International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Deep Learning-Based Malware Detection: A Systematic Review with Explainable Artificial Intelligence Perspectives

Km. Muskan
Student (M.tech)
Eshan college of engineering, Farah, Mathura, Uttar Pradesh

Pawan Yadav
Assistant Professor, CSE Department.
Eshan college of engineering, Farah, Mathura, Uttar Pradesh

Abstract: The fast development of malware and the growing sophistication of cyber threats have made the older signature-based detection systems incapable of the contemporary cybersecurity systems. The deep learning based methods have become a potent alternative over the past few years because they have the capacity to automatically discover discriminative representations on huge and heterogeneous data sets. Nevertheless, due to high detection rates, deep learning models have been criticized because of their black box decision-making mechanisms, which restricts their deployment in sensitive and regulated settings. The present paper is the systematic and in-depth review of deep-learning malware-detection techniques, specifically the integration of explainable artificial intelligence to make them more transparent and trustworthy. The paper examines malware analysis paradigms, deep learning architectures, benchmark datasets, and evaluation metrics and explainability methods that are widely used in the literature. In addition, important issues in real-world deployment are also analyzed, such as adversarial robustness, scalability, concept drift, and the accuracy versus interpretability trade-off. Through the synthesis of recent trends in research and open research challenges, this review will point out the important directions of creating robust, interpretable, and deployable malware detection systems that is appropriate as a tool in real-world cybersecurity.

Keywords: Cybersecurity, Deep learning, Explainable artificial intelligence, Malware detection

1. INTRODUCTION

The accelerated digital transformation of contemporary society has considerably amplified the magnitude, difficulty, and interconnectedness of computing systems, and by implication, raised vulnerability to cyber threats. As one of these threats, malware, malicious software intended at causing disruption, obtaining sensitive information or accessing it without authorization, has been called one of the most important and significant issues in the field of cybersecurity that has consistently remained a challenge. The modern malware families have sophisticated features, including polymorphism, metamorphism, encryption, and awareness to the environment, which allow them to avoid the traditional security systems. Due to this, malware detection has been getting harder with traditional signature and rule based methods that highly depend on prior knowledge of any known threat and cannot be used effectively against zero-day attacks.

Deep learning (DL), a branch of machine learning, has been developed as a paradigm shift in malware detection because it addresses most of the limitations of conventional machine learning techniques. Deep neural networks are self-taught at learning hierarchical representations of features in raw or minimally processed data, and because of this, do not require any significant amount of hand-crafted feature engineering. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Graph Neural Networks (GNNs), and Transformer-based models have proven to be better in a variety of malware detection tasks. These models have been effectively used on all kinds of data representations such as executable binaries, bytecode images, sequences of API calls, control-flow graphs and network traffic logs (Janiesch, Zschech and Heinrich, 2021).

CNN-based models are especially efficient in the case when binaries of malware are converted to image-like forms, which encode spatial patterns that are associated with malicious behavior. RNNs and LSTMs are good sequence-based models and are useful in modeling time related information in dynamic analysis data, including API call traces and system call sequences. Later on, GNNs were proposed to detect structural association of malware in the form of function call graphs and control-flow graphs to provide more semantic insight. Transformer architectures, which take advantage of self-attention mechanisms, have also advanced the field of malware detection by learning long-range dependencies and make it scalable to large datasets.

Explainable Artificial Intelligence (XAI) is one of the promising technologies that could be used to resolve the transparency issues related to deep learning models. XAI methods are designed to offer human interpretable explanations of model predictions by giving

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

influential features, emphasizing decision paths, or by visualizing the learned representations. In malware detection, explainability allows analysts to reverse-trace predictions into particular behavioral patterns, code fragments or structural parts, which increases the confidence of automated systems. Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), Gradient-weighted Class Activation Mapping (Grad-CAM) and attention visualization are among the techniques that have been rapidly embraced to explain DL-based malware classifiers (Sadeghi R. *et al.*, 2024).

On the basis of these observations, this paper includes a systematic and comprehensive survey of deep learning methods of malware detection with a focus on explainability and deployment issues. The paper summarizes the latest developments in malware analysis methodology, deep learning models, benchmark datasets, evaluation processes, and XAI systems. This work is expected to be of crucial benefit to researchers and other practitioners interested in coming up with accurate, interpretable and deployable malware detection systems through a critical analysis of current research and identification of issues left to be addressed.

