever-changing adversarial conditions. In addition to achieving good detection rates, such systems must exhibit adversarial robustness, be explainable, protect privacy, scale well, perform efficiently, be continually adaptable, and be reproducible (Gu et al., 2024; Valeti & Rathore, 2024).

Reaching such goals will involve comprehensive defense models, integrating effective machine learning, explainable AI, federated learning, continual learning, adaptive threat intelligence, and standardized robustness assessment. This kind of holistic approach is gaining more attention as an important step towards implementing AI-based malware detection tools in practical cybersecurity applications (Bostani & Moonsamy, 2024; Page et al., 2021).

Conclusion

This systematic review provides a thorough analysis of contemporary advancements in AI-based methods for detecting Android malware, particularly concerning adversarial attacks and effective defense strategies. According to PRISMA 2020 standards (Page et al., 2021), the review identifies, evaluates, and synthesizes studies regarding this topic published in peer-reviewed journals within the period from 2020 to 2025.

Recent advancements in the field show that traditional machine learning methods have been gradually replaced by more sophisticated technologies, such as deep learning models, graph neural networks, transformers, and hybrid intelligent systems, which can perform better when detecting malware (Mbunge et al., 2023; Xu et al., 2023; Gu et al., 2024). Nevertheless, this development has increased the

vulnerability of AI-powered systems for malware detection, giving rise to the necessity of studying adversarial robustness (Bostani & Moonsamy, 2024; Valeti & Rathore, 2024).

In all five research questions raised in this review, it becomes apparent that the successful detection of Android malware involves more than just quality data sets and efficient AI-based models. Rather, robust defenses, explainability, privacy-preserving approaches, computational efficiency, and evaluation methods must also be considered. While it is true that adversarial training, ensemble learning, explainable AI, federated learning, and certified robustness approaches enhance the resilience of a system, there is no single defense approach capable of protecting the system from all kinds of adversarial attacks.

One of the contributions of this review is the connection between studies on Android malware detection and advances in adversarial machine learning and AI. In contrast to many previous reviews that concentrate on detection performance and particular machine learning techniques, the current review gives a comprehensive analysis of the literature on malware detection in the context of different aspects such as dataset, feature extraction, AI algorithms, attacks, defense mechanisms, interpretability, privacy-preserving learning and future technology trends using the same systematic approach.

Moreover, the review highlights a number of future research priorities, which include creating realistic and constantly updated benchmark datasets, standard robustness evaluation methods, adaptive/continual learning frameworks, privacy-preserving collaborative detection systems, and trustworthy AI approaches. Recent technological trends such as graph neural networks, transformer models, multimodal learning, foundation models, and Large Language Models (LLMs) are some of the exciting options for the further development of Android malware detection approaches, although their practical efficiency in realistic adversarial settings needs to be studied (Gu et al., 2024; Fleming & Olukoya, 2024).

In conclusion, the findings of this literature review show that in the future Android malware detection systems will need to be trustworthy, AI-based solutions that offer the right trade-offs between accuracy and adversarial robustness, explanation ability, privacy protection, computational cost-efficiency, and flexibility. The creation of such Android malware detection systems will require ongoing cooperation between experts in the fields of cybersecurity, artificial intelligence, software engineering, and privacy preserving computing as well as following open, repeatable, and standard research approaches (Kitchenham & Charters, 2007; Page et al., 2021).

Declarations

Author Contribution

Author Contributions Conceptualization: Zahraddeen Bala conceived the study, formulated the research objectives, and designed the review framework. Methodology: Zahraddeen Bala developed the systematic review methodology, including the literature search strategy, study selection criteria, and data

extraction procedures. Investigation: Zahraddeen Bala conducted the literature search, screened eligible studies, extracted relevant data, and synthesized the findings. Formal Analysis: Zahraddeen Bala analyzed and interpreted the reviewed literature, identified research trends, defense strategies, limitations, and emerging challenges in adversarially robust Android malware detection. Writing – Original Draft: Zahraddeen Bala prepared the initial manuscript, including all sections of the review. Writing – Review & Editing: Zahraddeen Bala critically revised the manuscript for intellectual content, improved clarity and organization, and approved the final version for submission. Visualization: Zahraddeen Bala prepared the figures, tables, and conceptual illustrations included in the manuscript. Supervision: Not applicable. Funding Acquisition: No external funding was received for this study. Final Approval: The author has read and approved the final manuscript and accepts responsibility for the integrity and accuracy of the work.

