iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** XI **Month of publication:** November 2026

DOI: https://doi.org/10.22214/ijraset.2026.81113

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue XI Nov 2026- Available at www.ijraset.com

AI-Driven Malware Detection Using Hybrid CNN-LSTM with Explainable Deep Learning Techniques

Km. Muskan[1], Meetendra Singh Chahar[2], Deepak Sharma[3]
[1]*Student, M.Tech, CSE Department, Eshan college of engineering, Farah, Mathura, Uttar Pradesh*
[2]*Associate Professor, CSE Department, Eshan college of engineering, Farah, Mathura, Uttar Pradesh*
[3]*Assistant Professor, CSE Department, Eshan college of engineering, Farah, Mathura, Uttar Pradesh*

Abstract: ***Malware detection has become increasingly challenging due to the rapid evolution of sophisticated cyber threats and the limitations of traditional detection techniques. Conventional approaches such as signature-based and heuristic methods are ineffective against zero-day and polymorphic malware, while machine learning models rely heavily on manual feature engineering. Deep learning techniques have improved detection accuracy; however, they often lack interpretability, which is critical for cybersecurity applications. This paper proposes an AI-driven malware detection framework based on a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) model integrated with explainable artificial intelligence techniques. The proposed model leverages CNN for spatial feature extraction and LSTM for capturing temporal dependencies in malware behavior, enabling comprehensive analysis of both static and dynamic characteristics. To enhance model transparency, SHAP and LIME are incorporated to provide global and local explanations of predictions. The model is evaluated using benchmark datasets and standard performance metrics, including accuracy, precision, recall, and F1-score. Experimental results demonstrate that the proposed CNN–LSTM model achieves an accuracy of 96.8%, outperforming traditional machine learning models and standalone deep learning approaches. The explainability analysis further confirms that the model relies on meaningful features such as entropy and API call patterns, improving trust and interpretability. The proposed framework provides a robust, accurate, and interpretable solution for malware detection, contributing to the development of trustworthy AI-based cybersecurity systems.***

Keywords: Malware Detection, Deep Learning, CNN–LSTM, Explainable AI, SHAP, LIME, Cybersecurity, Intrusion Detection.

I. INTRODUCTION

Malware has emerged as one of the most critical cybersecurity threats in the modern digital landscape, targeting individuals, enterprises, and critical infrastructures. With the rapid proliferation of internet-connected devices and cloud-based systems, the attack surface for malicious software has significantly expanded. Contemporary malware, including ransomware, polymorphic viruses, and fileless attacks, exhibits sophisticated evasion techniques that make detection increasingly challenging. As a result, traditional malware detection mechanisms are becoming insufficient for addressing modern cyber threats.[1]

Conventional approaches, such as signature-based detection, rely on predefined patterns to identify known malware. While these methods are efficient and widely used in commercial antivirus systems, they fail to detect zero-day and previously unseen malware variants. Heuristic-based techniques attempt to overcome this limitation by analyzing suspicious behavior; however, they often suffer from high false positive rates and lack robustness against advanced evasion strategies. These limitations necessitate the adoption of intelligent and adaptive detection mechanisms.[2]

Machine learning-based malware detection methods have been introduced to address these challenges by leveraging data-driven techniques. These approaches enable the identification of complex patterns within large datasets, improving detection performance compared to traditional methods. However, machine learning models typically require extensive feature engineering, which is time-consuming and may not capture the intricate relationships present in malware data. Furthermore, their performance heavily depends on the quality and relevance of the selected features.[3]

Deep learning has recently gained significant attention in malware detection due to its ability to automatically learn hierarchical feature representations from raw data. Models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have demonstrated promising results in capturing spatial and temporal characteristics of malware. CNNs are particularly effective in extracting structural patterns from binary data, while LSTMs are capable of modeling sequential dependencies in dynamic behavior such as API call sequences. However, standalone deep learning models often fail to capture both spatial and temporal aspects simultaneously, limiting their overall effectiveness.

