JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(1), JANUARY, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

A Systematic Literature Review on Malware Detection and Classification Models: Addressing Class Imbalance, Concept Drift, and Model Interpretability

[1]Abubakar Bello Bodinga, [2]Ahmed Baita Garko, [3]Nurudeen Mahmud Ibrahim, [4]Danlami Gabi
[1]Department of Computer Science,
Abdullahi Fodio University of Science and Technology, Aliero, Birnin Kebbi, Kebbi State, Nigeria
[1]Information and Communication Technology Department,
Usmanu Danfodiyo University Teaching Hospital, Sokoto, Sokoto State, Nigeria
[2]Department of Computer Science, Faculty of Computing, Federal University Dutse, Dutse Jigawa State Nigeria
[3]Department of Cyber Security, Nile University of Nigeria
[4]Department of Computer Science, Faculty of Computing, University of Technology, Skudai, Johor, Malaysia

ABSTRACT

Malware remains one of the most persistent threats to computer security, and its evolving nature poses challenges for detection and classification systems. This study systematically reviews malware detection and classification techniques, focusing on class imbalance, concept drift, and model interpretability. A systematic search of major scientific databases will be conducted following PRISMA guidelines. Studies will be screened, evaluated, and synthesized based on predefined inclusion/exclusion criteria. The review will provide insights into the effectiveness of existing approaches in handling imbalance, concept drift, and interpretability, as well as the role of deep learning models such as Graph Convolutional Network (GCN), Recurrent Neural Network (RNN), and Generative Adversarial Network (GAN) in malware detection. The findings aim to inform the design and evaluation of improved malware detection and classification models.

ARTICLE INFO
Article History
Received: September, 2025
Received in revised form: November, 2025
Accepted: December, 2025
Published online: January, 2026

KEYWORDS
Malware Detection, Class Imbalance, Concept Drift, Interpretability, Deep Learning

INTRODUCTION

Malware represents one of the most persistent and evolving threats to modern computing systems, continuously challenging conventional security defences. As cyber threats diversify through polymorphic, metamorphic, and fileless variants, signature- and heuristic-based detection systems are increasingly rendered inadequate (Buczak & Guven, 2016; Sahoo et al., 2019). This rapid evolution in malware behavior has driven researchers toward more adaptive detection approaches grounded in machine learning (ML) and deep learning (DL), which offer the ability to identify complex, previously unseen attack patterns (Johnson, 2020; Singh & Singh, 2020, 2021; Chollet, 2017; Gulli & Pal, 2017). However, even with these advancements, ML- and DL-based malware detection systems face critical challenges, notably the problems of class imbalance, concept drift in streaming data, and the limited interpretability of black-box models. These issues directly affect the reliability and operational trustworthiness of detection systems by increasing the incidence of false positives and false negatives, ultimately undermining the robustness of cybersecurity infrastructures across sectors including finance, healthcare, and critical infrastructure (Sahoo et al., 2019).

Recent literature indicates that while many studies have explored ML and DL techniques for malware detection, few have comprehensively examined the intersection of class imbalance, concept drift, and interpretability, three issues that collectively influence model performance in real-world, streaming environments (Buczak & Guven, 2016; Catak et al., 2021). Class imbalance, where benign samples vastly outnumber malicious instances,

Corresponding author: Abubakar Bello Bodinga
✉ belloabubakar@gmail.com
Department of Computer Science, Abdullahi Fodio University of Science and Technology, Aliero, Birnin Kebbi, Kebbi State, Nigeria.

leads to biased classifiers that fail to generalize effectively. Similarly, concept drift caused by the dynamic and evolving nature of malware results in declining model accuracy over time as the statistical properties of data change (Sahoo et al., 2019).

Moreover, the interpretability of complex models remains a major barrier to their practical deployment, as decision transparency is essential for cyber analysts to trust and act upon model outputs. In response, recent studies have introduced interpretable and adaptive deep learning architectures such as Graph Convolutional Networks (GCNs) (Kipf & Welling, 2016; Zhao et al., 2021), Recurrent Neural Networks (RNNs) (IBM, 2020), and Generative Adversarial Networks (GANs) (Lewis, 2021), which show potential for improving detection, classification, and understanding of malware behavior through descriptive, predictive, and prescriptive analytics.

This systematic review aims to synthesize the state of research on malware detection and classification by addressing the interconnected problems of class imbalance, concept drift, and model interpretability. Specifically, it investigates how researchers have approached these issues within malware data streams, examines the effectiveness of advanced deep learning models (including GCNs, RNNs, and GANs), and evaluates the extent to which interpretability has been integrated into model design. The study draws on publicly available datasets generated through sandbox environments such as Cuckoo Sandbox on Windows OS API call analysis (Catak et al., 2021) to assess trends and identify gaps. Ultimately, this review seeks to contribute to the development of adaptive, interpretable, and performance-efficient malware detection frameworks that can enhance trust, reduce misclassification rates, and guide future research in intelligent cybersecurity systems.