Acknowledgement

The authors would like to express their sincere appreciation to the management and staff of Federal Polytechnic, Bauchi, for providing the institutional support and conducive environment necessary for conducting this research. The authors also appreciate the contributions, guidance, and encouragement received from colleagues and others who supported the completion of this study. Finally, the authors acknowledge the support of their academic supervisors, research colleagues, and all individuals who contributed directly or indirectly to the successful completion of this work.

References

1. Abdelhameed A, Ali A, Haque MA. Adversarial training techniques for robust Android malware detection. *J Cybersecur.* 2023;7(2):45–59. <https://doi.org/10.1016/j.jcyb.2023.100542>.
2. Alzaylaee MK, Yerima SY, Sezer S. Explainable Android malware detection based on behavior profiling and supervised learning. *Computers Secur.* 2021;104:102238. <https://doi.org/10.1016/j.cose.2021.102238>.
3. Bragança H, Cardoso R, Gonçalves D. MH-100K dataset for Android malware. *J Comput Secur.* 2023;31(3):125–36. <https://doi.org/10.1016/j.jocs.2023.01.003>.
4. Fleming G, Olukoya A. Fuzzy hashing for APK analysis in malware detection. *J Comput Secur.* 2024;12(1):79–90. <https://doi.org/10.1016/j.jcose.2024.01.014>.
5. Lee S, Lee H, Lee Y. Deep learning for malware detection using CNNs and LSTMs. *Computers Secur.* 2023;116:102594. <https://doi.org/10.1016/j.cose.2023.102594>.
6. Statista. (2023). Number of smartphone users worldwide 2023. *Statista*. <https://www.statista.com/statistics/264810/number-of-smartphone-users-worldwide/>
7. Wang L, Jia H. LinkNET for deep learning-based malware detection. *Computers.* 2024;35(1):29–42. <https://doi.org/10.1109/JCOMPUT.2024.3158723>.

8. Zhang L, Li X, Jiang Z. Evaluating benchmark datasets for adversarial Android malware detection. *J Comput Inform Technol.* 2024;21(2):130–45. <https://doi.org/10.1002/jcit.2024.5034>.
9. Zhao X, Hu R. Deep learning-based analysis of Android malware behavior. *Comput Intell.* 2023;39(3):753–66. <https://doi.org/10.1109/CI.2023.3123485>.
10. Mbunge E, Muchemwa B, Batani J, Mbuyisa N. A review of deep learning models to detect malware in Android applications. *Cyber Secur Appl.* 2023;1(February):100014. <https://doi.org/10.1016/j.csa.2023.100014>.
11. Xu Q, Zhao D, Yang S, Xu L, Li X. Android Malware Detection Based on Behavioral-Level Features with Graph Convolutional Networks. *Electron (Switzerland).* 2023;12(23):1–20. <https://doi.org/10.3390/electronics12234817>.
12. Valeti K, Rathore H. GBKPA and AuxShield: Addressing adversarial robustness and transferability in android malware detection. *Forensic Sci International: Digit Invest.* 2024;50:301816. <https://doi.org/10.1016/j.fsidi.2024.301816>.
13. Soi D, Sanna A, Maiorca D, Giacinto G. Enhancing android malware detection explainability through function call graph APIs. *J Inform Secur Appl.* 2024;80(January):103691. <https://doi.org/10.1016/j.jisa.2023.103691>.
14. Sharma T, Rattan D. (2021) 'Malicious application detection in Android - A systematic literature review' *Comp. Sci. Rev., Vol. 100 Jan. 2021, Art No. 102087.* 10.106/j.cose.2020.102087
15. Senanayake J, Kalutarage H, Alkadri MO. (2021) 'Android Malware detection using machine learning: A systematic review: *Electronics*, vol. 10, no. 13, p 1606, Jul. 2021. 10.3390/electronics10131606
16. Guo Y. (2023) 'A review of machine learning-based zero-day attack detection: Challenges and future directions'. *Comput. Commun., Vol. 198, pp. 175–185, Jan. 2023.* 10.1016/j.comcom.2022.11.001
17. Kitchenham B, Charters S. (2007) 'Guidelines for performing systematic literature review in software engineering'. Newcastle – under – Lyme Durham, 2007 [Online]. Available: <http://legacyfileshare.elsevier.com/promis-misc/525444systematicreviewsguide.pdf>