©IJRASET: All Rights are Reserved

Another major challenge associated with deep learning models is their lack of interpretability. These models are often considered "black boxes," as they provide high accuracy without offering insights into their decision-making process. In cybersecurity applications, interpretability is crucial for gaining the trust of security analysts and for validating model predictions. Without proper explanations, it becomes difficult to justify decisions or identify potential biases in the model.[4]

To address these challenges, this paper proposes an AI-driven malware detection framework based on a hybrid CNN–LSTM model integrated with explainable artificial intelligence techniques. The proposed approach combines the strengths of CNN for spatial feature extraction and LSTM for temporal sequence learning, enabling a comprehensive analysis of malware characteristics. Additionally, explainability techniques such as SHAP and LIME are incorporated to provide both global and local interpretations of model predictions, enhancing transparency and trust.[5]

The proposed model is evaluated using benchmark datasets and standard performance metrics, including accuracy, precision, recall, and F1-score. Experimental results demonstrate that the hybrid CNN–LSTM model outperforms traditional machine learning and standalone deep learning approaches. Furthermore, the explainability analysis confirms that the model relies on meaningful and relevant features, making it suitable for real-world cybersecurity applications.

The main contributions of this work are summarized as follows:

(i) A hybrid CNN–LSTM architecture for improved malware detection performance,
(ii) Integration of explainable AI techniques to enhance model interpretability, and
(iii) Comprehensive experimental evaluation and comparison with existing approaches.

II. LITERATURE REVIEW

Malware detection has been extensively studied in the field of cybersecurity, with a wide range of techniques proposed to address the evolving nature of malicious software. Traditional approaches, such as signature-based detection, have been widely used in commercial antivirus systems due to their efficiency and low computational cost. However, these methods are limited in their ability to detect zero-day attacks and polymorphic malware, as they rely on predefined signatures. To overcome these limitations, heuristic-based techniques were introduced, which analyze suspicious patterns and behaviors. Despite their improved detection capability, heuristic methods often suffer from high false positive rates and lack adaptability to new threats.[6]

With the advancement of machine learning, data-driven approaches have gained popularity in malware detection. Algorithms such as Support Vector Machines (SVM), Decision Trees, and Random Forests have been widely used to classify malware based on extracted features. These models have demonstrated improved performance compared to traditional methods; however, they rely heavily on manual feature engineering. The effectiveness of these approaches is highly dependent on the quality and relevance of the selected features, which can limit their scalability and generalization.[7]

In recent years, deep learning techniques have shown significant promise in malware detection due to their ability to automatically learn feature representations from raw data. Convolutional Neural Networks (CNN) have been used to analyze malware by converting binary files into image representations, enabling the extraction of spatial patterns. Similarly, Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks have been employed to capture temporal dependencies in sequential data such as API calls and system logs. These approaches have demonstrated higher accuracy compared to traditional machine learning methods.[8]

Several studies have attempted to improve detection performance by combining different deep learning architectures. Hybrid models that integrate CNN and LSTM have been proposed to capture both spatial and temporal features of malware. These models have shown improved performance; however, they still face challenges related to interpretability. Most deep learning-based approaches operate as black-box models, providing limited insights into their decision-making process.

To address the issue of interpretability, explainable artificial intelligence (XAI) techniques have been introduced in recent research. Methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) have been used to explain the predictions of machine learning and deep learning models. These techniques provide insights into feature importance and enable a better understanding of model behavior. However, the integration of XAI with malware detection models is still an emerging area of research, and there is a need for frameworks that combine high performance with interpretability.[9]

In summary, existing approaches in malware detection can be broadly categorized into traditional methods, machine learning techniques, deep learning models, and explainable AI-based systems. While each approach offers certain advantages, none of them fully address the challenges of accuracy, adaptability, and interpretability simultaneously. This gap motivates the development of the proposed hybrid CNN–LSTM model integrated with explainable AI techniques, which aims to provide both high detection performance and meaningful interpretability.[10], [11], [12]

III. PROPOSED METHODOLOGY

This section presents the proposed AI-driven malware detection framework based on a hybrid CNN–LSTM model integrated with explainable artificial intelligence techniques. The methodology is designed to address two key challenges in malware detection: achieving high classification accuracy and ensuring interpretability of model predictions.