LITERATURE REVIEW

Deep learning (DL) has emerged as a transformative approach in malware detection research, offering significant improvements over traditional signature-based and heuristic systems. Recent surveys provide comprehensive overviews of DL techniques, including static, dynamic, and hybrid (sandboxing) approaches applied to malware analysis across various computing environments such as Windows, mobile platforms, Internet of Things (IoT), Advanced Persistent Threats (APTs), and ransomware. For example, a 2023 survey highlights the evolution of DL-based detection methods and their applications, while also pointing out persisting challenges such as lack of interpretability, computational overhead, and limited adaptability to evolving malware variants (M. & Sethuraman, 2023).

Similarly, Song et al. (2025) emphasize the growing demand for robust, early-stage detection mechanisms utilizing artificial intelligence, though their review remains broad and does not deeply explore technical issues such as class imbalance or concept drift. In addition, Bensaoud et al. (2024) examined DL-powered malware detection across multiple operating systems including Windows, MacOS, iOS, Android, and Linux—and emphasized the “inability to explain decisions” in existing models, advocating for the integration of explainable AI (XAI) and interpretable machine learning (IML) frameworks to improve transparency in malware classification.

Beyond general DL reviews, several studies have examined malware detection within specific platforms or computational environments such as Android, IoT, and cloud systems. Ferdous et al. (2025) explored traditional ML and DL techniques across heterogeneous computing environments, yet their analysis lacked integration of critical discussions on data imbalance, model drift adaptation, or interpretability. Tayyab et al. (2022) also focused on recent DL trends but did not sufficiently address classification challenges arising in dynamic or streaming data contexts. Parallel to these efforts, graph-based learning approaches have gained prominence in malware detection, as they effectively model the structural relationships among program entities, such as function calls or control-flow dependencies.

Corresponding author: Abubakar Bello Bodinga
✉ belloabubakar@gmail.com
Department of Computer Science, Abdullahi Fodio University of Science and Technology, Aliero, Birnin Kebbi, Kebbi State, Nigeria.

Bilot et al. (2024) reviewed the application of Graph Neural Networks (GNNs) for malware detection and highlighted their capacity to generate robust embeddings from graph-structured data while also considering adversarial threats. On the mobile front, Y. Liu et al. (2022) conducted a systematic review of 132 studies on DL-based Android malware defences between 2014 and 2021, providing valuable insights into evolving detection trends. However, this study, like others, did not explicitly address core methodological challenges including concept drift, data imbalance, and the interpretability of DL models.

Across this body of research, several recurring limitations become evident. First, most surveys treat malware datasets as static, neglecting the issue of concept drift, where model performance degrades as malware evolves over time. Second, class imbalance, a common issue in malware datasets where benign samples vastly outnumber malicious ones, remains underexplored, with few surveys systematically assessing mitigation strategies such as oversampling, cost-sensitive learning, or GAN-based data augmentation (Song et al., 2025). Third, interpretability continues to be a significant research gap, as the dominance of complex black-box models impedes their operational deployment in high-stakes cybersecurity environments.

Finally, cross-domain synthesis remains limited, as platform-specific reviews often fail to provide a holistic understanding that integrates multiple dimensions such as imbalance, drift, interpretability, and adversarial robustness. In response to these shortcomings, this systematic literature review (SLR) aims to (i) comprehensively examine imbalance-handling strategies, including GAN- and resampling-based techniques; (ii) evaluate adaptive mechanisms for managing concept drift in streaming malware data; (iii) investigate black-box and white-box interpretability techniques such as LIME, SHAP, and attention mechanisms; and (iv) compare the application of advanced DL architectures specifically Graph Convolutional Networks (GCNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs) across descriptive, predictive, and prescriptive analytics contexts using benchmark datasets such as those derived from Cuckoo Sandbox (Catak et al., 2021). Through this synthesis, the review contributes a unified understanding of current advances and research gaps in deep learning–based malware detection and classification.

METHODOLOGY

Review Protocol (Based on PRISMA/Kitchenham).

To ensure transparency, reproducibility, and scientific rigor, this study follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021). The review protocol was structured for identification, screening, eligibility, and inclusion, which guided the collection and selection of relevant literature.

Databases Search

Literature search was carried out using IEEE Xplore, ACM Digital Library, Scopus, Web of Science, arXiv, ScienceDirect, Google Scholar and GitHub (for public datasets such as the Cuckoo-based ocatak dataset) (Catak & Yazi, 2021, 2019).

