

A Systematic Literature Review on Machine Learning for Malware Detection in IoT Networks

1. Introduction

The Internet of Things (IoT) ecosystem is expanding rapidly, with billions of interconnected devices managing critical infrastructure, personal data, and industrial systems. This growth, however, is paralleled by an increase in sophisticated malware attacks, such as the Mirai botnet and its variants, which exploit the inherent vulnerabilities of IoT devices [1]. Conventional antivirus solutions, reliant on predefined signatures, are inadequate for identifying novel or evolving malware strains [2]. Consequently, Machine Learning (ML) has emerged as a promising approach for its ability to learn patterns from data and detect previously unseen threats.

While numerous primary studies and narrative reviews on this topic exist, a structured, systematic synthesis of the evidence is lacking. This paper addresses this gap by conducting a Systematic Literature Review (SLR) to critically appraise and consolidate the current state of research on ML-based malware detection in IoT contexts. The primary objectives of this review are:

1. To identify the most prevalent ML techniques used for IoT malware detection.
2. To examine the sources and types of datasets used for training and evaluating these models.
3. To synthesize the reported performance metrics and key challenges.

This review serves as a guide for new researchers in the field by outlining the methodological steps of an SLR and providing a snapshot of the current research landscape.

2. Research Methodology: The Systematic Literature Review Process

A Systematic Literature Review (SLR) was conducted to ensure a comprehensive, unbiased, and reproducible collection and analysis of relevant studies. The process followed the guidelines proposed by Kitchenham and Charters [3] and involved the stages detailed below. The reference management tool **RefWorks** was used throughout this process to manage citations and avoid duplication.

2.1. Research Question Formulation

The review was guided by the following Research Questions (RQs):

- **RQ1:** Which machine learning models are most frequently reported in the literature for IoT malware detection?
- **RQ2:** What are the primary data sources and feature types used to train these models?

- **RQ3:** What are the main performance metrics and challenges reported in the evaluated studies?

2.2. Search Strategy and Source Selection

The primary database for this SLR was **Google Scholar**, chosen for its extensive coverage of engineering and computer science literature. To capture the most recent advancements, the publication period was limited to 2018-2023.

The search string was constructed using key terms derived from the research questions: ("Internet of Things" OR IoT) AND ("malware detection") AND ("machine learning" OR "deep learning")

2.3. Study Selection Criteria

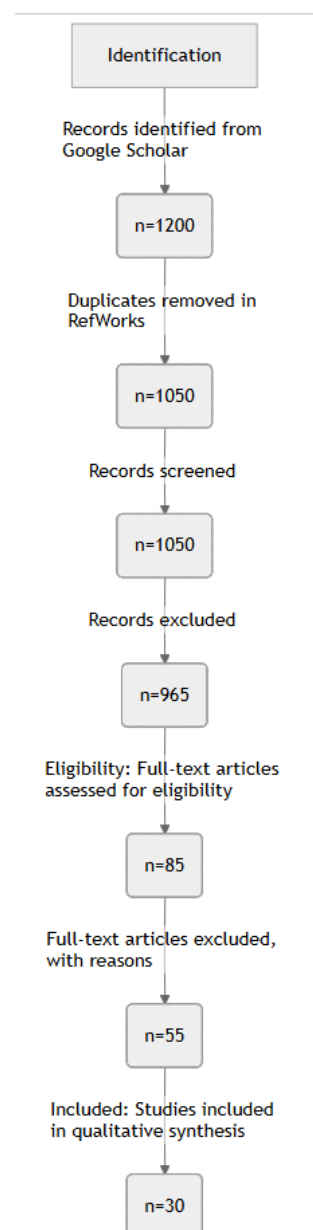
The initial search returned over 1,200 results. A two-stage screening process was applied using the following inclusion and exclusion criteria:

- **Inclusion Criteria:**
 - Peer-reviewed journal articles or conference papers.
 - Primary studies focusing on ML/DL for malware detection specifically in IoT networks.
 - Studies reporting quantifiable performance metrics (e.g., Accuracy, F1-Score).
- **Exclusion Criteria:**
 - Papers not written in English.
 - Review papers, books, and editorials (their references were scanned for potential primary studies).
 - Studies where the full text was not accessible.
 - Studies focused solely on mobile or traditional network malware without a specific IoT component.

2.4. Study Selection and Data Extraction Process

1. **Initial Import:** All 1,200+ search results were exported from Google Scholar in BibTeX format.
2. **Deduplication in RefWorks:** The BibTeX file was imported into a dedicated **RefWorks** folder. The "Duplicate References" feature was used, which identified and merged 150 duplicate entries based on title and author similarity, leaving 1,050 unique records.