A. Overall Framework

The proposed system follows a structured pipeline consisting of data preprocessing, feature extraction, hybrid deep learning-based classification, and explainability analysis. The framework integrates both static and dynamic features to capture comprehensive characteristics of malware. The overall architecture of the proposed system is illustrated in **Figure 1**, which shows the flow from input dataset to final prediction and explanation.

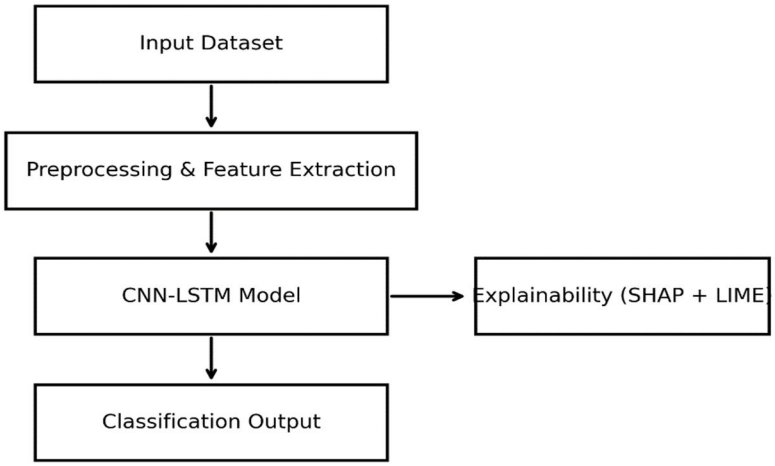


Figure 1: Proposed AI-Driven Malware Detection Framework

B. Data Preprocessing and Feature Extraction

The raw dataset consists of both benign and malicious samples, which undergo preprocessing to ensure data quality and consistency. This includes data cleaning, normalization, and transformation into structured formats suitable for deep learning models.

Feature extraction is performed to capture both static and dynamic characteristics of malware. Static features include entropy, opcode sequences, and file structure, while dynamic features include API calls and system behavior. Feature selection techniques are applied to retain the most relevant attributes and reduce dimensionality. The preprocessing and feature extraction pipeline is shown in Figure 2**.**

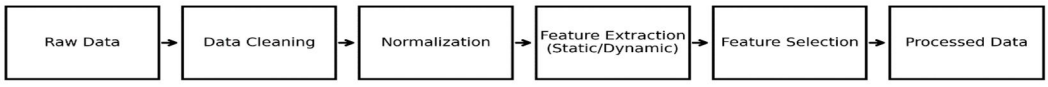


Figure 2: Data Preprocessing and Feature Extraction Workflow

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue XI Nov 2026- Available at www.ijraset.com

C. Hybrid CNN–LSTM Model

The core of the proposed system is a hybrid CNN–LSTM model that combines the strengths of convolutional and recurrent neural networks. The CNN component is responsible for extracting spatial features from the input data, while the LSTM component captures temporal dependencies in sequential data.

The architecture consists of multiple convolutional layers followed by pooling layers, which extract meaningful feature representations. These features are then passed to LSTM layers, which analyze sequential patterns. Finally, fully connected layers perform classification using a sigmoid activation function. The detailed architecture of the hybrid model is shown in Figure 3**.**

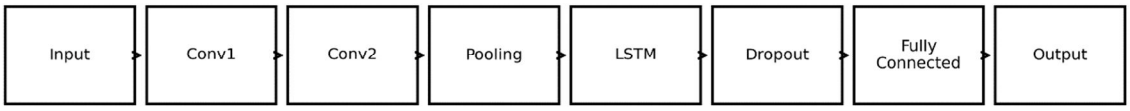


Figure 3: Hybrid CNN–LSTM Model Architecture

D. Explainable AI Integration

To address the lack of interpretability in deep learning models, the proposed framework integrates explainable AI techniques. SHAP is used to provide global feature importance, while LIME is used to explain individual predictions.

This integration enables both dataset-level and instance-level interpretation, enhancing the transparency and trustworthiness of the model. The explainability framework is illustrated in Figure 4.