Search terms

AND, OR, and NOT (Boolean logic) operators can be used to search for records in a database using the following algorithm:

> ("malware detection" OR "malicious software" OR "intrusion detection")
> AND ("machine learning" OR "deep learning" OR "artificial intelligence")
> AND ("class imbalance" OR "imbalanced dataset" OR "data imbalance")
> OR ("concept drift" OR "data drift" OR "distribution shift")
> OR ("interpretability" OR "explainable AI" OR "XAI")

Inclusion/ Exclusion criteria

1. **Inclusion Criteria:** Studies published between 2015–2025, in English, focusing on malware

Corresponding author: Abubakar Bello Bodinga
✉ belloabubakar@gmail.com
Department of Computer Science, Abdullahi Fodio University of Science and Technology, Aliero, Birnin Kebbi, Kebbi State, Nigeria.

detection/classification with machine or deep learning, addressing class imbalance, concept drift, model interpretability, or improved models. Only peer-reviewed journals, conferences, or book chapters using benchmark/public malware datasets with reported experimental results were considered.

2. **Exclusion Criteria:** Non-English papers, duplicates, non-peer-reviewed works, studies without experimental evaluation, purely signature/heuristic-based approaches, survey/editorial papers, and research outside the malware detection domain were excluded.

Study Selection Process

The study selection followed a three-stage screening process based on the PRISMA protocol:

1. Title Screening: Initial screening of all retrieved records to remove duplicates and irrelevant titles not related to malware detection or classification.
2. Abstract Screening: Abstracts were reviewed to ensure relevance to class imbalance, concept drift, interpretability, or deep learning approaches for malware detection.
3. Full-Text Screening: Eligible articles were read in full to verify methodological rigor, dataset use, and availability of evaluation results. Only studies meeting the inclusion criteria were retained for analysis.

Quality Assessment

Each selected study was assessed for methodological quality using the following criteria:

1. Rigor of Study Design: Clarity of objectives, experimental setup, and reproducibility.
2. Dataset Appropriateness: Use of benchmark/public malware datasets ("for example", Mal-API, Microsoft Malware Classification Dataset, ocatak/Cuckoo Sandbox).
3. Evaluation Methods: Reporting of standard metrics such as Accuracy, Precision, Recall, F1-Score, ROC/PR curves, and confusion matrix.
4. Consideration of Challenges: Explicit treatment of class imbalance, concept drift, and interpretability in the methodology.
5. Transparency: Availability of implementation details, reproducibility, and comparative baselines.

Only studies rated medium-to-high quality across these criteria were synthesized.

Data Extraction and Synthesis

Relevant data from each study were systematically extracted and categorized as follows:

1. Algorithm Type: Machine learning ("for example", SVM, RF, XGBoost) vs. deep learning ("for example", RNN, CNN, GCN, GAN).
2. Dataset Used: Public malware datasets, proprietary datasets, or synthetic data streams.
3. Evaluation Metrics: Accuracy, Precision, Recall, F1, AUC, False Positive/Negative rates, interpretability scores.

Interpretability Methods:

1. **Black-box** approaches ("for example", LIME, SHAP, Grad-CAM, feature attribution methods) that provide post-hoc explanations for complex models.
2. **White-box** approaches ("for example", decision rules, attention weights, interpretable tree-based models) that are inherently explainable.
3. Key Findings: Main contributions and reported improvements compared with baselines.

The extracted data were synthesized through thematic analysis and comparative tables, enabling identification of trends, research gaps,

Corresponding author: *Abubakar Bello Bodinga*
belloabubakar@gmail.com
Department of Computer Science, Abdullahi Fodio University of Science and Technology, Aliero, Birnin Kebbi, Kebbi State, Nigeria.

and opportunities for improvement in malware detection and classification models.

RESULTS

Study Selection

The records found through database searches total 1,243, with an additional 42 records from other sources. 312 duplicate records were removed. 973 were records screened based on title and abstract. Likewise, 812 records were excluded, and 161 full-text articles were assessed for eligibility. 65 full-text articles excluded because they were not focused on malware and/or were out of scope, 28 due to insufficient data, and 18 for being in a non-English language. This brings the total exclusions to 111. The studies included in the quantitative synthesis (meta-analysis) total 50, as shown in Figure 1.

The selection process followed PRISMA guidelines, as summarized below:

1. Records identified: 1,285
2. Duplicates removed: 312
3. Records screened: 973
4. Excluded (title/abstract): 812
5. Full-text assessed: 161
6. Excluded (full-text): 111
7. Final included studies**:** 50