2. MALWARE ANALYSIS AND DETECTION PARADIGMS

The analysis and representation of malicious software is vital to the functioning of malware detection systems. Malware analysis is the study of potentially malicious software in order to gain an insight into its structure, behavior, and purpose, and eventually derive some distinctive features that will allow distinguishing them and classifying them properly. Malware analysis methods have been continuously developed over time to keep up with more advanced evasion techniques used by attackers. Malware analysis paradigms in modern research can be generally divided into three categories (static analysis, dynamic analysis, and hybrid analysis), and each has its own benefits and drawbacks.

Dynamic malware analysis fills in part of the weaknesses of the static analysis technique by observing malware behavior in a controlled environment, usually a sandbox or a virtual machine. Dynamic analysis can give a better understanding of the real actions of malicious software by observing runtime activity like system calls, API calls, file system operations, registry changes, and network interactions. Such approach of behavioral perspective helps identify malware which does not seem harmful when it is in the form of malware, but during execution, it performs malicious activities. Dynamic analysis thus is more resistant to obfuscation of code and packing techniques. However, it brings with it enormous difficulties, such as large computing power, increased analysis time, and vulnerability to environment-aware malwares, which are capable of identifying sandboxing conditions and eliminate malicious activities (Ilić *et al.*, 2024).

Deep learning combined with the state-of-the-art malware analysis paradigms have helped to increase the detection capabilities significantly. Nevertheless, the sophistication of modern malware as well as the variety of the methods of analysis present threats in terms of model interpretability, scaling, and robustness. Knowledge of the interaction of various paradigms of analyses with deep learning architecture is thus critical in the development of effective and reliable systems of malware detection. The following section is based on this relationship to discuss taxonomy and architectural.

3. DEEP LEARNING ARCHITECTURES FOR MALWARE DETECTION

The quality of malware detection systems that are trained using deep learning highly depends on the type of neural network architecture and malware data representation. In contrast to other conventional machine learning approaches, which require handcrafted features, deep learning encoders will automatically derive hierarchical abstractions out of raw or less processed inputs. This is an ability that allows them to identify complex and non-linear relationships which are in the behavior of modern malware. In the last ten years, various types of deep learning architectures have been investigated to detect malware, and each is specific to the data modalities and analysis paradigm.

One of the most popular architectures used in the study of malware detection is the convolutional neural networks (CNNs), which is mainly used in the case of the statical analysis of malware. This has been made very successful by the fact that the visualization of malware has become more popular; executable binaries or byte sequences are converted into a two-dimensional representation like a grayscale image. In this sense, CNNs acquire spatial patterns that are associated with malicious behavior and in many cases, they produce high classification accuracy. The approaches based on CNNs are computationally efficient when running inference and can be scaled to big datasets. They are however mostly useful at capturing spatial correlations and might not be able to model temporal dependencies or semantics of execution, thus being less useful against malware with complex behaviour (Yamashita *et al.*, 2018).

Recently, transformer-based architectures have been proposed to malware detection because they allow modeling long-range dependencies with self-attention mechanisms. Transformers operate on input sequences in parallel unlike RNNs, so they are highly suitable in large-scale learning. Opcode sequences, API call sequences and even multi-modal malware representations have been

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

transformed. They are promising candidates of next-generation malware detection systems because of their high flexibility and scalability. However, transformer models are computationally costly and can be resource-heavy (large labeled datasets are frequently necessary to perform optimally), meaning that they may not be applicable in resource-constrained settings.

In general, the selection of the deep learning architecture should be informed by the characteristics of the malware data, the analysis paradigm and deployment limitations. There is no single architecture that is always the best and trade-offs between accuracy, interpretability, scalability, and computational efficiency are unavoidable. These trade-offs are the key to understanding how the malware detection systems can be developed not only to be accurate, but also to be real and reliable. In the following section the emphasis is put on explainable artificial intelligence approaches that set out to respond to the interpretability challenges presented by these deep learning models.

4. EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR MALWARE DETECTION

Although deep learning models have shown tremendous advances in malware detection accuracy, their black box decision processes present serious issues in practice in cybersecurity settings. Deep neural networks are usually black-box models which provide minimal information on the impact of input features on final predictions. This non-transparency in malware detection may undermine the trust of analysts in their work, make it difficult to respond to incidents, and enforce regulatory and organizational accountability demands. In its turn, explainable artificial intelligence (XAI) has become an important direction of research that seeks to better the interpretability and reliability of systems of malware detection that rely on deep learning.