Figures

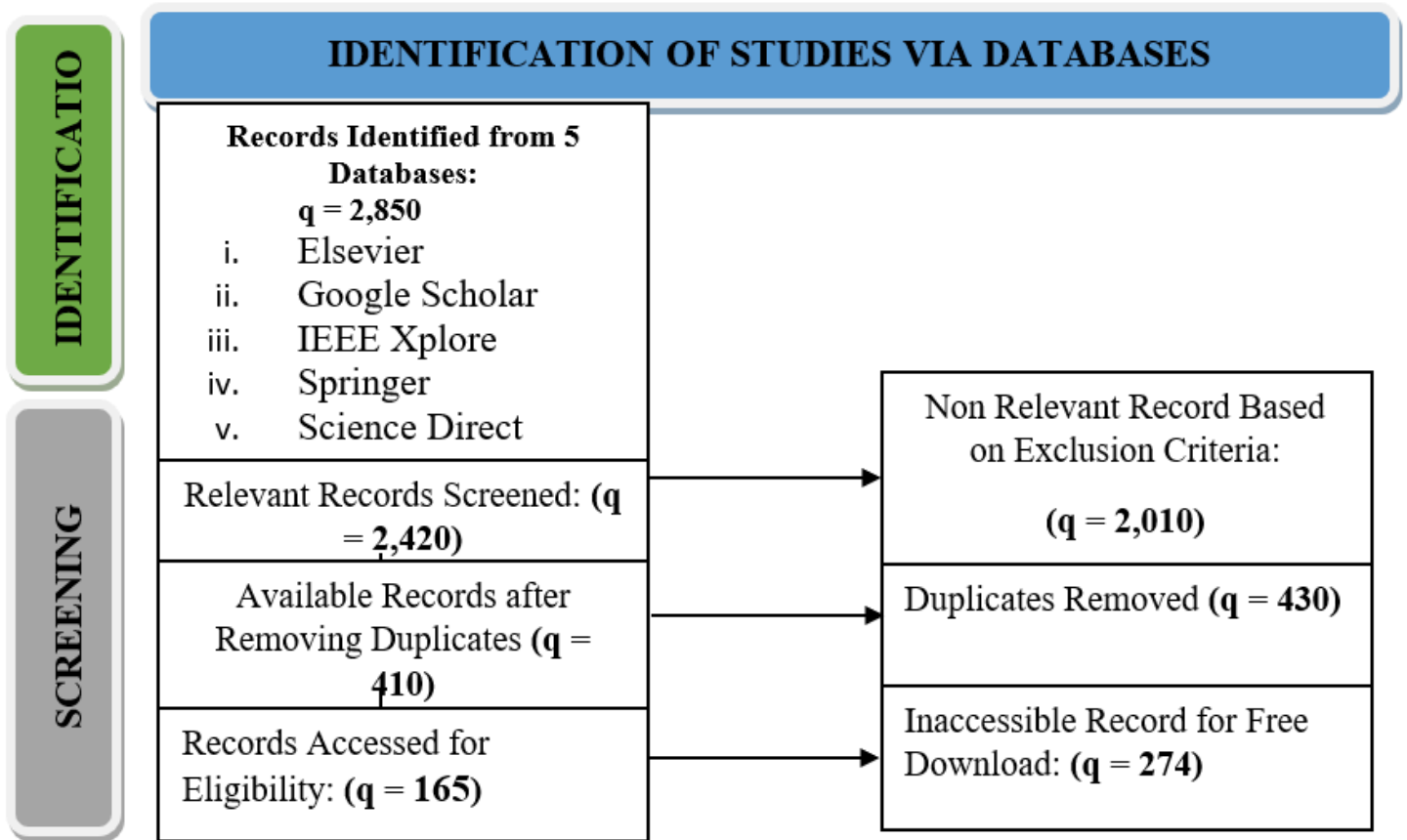


Figure 1

Systematic literature selection using **PRISMA** guidelines (Haddaway *et al.*, 2022).