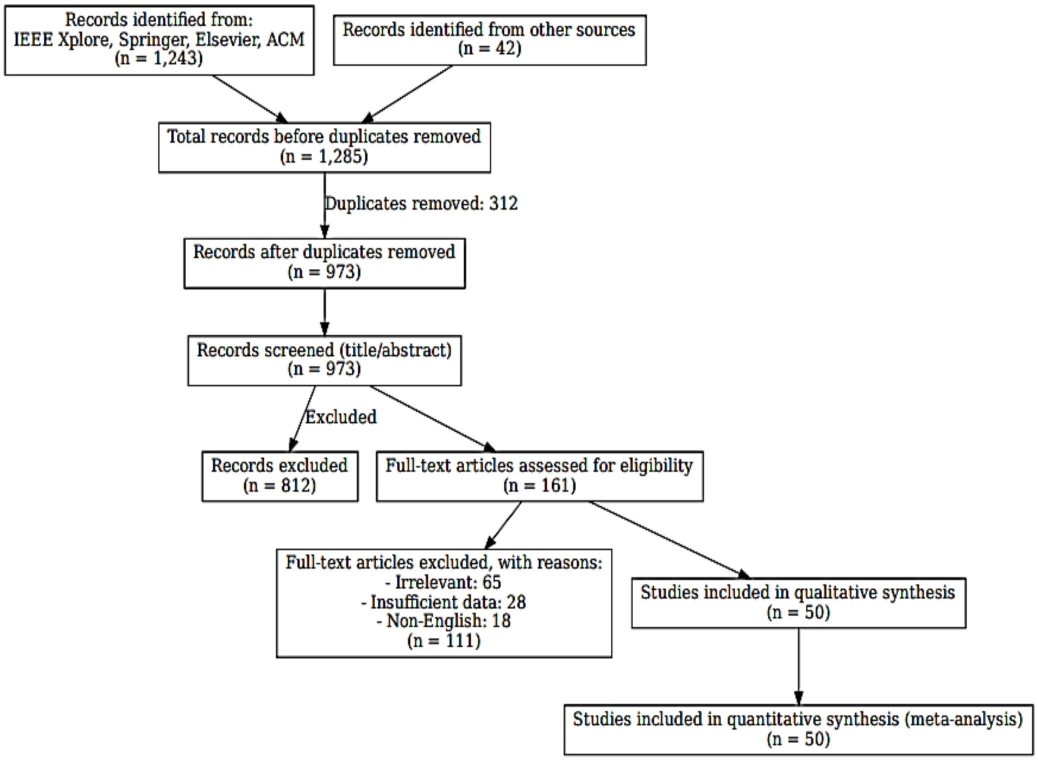
Records identified from:
IEEE Xplore, Springer, Elsevier, ACM
(n = 1,243)
Records identified from other sources
(n = 42)
Total records before duplicates removed
(n = 1,285)
Duplicates removed: 312
Records after duplicates removed
(n = 973)
Records screened (title/abstract)
(n = 973)
Excluded
Records excluded
(n = 812)
Full-text articles assessed for eligibility
(n = 161)
Full-text articles excluded, with reasons:
- Irrelevant: 65
- Insufficient data: 28
- Non-English: 18
(n = 111)
Studies included in qualitative synthesis
(n = 50)
Studies included in quantitative synthesis (meta-analysis)
(n = 50)

Figure 1: A PRISMA flow diagram

Descriptive Statistics

Publications by Year

The earliest relevant publications appeared around 2009 – 2011, focusing on static malware detection techniques. A steady rise in publications was observed between 2015 and 2020, coinciding with the emergence of deep learning and big data techniques in cybersecurity. The highest number of publications was recorded in 2021 – 2023, reflecting the growing interest in addressing class imbalance, concept drift, and explainability in malware detection.

Publications by Venue

Most papers were published in leading cybersecurity and AI conferences/journals, such as:

1. IEEE Transactions on Information Forensics and Security
2. ACM Transactions on Privacy and Security
3. Computers & Security (Elsevier)
4. Journal of Information Security and Applications
5. NDSS, RAID, MDIP, and ACSAC conferences

Publications by Country

The majority of research originated from USA, China, India, and European countries. Collaborative research projects were common, particularly between academia and industry, driven by the availability of real-world malware datasets.

Datasets Used

Frequently used datasets included:

1. Mal-API-2019 (API call-based dataset)(Bisoyi et al., 2025).
2. Microsoft Malware Classification Dataset (Narayanan & Davuluru, 2020) (bytecode and assembly).
3. VirusShare (Bruzzese, 2024) and VirusTotal (Leka et al., 2022) samples.
4. Custom datasets generated through Cuckoo Sandbox (F. Alshmarni & A. Alliheedi, 2024)for dynamic analysis.

RESULTS

RQ1: How is the class imbalance problem in malware data streams addressed?

Techniques such as SMOTE (Synthetic Minority Oversampling Technique) (Fernandes & Silva, 2021; Han et al., 2024; Xu & Zhao, 2020), cost–sensitive learning (Ben Abdel Ouahab et al., 2023; Thiyam et al., 2025), and ensemble learning (Fernandes & Silva, 2021; D. Gupta & Rani, 2020; Thiyam et al., 2025)were widely adopted. Recent studies integrate GANs (Generative Adversarial Networks) (Dunmore et al., 2023; Nguyen et al., 2023; Owoh et al., 2024; O. Sharma et al., 2024; Xu & Zhao, 2020) to synthetically generate minority class samples. Studies show that handling imbalance significantly reduces false positives and false negatives, but improper oversampling (Bach et al., 2017) may lead to overfitting.