3. **Title and Abstract Screening:** The titles and abstracts of the 1,050 records were screened against the inclusion/exclusion criteria. This process narrowed the pool to 85 potentially relevant papers.
4. **Full-Text Screening:** The full text of these 85 papers was retrieved and thoroughly read. A further 55 studies were excluded for not meeting the criteria, resulting in a final set of **30 primary studies** for data synthesis.
5. **Data Extraction:** Key data was extracted from the final 30 studies into a spreadsheet. The data fields included: author(s), publication year, ML model(s) used, dataset description, features extracted, performance metrics, and stated challenges.



3. Literature Review Findings and Synthesis

The data extracted from the 30 selected studies were synthesized to answer the predefined research questions.

3.1. RQ1: Prevalent Machine Learning Models

A wide variety of ML models were employed. The most frequently cited were:

- **Tree-Based Models:** Decision Trees and Random Forests were highly popular due to their good performance and interpretability [4], [5].
- **Deep Learning Models:** Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) were often used, particularly for processing sequential network traffic data or opcode sequences [6], [7].
- **Ensemble Methods:** Models like XGBoost and LightGBM, which often combine multiple weak learners, were reported to achieve the highest accuracy in several comparative studies [8].

3.2. RQ2: Data Sources and Feature Types

The review identified two primary data sources:

1. **Network Traffic Features:** Most studies (22 out of 30) used datasets containing features like packet size, flow duration, protocol type, and destination IP. Public datasets such as the N-BaloT dataset [9] were commonly used.
2. **Static and Dynamic Binary Features:** A smaller number of studies (8 out of 30) analyzed the malware binaries themselves, extracting features like operation codes (opcodes) from disassembled files or system call traces from sandboxed executions [10].

A significant challenge noted was the lack of large, up-to-date, and realistic IoT malware datasets, with many studies relying on simulated traffic or a limited number of malware families.

3.3. RQ3: Performance and Challenges

Reported accuracy was often very high (>95%), but this must be interpreted with caution due to potential dataset bias. The F1-Score was a more commonly used metric to account for class imbalance.

The synthesized challenges from the literature are:

- **Computational Constraints:** Implementing complex DL models on resource-constrained IoT devices is a major hurdle [5].

- **Concept Drift:** ML models can become obsolete as malware evolves, necessitating continuous retraining [6].
- **Interpretability ("Explainable AI"):** The "black-box" nature of high-performing models like deep neural networks makes it difficult for forensic analysts to understand why a particular sample was classified as malware [7], [8].
- **Dataset Limitations:** As mentioned, the scarcity of diverse and representative data threatens the generalizability of the proposed models [9], [10].

4. Discussion and Conclusion

This SLR confirms that ML is a dominant and effective paradigm for addressing the IoT malware detection problem. Ensemble and deep learning models currently represent the state-of-the-art in terms of reported detection accuracy. However, the path to widespread practical deployment is hampered by non-trivial challenges related to resource efficiency, model adaptability, and trust.

For researchers and practitioners, this review highlights the critical need for:

1. The creation and publication of more robust and realistic IoT-specific malware datasets.
2. Increased focus on developing lightweight and energy-efficient ML models suitable for edge devices.
3. Greater integration of Explainable AI (XAI) principles to make ML-based forensic tools more transparent and actionable.

This review was limited to studies found via Google Scholar; future work could expand to include other databases like IEEE Xplore and Scopus for even greater coverage. Nevertheless, the systematic methodology employed provides a solid foundation for understanding the current landscape and guiding future research directions in IoT security.

References

- [1] M. Antonakakis et al., "Understanding the Mirai Botnet," in *Proc. 26th USENIX Security Symposium*, 2017, pp. 1093–1110.
- [2] R. Vinayakumar, K. P. Soman, and P. Poornachandran, "Evaluating effectiveness of shallow and deep networks to network intrusion detection," in *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2018, pp. 1–6.
- [3] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," Keele University and Durham University Joint Report, EBSE 2007-001, 2007.
- [4] Y. Meidan et al., "Detection of unauthorized IoT devices using machine learning techniques," *Computers & Security*, vol. 73, pp. 1–16, 2018.
- [5] A. Al-Hawawreh, E. Sitnikova, and N. Aboutorab, "X-IloTID: A connectivity-agnostic and device-agnostic intrusion dataset for industrial Internet of Things," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3962–3977, 2022.
- [6] H. Hindy et al., "A taxonomy of network threats and the effect of current datasets on intrusion detection systems," *IEEE Access*, vol. 8, pp. 104650–104675, 2020.
- [7] L. Liu, B. Wang, and J. Liu, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261–318, 2020.
- [8] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [9] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection," in *Proc. NDSS*, 2018.
- [10] J. U. A. M. C. L. D. L. P. E. Raff et al., "Malware detection by eating a whole EXE," in *Proc. 31st AAAI Conference on Artificial Intelligence*, 2017.