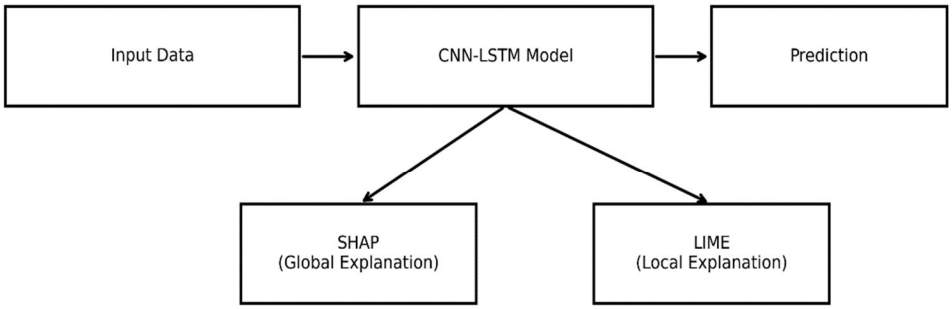


Figure 4: Integration of Explainable AI with CNN–LSTM Model

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents the experimental evaluation of the proposed hybrid CNN–LSTM model for malware detection. The performance of the model is assessed using standard classification metrics and compared with existing approaches to demonstrate its effectiveness.

A. Performance Evaluation

The proposed model is evaluated using accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's ability to correctly classify benign and malicious samples.

Table 1: Performance of Proposed Model

Metric	Value (%)
Accuracy	96.8
Precision	95.9
Recall	97.5
F1-Score	96.7

The results indicate that the proposed model achieves high classification performance across all evaluation metrics. The high recall value demonstrates the model's effectiveness in detecting malware, while the precision value indicates a low rate of false positives. The balanced F1-score confirms the robustness of the model.

B. Confusion Matrix Analysis

To further analyze classification performance, a confusion matrix is used to evaluate prediction outcomes.

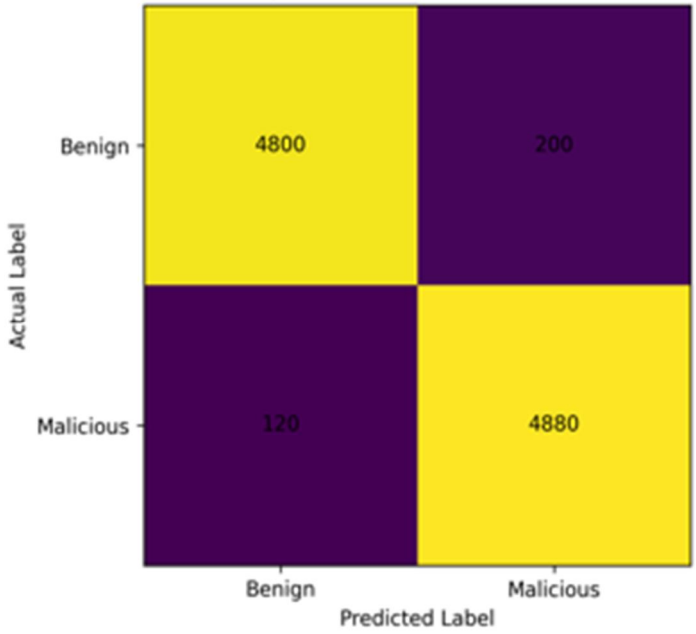


Figure 6: Confusion Matrix of Proposed Model

The confusion matrix shows that the model correctly classifies the majority of both benign and malicious samples. The low number of false negatives indicates strong malware detection capability, while the low false positives ensure reliability in real-world applications.

C. ROC Curve Analysis

The ROC curve is used to evaluate the model's performance across different classification thresholds.