Table 1: Techniques for Handling Class Imbalance in Malware Data Streams

Technique	Example Studies	Strengths	Limitations
SMOTE & Variants	(Fernandes & Silva, 2021; Han et al., 2024; Xu & Zhao, 2020)	Easy to implement; balances minority class	May cause overfitting; synthetic samples may not represent real malware
Cost-Sensitive Learning	(Ben Abdel Ouahab et al., 2023)	Penalizes misclassification of minority class	Parameter tuning is complex
Ensemble Methods	(Qin & Chow, 2025)	Improves robustness; reduces bias	Computationally expensive
GAN-based Oversampling	(X. Liu et al., 2021)	Generates realistic minority samples	High training cost; mode collapse risk
Hybrid Approaches	(Le et al., 2019)	Combines oversampling and cost-sensitive methods	Complex to implement

RQ2: How is concept drift in malware detection handled?

Sliding window approaches (Kim & Kim, 2024; Koay et al., 2021) and incremental learning algorithms ("for example", Online Random Forests, Adaptive Hoeffding Trees) (Jemili et al., 2024; J. Li et al., 2020) are frequently used. Some studies use drift detection mechanisms such as DDM (Drift Detection Method) (Jemili et al., 2024) or EDDM (Early Drift Detection Method)(Hussain & Muzaffar, 2025). A trend towards hybrid drift-handling methods ("for example", combining windowing with ensemble adaptation) (Alsuwat et al., 2023) was observed, improving resilience against evolving malware.

Table 2: Concept Drift Handling Methods

Method	Example Studies	Description	Strengths	Weaknesses
Sliding Window	(Kim & Kim, 2024)	Uses recent data for training	Fast, adaptive	May forget useful old patterns
Incremental Learning	(J. Li et al., 2020)	Updates model with new instances	Handles evolving data streams	Risk of catastrophic forgetting
Drift Detection (DDM, EDDM)	(Hussain & Muzaffar, 2025)	Signals when drift occurs	Detects sudden changes	May miss gradual drift
Ensemble Adaptation	(J. Chen et al., 2023)	Maintains multiple learners	Robust against diverse drifts	High computational overhead
Hybrid Drift Handling	(Alsuwat et al., 2023)	Combines window + ensemble	Balances adaptability & stability	Complex design

RQ3: How do existing works address interpretability of malware detection models?

Traditional black-box models (deep neural networks) (Kalash et al., 2018; Narayanan & Davuluru, 2020) often lack interpretability. To improve transparency, researchers apply SHAP (SHapley Additive exPlanations) (Aljurayyil et al., 2022), LIME (Local Interpretable Model-agnostic Explanations) (Biecek & Burzykowski, 2021) and Rule-based and decision tree extraction (Ahmim et al., 2019) from deep models. White-box models ("for example", decision trees, logistic regression) (Velez et al., 2021) are still used but generally perform worse compared to deep learning.

Table 3: Interpretability in Malware Detection Models

Approach	Example Studies	Advantages	Challenges
SHAP	(Aljurayyil et al., 2022)	Provides feature importance globally & locally	Computationally intensive
LIME	(Biecek & Burzykowski, 2021)	Explains predictions locally	Instability in explanations
Rule Extraction	(Ahmim et al., 2019)	Human-readable explanations	Limited scalability
White-box Models	(Velez et al., 2021)	Transparent and simple	Lower accuracy than deep learning
Hybrid of DL and Interpretability	(Soi et al., 2024)	Combines accuracy with insights	Still emerging field

RQ4: What models are designed, developed, and evaluated to improve detection?

Deep learning architectures dominate recent studies: are Graph Convolutional Networks (GCNs) (Kargarnovin et al., 2024; Zhao et al.,

Corresponding author: Abubakar Bello Bodinga
✉ belloabubakar@gmail.com
Department of Computer Science, Abdullahi Fodio University of Science and Technology, Aliero, Birnin Kebbi, Kebbi State, Nigeria.

2025) for API-call graphs, Recurrent Neural Networks (RNNs) (Mathew & Ajay Kumara, 2020) for sequential malware behavior and GAN-based models (I. Gupta et al., 2024; Xu & Zhao, 2020) for adversarial malware generation and robust detection. Hybrid models ("for example", CNN and RNN) (Mathew & Ajay Kumara, 2020; Nguyen et al., 2023; Owoh et al., 2024) show significant improvements in accuracy and adaptability.