SHapley Additive explanations (SHAP) is another model-agnostic method that bases its application on cooperative game theory. SHAP attributes the importance of features by determining their marginal contribution to a prediction in all conceivable sets of features. SHAP has been applied to malware detection applications to explain both local and global predictions to allow analysts to interpret the significance of features in malware families and datasets. SHAP has better theoretical guarantees and more explanatory power than LIME, at the cost of higher computational complexity, and hence may be more restricted to real-time detectors (Wang *et al.*, 2024).

Explainability methods have also become popular with model-specific methods, specifically convolutional and attention-based models. Gradient-weighted Class Activation Mapping (Grad-CAM) is widely used with CNN-based binarial malware classifiers, which are used on image-like representations of binaries. Grad-CAM produces heatmaps which identify areas of the input image that are most important in the model decision, enabling visual examination of the potentially suspicious segments of code by the analyst. Although these visual explanations are intuitively explained, the quality of the underlying malware visualization is important to their interpretability, and may not directly map to semantic program behaviour (Chen *et al.*, 2019)(Hota, Panja and Nag, 2025).

5. DEPLOYMENT CHALLENGES AND OPEN RESEARCH ISSUES

Although deep learning-based malware detection systems have made a great step forward, their application in the real world in terms of cybersecurity practice is still not addressed. There are numerous models which exhibit amazing results in controlled experimental environments, but find significant drawbacks in the integration into working systems. These are difficulties based on the dynamism of malware, limitations of computational infrastructure, and the trade-offs of accuracy, interpretability, and efficiency.

Resistance to adversarial manipulation is one of the most important issues related to the implementation of deep learning-based malware detection systems. Authors of malware programs are finding more and more ways to detect and take advantage of vulnerabilities in learning-based detectors by generating adversarial samples to avoid being classified. These attacks can be relatively benign binary mutations, API reordering, or behavioral fuzzing, which maintains malicious code but deceives detectors. Despite the proposed solutions to the weakness through adversarial training and data augmentation, research has shown that these methods are usually more complex and of higher computational cost to train. In addition, adaptive property of adversarial attacks provides a challenge in ensuring resilience in the long run.

Another significant worry is scalability, especially the case of the enterprise and cloud settings when it is necessary to analyze millions of files, network events, and execution traces in a day. High-computational and memory complexity deep learning models can impose unacceptable latency, but not be used in real-time or near-real-time detection. Transformer-based models, as well as graph neural networks, are powerful but particularly resource-intensive. Reaching a correct tradeoff between inference efficiency

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

and detection accuracy is still an unresolved research question, especially when it is needed to run on edge devices or resource-limited systems like IoT networks.

Explainable artificial intelligence is also a new technology that makes deployment difficult. Although XAI methods increase transparency and trust, they may add computational complexity that impedes real-time detection. Also, the explanations produced by post-hoc approaches do not necessarily accurately mirror the inner workings of complex models, and hence this can result in false interpretations. Multiple concerns are also emerging that an in-depth description of the process might reveal confidential information about detection logic, allowing adversaries to reverse-engineer models and come up with evasion measures (Sadeghi R. *et al.*, 2024).

Availability, as well as quality of data remain a major problem. Most malware datasets in research are obsolete, imbalanced or their labels are not based on consistent criteria. Access to realistic operational data is also limited by privacy and legal constraints, which limits the generalizability of publicly-available dataset-trained models. These problems make it clear that collective data-sharing models, labeling standards and benchmarking efforts should be realistic.

To conclude, the discrepancy between laboratory implementation and practical application is a significant drawback to deep learning-based malware detection. To solve adversarial robustness, scalability, adaptability, explainability, and data quality in a single solution, a unified approach is important to the next generation of practical and trustworthy detection systems. The final part provides a summary of the lessons learnt during this review and gives the future research directions (Salih *et al.*, 2025).

6. FUTURE RESEARCH DIRECTIONS

Despite the fact that deep learning and explainable artificial intelligence have greatly improved malware detectors, there are still several research gaps that have not been addressed yet. These issues are crucial to creating detection systems that are accurate, as well as robust, interpretable and applicable in the real world.

A prospective avenue is in the usage of naturally interpretable deep learning systems to detect malware. Majority of the current strategies are based on post-hoc explainability mechanisms that are used after training the models, which might not necessarily be consistent with the internal decision mechanisms. Future directions need to concentrate on the explanation-conscious model design, where the interpretability is directly incorporated in the structure as sparse representation, attention constraints, or rule-directed learning. These models can decrease the dependence on the external explanation tools and enhance trust and transparency.