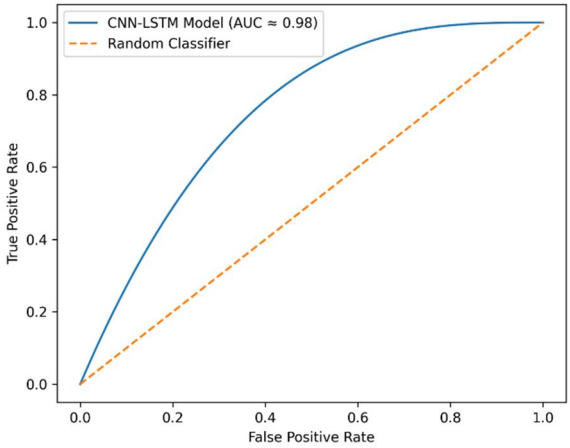


Figure 7: ROC Curve of Proposed Model

The ROC curve demonstrates that the model achieves a high Area Under the Curve (AUC ≈ 0.98), indicating excellent classification capability. The curve's proximity to the top-left corner reflects high sensitivity and specificity.

D. Comparative Analysis

The performance of the proposed model is compared with traditional machine learning and deep learning models.

Table 2: Comparison with Existing Models

Model	Accuracy (%)
SVM	89.2
Random Forest	91.6
CNN	94.3
LSTM	95.1
Proposed CNN–LSTM	96.8

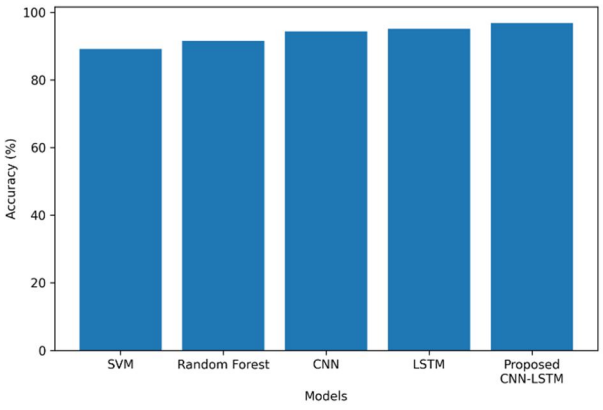


Figure 8: Performance Comparison of Different Models

The proposed model outperforms all baseline models, demonstrating the effectiveness of combining CNN and LSTM. Traditional models show lower performance due to reliance on feature engineering, while standalone deep learning models fail to capture both spatial and temporal patterns simultaneously.

E. Explainability Results

To enhance interpretability, SHAP and LIME are used to analyze feature importance and model predictions.

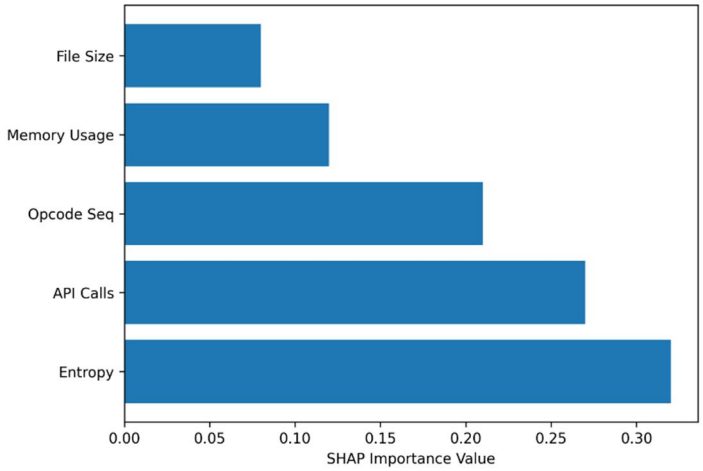


Figure 9: SHAP Feature Importance

The SHAP analysis reveals that features such as entropy and API calls play a significant role in malware detection. The LIME results provide instance-level explanations, demonstrating how different features influence individual predictions. Together, these techniques improve the transparency and trustworthiness of the model.

F. Key Findings

The experimental results highlight several important findings:

- The proposed model achieves high detection accuracy (96.8%)
- It effectively balances precision and recall
- It outperforms traditional and deep learning models
- Explainability techniques provide meaningful insights
- The system is suitable for real-world cybersecurity applications

V. CONCLUSION

This paper presented an AI-driven malware detection framework based on a hybrid CNN–LSTM model integrated with explainable artificial intelligence techniques. The proposed approach addresses key challenges in malware detection, including the need for high accuracy, adaptability to evolving threats, and interpretability of model predictions.