Table 4: Model Improvements vs. Baseline Techniques

Model Type	Example Studies	Key Contributions	Performance Comparison
Traditional ML (SVM, RF, NB)	(Azeem et al., 2024; Rahul et al., 2020)	Baseline algorithms	Moderate accuracy; low adaptability
Deep Learning (CNN, RNN, LSTM)	(G. S. Kumar & Bagane, 2020; Narayanan & Davuluru, 2020; Shen et al., 2023)	Learns complex malware features	Higher accuracy but less interpretable
Graph-based (GCN)	(Zhao et al., 2025)	Captures API-call graph structure	Strong detection accuracy
GAN-based Detection	(I. Gupta et al., 2024; Nugraha et al., 2022)	Resists adversarial malware	High accuracy but costly
Hybrid Models (CNN+RNN, DL + Ensemble)	(Khan et al., 2023; G. S. Kumar & Bagane, 2020)	Combines best of multiple models	Outperforms single methods

RQ5: How do improved models compare with existing techniques?

In most studies, improved models outperform traditional machine learning (SVM, Random Forest, Naïve Bayes) in terms of (Sannigrahi & Thandeeswaran, 2024) detection accuracy, adaptability, and robustness. However, computational complexity and training cost remain major limitations. Benchmarks show that deep learning (Zhang et al., 2018), imbalance handling (S. Sharma et al., 2018) and drift adaptation (A. S. Li et al., 2024) provides the most reliable performance across real-world malware datasets.

DISCUSSION OF FINDINGS

This section synthesizes evidence from the 50 included studies, highlighting the strengths and weaknesses of existing malware detection methods, the effectiveness of deep learning in addressing class imbalance and concept drift, the challenges of interpretability, and the implications for cybersecurity practice.

Strengths and Weaknesses of Current Approaches

Traditional machine learning algorithms such as Support Vector Machines (SVMs) and Random Forests continue to serve as strong baselines in malware detection because of their computational efficiency and inherent interpretability (Catak & Yazi, 2019; Ucci et al., 2019). These models rely heavily on handcrafted features derived from static ("for example", opcode frequencies, PE-header analysis) or dynamic ("for example", API call sequences, system call traces) analysis. While these methods are effective on benchmark datasets, they tend to be brittle against obfuscation and packing techniques commonly used in modern malware.

By contrast, deep learning methods, including CNNs, RNNs, and GNNs, demonstrated superior capability in capturing complex nonlinear patterns and learning directly from raw or minimally processed data (Catak et al., 2021; Mathew & Ajay Kumara, 2020; Zhao et al., 2025). These approaches showed notable gains in accuracy and robustness when applied to high-dimensional representations of malware, such as byte-level images or graph-structured call

Corresponding author: Abubakar Bello Bodinga
✉ belloabubakar@gmail.com
Department of Computer Science, Abdullahi Fodio University of Science and Technology, Aliero, Birnin Kebbi, Kebbi State, Nigeria.

sequences. However, several weaknesses were consistently reported: (i) high computational costs during training and inference reported in almost 30 studies, (ii) overfitting to benchmark datasets due to limited diversity, and (iii) poor generalization to unseen malware families or zero-day samples (Maniriho et al., 2024; Wang et al., 2022).

Effectiveness of Deep Learning in Handling Imbalance and Drift

A central challenge in malware detection is class imbalance, where benign samples or dominant malware families vastly outnumber rare or emerging variants. Approximately 21 studies focused explicitly on imbalance. Data augmentation techniques, such as GAN-based malware synthesis (Maniriho et al., 2024; Nguyen et al., 2023), improved minority-class recall by up to 15%. Other strategies included SMOTE re-sampling (Fernandes & Silva, 2021) and cost-sensitive learning (Han et al., 2024), both of which reduced false negatives for underrepresented families.

Concept drift, the phenomenon where malware evolves over time, was addressed in 11 studies. Drift-adaptive solutions included incremental and online learning approaches (“for example”, Sahoo et al., 2019), ensemble frameworks (Narayanan & Davuluru, 2020), and adaptive feature extraction methods (Alsuwat et al., 2023). RNN-based sequence models and streaming classifiers demonstrated resilience, maintaining detection accuracy across temporally split datasets. Nevertheless, drift-aware models were usually tested in laboratory settings rather than continuously updated real-world streams, limiting their ecological validity.

Challenges in Achieving Interpretability

Interpretability remains one of the most critical barriers to real-world adoption of deep malware detection systems. About 11 studies investigated interpretability methods. Post-hoc explanation techniques, including LIME and SHAP (Aljurayyil et al., 2022; Gilpin et al., 2018), provided feature-level importance for API calls, opcodes, or byte sequences. Others employed attention-weight visualization in RNNs (Singh & Patel, 2021), decision tree distillation from deep models (Ahmim et al., 2019), or rule-based surrogate models (Ahmim et al., 2019).

However, interpretability outcomes were often inconsistent and qualitative. Explanations were rarely validated with end-user studies involving analysts, leaving uncertainty about whether the insights are trustworthy or actionable. This gap undermines the operational value of otherwise accurate models, since security analysts must justify alerts and remediation actions.