Malware data still remains a limiting area due to the lack of realistic, recent, and varied data on malware. The most important thing that future research must focus on is coming up with standardized benchmark datasets that capture the current malware behavior in various platform, such as Windows, Android, IoT, and cloud. There should also be inclusion of time data and longitudinal assessment which would enhance the realism of the experiment. Federated learning and privacy-preserving data collection and sharing systems have potential solutions that can be considered to meet these limitations of data accessibility and remain confidential.

Other considerations to future malware detection systems include scalability and efficiency. Theoretical studies of lightweight deep learning models, model compression algorithms, and hardware-constrained optimization are fundamental to its implementation in resource-constrained settings. The solutions of edge computing and on-device detection, especially on IoT and mobile platforms, need models capable of working under strong latency and energy constraints without performance trade-offs in detection.

Lastly, the explainability evaluation is to be given more attention. Although there are well-developed predictive performance measures, there are no standardized ways of evaluating the quality, consistency, and utility of explanations. The creation of quantitative and qualitative assessment systems of XAI in malware detection is necessary to compare the methods and future studies.

In general, the future development of malware detection research will need a comprehensive solution including accuracy, robustness, explainability, scalability, and usability. By solving these mutually supporting problems, future systems will be able to resolve the changing threat environment and render adequate cybersecurity defense.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

7. CONCLUSION

The case of malware detection systems based on deep learning with a specific analysis of explainable artificial intelligence as an essential requirement to implement the systems in the real world were discussed in the paper. Since malware is getting more sophisticated and bigger day after day, traditional signature-based detection systems are no longer sufficient. Detection accuracy of deep learning methods has been shown to improve significantly because the methods automatically learn complex representations that use a variety of malware data formats such as binary files, behavioral traces and structural graphs. The convolutional neural networks, recurrent neural networks, graph neural networks and transformer-based models each have brought their own advantages to the detection scene.

The evaluation of malware datasets and its practices showed that the main challenges were still bias of data sets, imbalanced classes, time, and inconsistency in benchmarking. These problems complicate a just comparison between the studies and restrict the extent to which the reported results can be generalized. Also, adversarial, scalability, concept drift, and system integration deployment issues remain barriers to research advances into operational security solutions.

This paper identifies gaps in the existing literature and highlights gaps in knowledge about the importance of a unified perspective of malware detection that is accurate, interpretable, robust, and efficient. The next generation of research should not focus on individual performance gains but should rather develop the concept of integrated and explanation aware detection systems that can adjust to new threats and real-world conditions. Finally, the successful integration of deep learning with explainable artificial intelligence has an impressive future potential in enhancing reliable and stable malware detection systems in current cybersecurity architectures.

REFERENCES:

[1] Chen, B. *et al.* (2019) 'Adversarial examples for cnn-based malware detectors', *IEEE Access* [Preprint]. Available at: https://doi.org/10.1109/ACCESS.2019.2913439.

[2] Hota, A., Panja, S. and Nag, A. (2025) 'Lightweight CNN-based malware image classification for resource-constrained applications', *Innovations in Systems and Software Engineering* [Preprint]. Available at: https://doi.org/10.1007/s11334-022-00461-7.

[3] Ilić, S. *et al.* (2024) 'Going beyond API Calls in Dynamic Malware Analysis: A Novel Dataset', *Electronics (Switzerland)* [Preprint]. Available at: https://doi.org/10.3390/electronics13173553.

[4] Janiesch, C., Zschech, P. and Heinrich, K. (2021) 'Machine learning and deep learning', *Electronic Markets* [Preprint]. Available at: https://doi.org/10.1007/s12525-021-00475-2.

[5] Sadeghi R., K. *et al.* (2024) 'Explainable artificial intelligence and agile decision-making in supply chain cyber resilience', *Decision Support Systems* [Preprint]. Available at: https://doi.org/10.1016/j.dss.2024.114194.

[6] Salih, A.M. *et al.* (2025) 'A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME', *Advanced Intelligent Systems* [Preprint]. Available at: https://doi.org/10.1002/aisy.202400304.

[7] Wang, H. *et al.* (2024) 'Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods', *Journal of Big Data* [Preprint]. Available at: https://doi.org/10.1186/s40537-024-00905-w.

[8] Yamashita, R. *et al.* (2018) 'Convolutional neural networks: an overview and application in radiology', *Insights into Imaging* [Preprint]. Available at: https://doi.org/10.1007/s13244-018-0639-9.