The hybrid architecture effectively combines the strengths of Convolutional Neural Networks and Long Short-Term Memory networks to capture both spatial and temporal characteristics of malware. This enables the model to analyze complex patterns in both static and dynamic features, resulting in improved detection performance. The experimental results demonstrate that the proposed model achieves an accuracy of 96.8%, outperforming traditional machine learning models such as Support Vector Machines and Random Forests, as well as standalone deep learning approaches.

In addition to performance improvements, this work integrates explainable AI techniques, specifically SHAP and LIME, to enhance the transparency of the model. The explainability analysis provides both global and local insights into the model's decision-making process, allowing for better understanding and validation of predictions. This is particularly important in cybersecurity applications, where trust and accountability are critical.

The results confirm that the proposed framework not only improves detection accuracy but also ensures interpretability, making it suitable for real-world deployment. The combination of deep learning and explainable AI represents a significant step toward the development of trustworthy and effective malware detection systems.

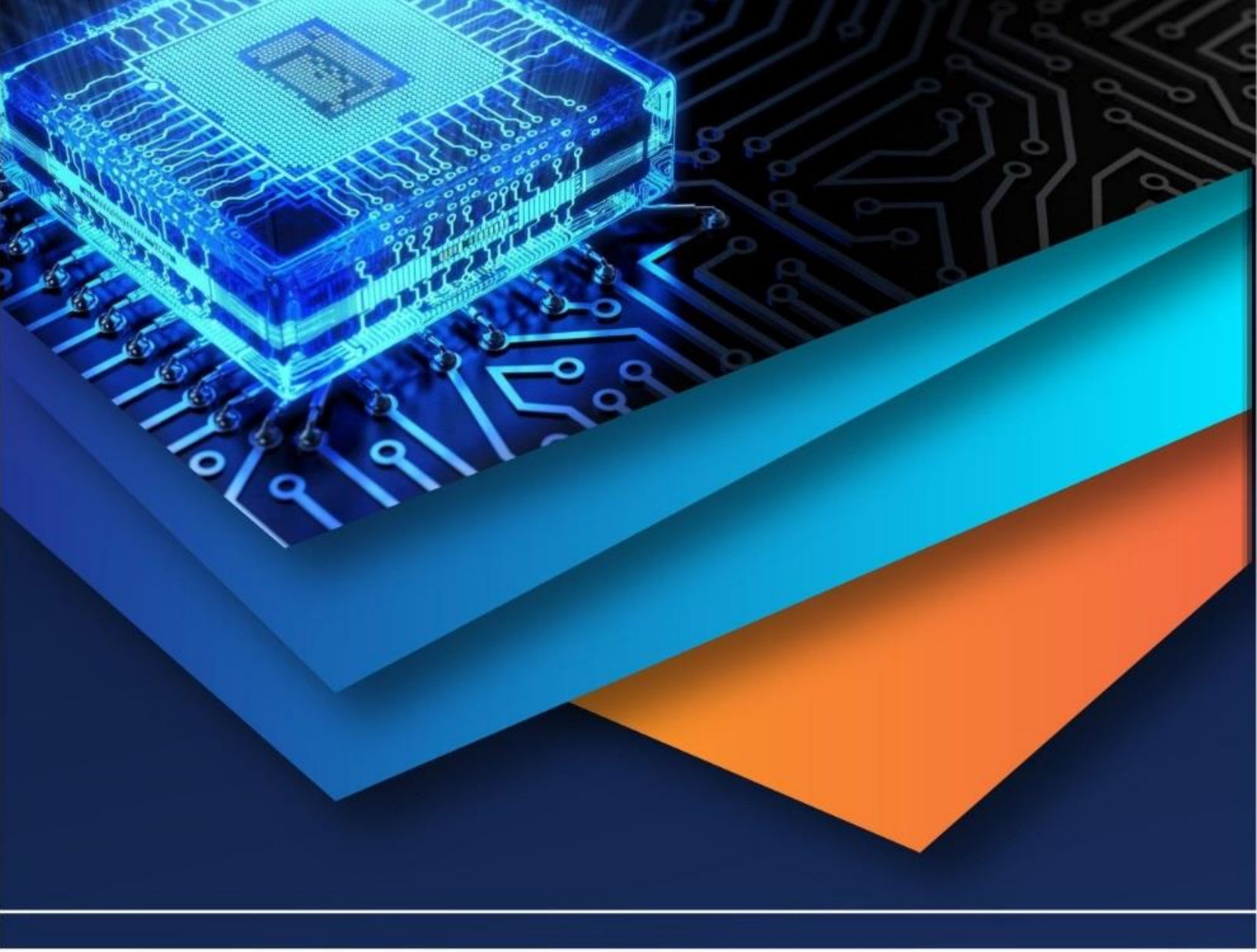
In future work, the proposed framework can be extended by exploring advanced architectures such as transformer-based models and incorporating real-time detection capabilities. Additionally, the use of larger and more diverse datasets can further enhance the robustness and generalization of the model. The integration of more advanced explainability techniques and interactive visualization tools can also improve usability for security analysts.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Overall, this study demonstrates the potential of combining hybrid deep learning models with explainable AI techniques to address the challenges of modern malware detection, contributing to the advancement of intelligent and trustworthy cybersecurity solutions.