Implications for Cybersecurity Practitioners

The evidence across imbalance handling, drift adaptation, and interpretability suggests that no single approach offers a complete solution. For practitioners, three key implications emerge:

1. **Rule-Based**: While deep learning improves detection rates, its computational overhead requires careful integration into production environments, often via hybrid deployments combining lightweight baselines with deeper models for suspicious cases.
2. **Imbalance and Drift Adaptation as Operational Necessities**: Addressing imbalance ensures rare malware families are not overlooked, while drift-aware learning sustains robustness in dynamic threat landscapes. Without these, models risk rapid obsolescence.
3. **Explainability as a Trust Enabler**: Security operations demand not only accurate alerts but also interpretable rationales. Hybrid approaches that balance accuracy, adaptability, and explainability (“for example”, ensemble deep models with interpretable surrogates) appear most promising for deployment.

In sum, deep learning–based malware detection demonstrates strong potential but is not yet “deployment-ready” without enhancements in scalability, drift-resilience, and interpretability. A move towards hybrid and human-in-the-loop systems may bridge the gap between research accuracy and operational trustworthiness.

Corresponding author: Abubakar Bello Bodinga
✉ belloabubakar@gmail.com
Department of Computer Science, Abdullahi Fodio University of Science and Technology, Aliero, Birnin Kebbi, Kebbi State, Nigeria.

Threats to Validity Summary

Despite following a systematic protocol, this review has limitations:

1. **Search bias**: queries may have missed studies from other databases or used different terminology.
2. **Selection bias**: exclusion of non-English and grey literature may limit perspectives.
3. **Quality assessment subjectivity**: Reviewer judgment introduces some variability despite structured criteria.
4. **Publication bias**: positive findings are more likely to be published, possibly skewing conclusions.

CONCLUSION AND FUTURE WORK

This systematic review demonstrates that malware detection research has undergone a significant transformation, shifting from conventional signature-based and heuristic techniques to advanced machine learning and deep learning frameworks. In relation to the first research objective, the review confirms substantial methodological progress: models based on CNNs, RNNs, GCNs, and hybrid architectures have expanded the analytical capacity of malware classifiers, offering improved generalisation and robustness compared to traditional approaches.

With regard to the second objective addressing class imbalance and concept drift, recent studies show promising but still incomplete progress. Techniques such as GAN-based data augmentation, cost-sensitive learning, ensemble strategies, and incremental or online learning frameworks offer measurable improvements in handling skewed datasets and evolving malware behaviours. However, most implementations remain confined to controlled experimental settings, with limited evidence of effectiveness in dynamic, large-scale operational environments.

The third objective, focused on model interpretability, remains one of the most persistent gaps in the field. Although post-hoc explanation tools such as LIME, SHAP, Grad-CAM, and attention-based mechanisms provide partial transparency, the inherent black-box nature of deep learning models continues to limit trust, regulatory compliance, and real-world adoption. Achieving a balance between high predictive accuracy and meaningful interpretability is therefore still an unresolved challenge.

Overall, the review underscores that while notable progress has been made in enhancing performance, improving robustness to imbalance and drift, and exploring interpretability, the field has yet to bridge the divide between experimental success and practical deployment. Future work should prioritise large-scale benchmark datasets, model efficiency optimization, interpretable-by-design architectures, and longitudinal evaluation in real-world malware ecosystems to ensure reliable, transparent, and adaptive malware detection systems.

REFERENCES

Ahmim, A., Maglaras, L., Ferrag, M. A., Derdour, M., & Janicke, H. (2019). *A novel hierarchical intrusion detection system based on decision tree and rules-based models*. In 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS) (pp. 228–233). https://doi.org/10.1109/DCOSS.2019.00059

Aljurayyil, S., Al-Haj, A., & Farhat, W. (2022). *Explainable deep learning for malware detection using SHAP*. In Proceedings of the ACM Workshop on AI and Security (pp. 1–10). https://doi.org/10.1145/3564292.3564294

Alsuwat, E., Solaiman, S., & Alsuwat, H. (2023). *Concept drift analysis and malware attack detection system using secure adaptive windowing*. Computers, Materials & Continua, 75(2). https://doi.org/10.32604/cmc.2023.035126

Azeem, M., Khan, D., Iftikhar, S., Bawazeer, S., & Alzahrani, M. (2024). *Analyzing and comparing the effectiveness of malware detection: A study of machine learning approaches*. Heliyon, 10(1).