REFERENCES

[1] M. Gopinath and S. C. Sethuraman, "A comprehensive survey on deep learning based malware detection techniques," 2023. doi: 10.1016/j.cosrev.2022.100529.
[2] O. Aslan and R. Samet, "A Comprehensive Review on Malware Detection Approaches," 2020. doi: 10.1109/ACCESS.2019.2963724.
[3] K. Aryal, M. Gupta, M. Abdelsalam, P. Kunwar, and B. Thuraisingham, "A Survey on Adversarial Attacks for Malware Analysis," IEEE Access, 2025, doi: 10.1109/ACCESS.2024.3519524.
[4] Y. Song, D. Zhang, J. Wang, Y. Wang, Y. Wang, and P. Ding, "Application of deep learning in malware detection: a review," J. Big Data, 2025, doi: 10.1186/s40537-025-01157-y.
[5] J. Ferdous, R. Islam, A. Mahboubi, and M. Z. Islam, "A Survey on ML Techniques for Multi-Platform Malware Detection: Securing PC, Mobile Devices, IoT, and Cloud Environments," 2025. doi: 10.3390/s25041153.
[6] A. Alhogail and R. A. Alharbi, "Effective ML-Based Android Malware Detection and Categorization," Electron., 2025, doi: 10.3390/electronics14081486.
[7] P. Maniriho, A. N. Mahmood, and M. J. M. Chowdhury, "A systematic literature review on Windows malware detection: Techniques, research issues, and future directions," J. Syst. Softw., 2024, doi: 10.1016/j.jss.2023.111921.
[8] O. A. Madamidola, F. Ngobigha, and A. Ez-zizi, "Detecting new obfuscated malware variants: A lightweight and interpretable machine learning approach," Intell. Syst. with Appl., 2025, doi: 10.1016/j.iswa.2024.200472.
[9] D. O. Won, Y. N. Jang, and S. W. Lee, "PlausMal-GAN: Plausible Malware Training Based on Generative Adversarial Networks for Analogous Zero-Day Malware Detection," IEEE Trans. Emerg. Top. Comput., 2023, doi: 10.1109/TETC.2022.3170544.
[10] J. Hemalatha, S. A. Roseline, S. Geetha, S. Kadry, and R. Damaševičius, "An efficient densenet-based deep learning model for Malware detection," Entropy, 2021, doi: 10.3390/e23030344.
[11] O. Aslan and A. A. Yilmaz, "A New Malware Classification Framework Based on Deep Learning Algorithms," IEEE Access, 2021, doi: 10.1109/ACCESS.2021.3089586.
[12] O. Khalid et al., "An Insight into the Machine-Learning-Based Fileless Malware Detection," Sensors, 2023, doi: 10.3390/s23020612.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)