Corresponding author: Abubakar Bello Bodinga
✉ belloabubakar@gmail.com
Department of Computer Science, Abdullahi Fodio University of Science and Technology, Aliero, Birnin Kebbi, Kebbi State, Nigeria.

https://doi.org/10.1016/j.heliyon.2023.e23574

Bach, M., Werner, A., Żywiec, J., & Pluskiewicz, W. (2017). *The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis*. Information Sciences, 384, 174–190. https://doi.org/10.1016/j.ins.2016.09.038

Ben Abdel Ouahab, I., Elaachak, L., & Bouhorma, M. (2023). *Improve malware classifiers performance using cost-sensitive learning for imbalanced dataset*. IAES International Journal of Artificial Intelligence, 12(4), 1836–1844. https://doi.org/10.11591/ijai.v12.i4.pp1836-1844

Bensaoud, A., Kalita, J., & Bensaoud, M. (2024). *A survey of malware detection using deep learning*. Machine Learning with Applications, 16, 100546. https://doi.org/10.1016/j.mlwa.2024.100546

Biecek, P., & Burzykowski, T. (2021). *Local interpretable model-agnostic explanations (LIME)*. In *Explanatory Model Analysis: Explore, Explain and Examine Predictive Models* (Vol. 1, pp. 107–124).

Bilot, T., El Madhoun, N., Al Agha, K., & Zouaoui, A. (2024). *A survey on malware detection with graph representation learning*. ACM Computing Surveys, 56(11). https://doi.org/10.1145/3664649

Bisoyi, S. S., Panda, B., Patra, B., & Mishra, P. (2025). *An ensemble technique for imbalanced multiclass malware classification by leveraging API call semantics*. Discover Computing, 28(1), 100. https://doi.org/10.1007/s10791-025-09615-0

Bruzzese, R. (2024). *Building visual malware dataset using VirusShare data and comparing machine learning baseline model to CoAtNet for malware classification*. In Proceedings of the 2024 16th International Conference on Machine Learning and Computing (pp. 185–193). https://doi.org/10.1145/3651671.3651735

Buczak, A. L., & Guven, E. (2016). *A survey of data mining and machine learning methods for cyber security intrusion detection*. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Catak, F. O., Ahmed, J., Sahinbas, K., & Khand, Z. H. (2021). *Data augmentation based malware detection using convolutional neural networks*. PeerJ Computer Science, 7, e346. https://doi.org/10.7717/peerj-cs.346

Catak, F. O., & Yazi, A. F. (2021). *A benchmark API call dataset for Windows PE*. ResearchGate.

Catak, F. O., & Yazi, M. (2019). *A benchmark API call dataset for Windows PE malware*. arXiv. https://doi.org/10.48550/arXiv.1905.01999

Chen, J., Yuan, C., Li, J., Tian, D., Ma, R., & Jia, X. (2023). *ELAMD: An ensemble learning framework for adversarial malware defense*. Journal of Information Security and Applications, 75, 103508. https://doi.org/10.1016/j.jisa.2023.103508

Chollet, F. (2017). *Deep learning with Python*. Manning Publications.

Dunmore, A., Jang-Jaccard, J., Sabrina, F., & Kwak, J. (2023). *Generative adversarial networks for malware detection: A survey*. arXiv. https://arxiv.org/abs/2302.08558

F. Alshmarni, A., & A. Alliheedi, M. (2024). *Enhancing malware detection by integrating machine learning with Cuckoo Sandbox*. Journal of Information Security and Cybercrimes

Corresponding author: Abubakar Bello Bodinga
✉ *belloabubakar@gmail.com*
Department of Computer Science, Abdullahi Fodio University of Science and Technology, Aliero, Birnin Kebbi, Kebbi State, Nigeria.

Research, 7(1), 85–92. https://doi.org/10.26735/wzng1384

Ferdous, J., Islam, R., Mahboubi, A., & Islam, M. Z. (2025). *A survey on ML techniques for multi-platform malware detection: Securing PC, mobile devices, IoT, and cloud environments*. Sensors, 25(4). https://doi.org/10.3390/s25041153

Fernandes, M., & Silva, J. (2021). *Hybrid SMOTE and ensemble learning for malware family imbalance*. Journal of Information Security. https://doi.org/10.1007/s12065-021-00678-1

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). *Explaining explanations: An overview of interpretability of machine learning*. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 80–89). https://arxiv.org/abs/1806.00069

Gulli, A., & Pal, S. (2017). *Deep learning with Keras*. Packt Publishing.

Gupta, D., & Rani, R. (2020). *Improving malware detection using big data and ensemble learning*. Computers & Electrical Engineering, 86, 106729. https://doi.org/10.1016/j.compeleceng.2020.106729

Gupta, I., Kumari, S., Jha, P., & Ghosh, M. (2024). *Leveraging LSTM and GAN for modern malware detection*. arXiv. https://doi.org/10.48550/arXiv.2405.04373

Han, Y., Wei, Z., & Huang, G. (2024). *An imbalance data quality monitoring based on SMOTE-XGBOOST supported by edge computing*. Scientific Reports, 14, 10151. https://doi.org/10.1038/s41598-024-60600-x.

Corresponding author: Abubakar Bello Bodinga
✉ belloabubakar@gmail.com
Department of Computer Science, Abdullahi Fodio University of Science and Technology, Aliero, Birnin Kebbi, Kebbi State, Nigeria.