

EXPLAINX-MALDETECT: ENHANCING MALWARE DETECTION WITH INTERPRETABLE

Mrs. P. Swathi¹, Lohith.M², Brijesh Kumar goud. T³, Harshini.K⁴, Charitharth swamy.M⁵

¹Assistant Professor, Department of IT, TKR College of Engineering and Technology, Telangana, India

^{2,3,4,5}B.Tech Students, Department of IT, TKR College of Engineering and Technology, Telangana, India

Abstract – Explain X-Mal Detect is an AI-driven malware detection system designed to address the limitations of traditional signature-based antivirus tools, which struggle to detect rapidly evolving and zero-day threats. Existing malware detection systems often operate as black-box models, offering high accuracy but no transparency into how decisions are made. This lack of interpretability reduces trust and makes detailed threat analysis difficult for security professionals. Explain X-Mal Detect overcomes these challenges by integrating Machine Learning and Deep Learning models—such as Random Forest, Decision Tree, Logistic Regression, MLP, and a custom deep neural network—with Explainable AI (XAI) methods like SHAP and LIME. The system classifies files or dataset samples as benign or malicious and automatically removes malware-classified files to prevent system compromise. XAI techniques provide clear, feature-level explanations that help analysts understand why a file was flagged, ensuring transparency and informed decision-making. Technically, the system uses Python-based ML workflows, a modular architecture for easy updates, and visualization tools for interpretability. Overall, Explain X-Mal Detect improves detection accuracy, enhances trust through explainability, reduces manual analysis time, and contributes to a more secure and insight-driven malware defense ecosystem.

Key Words: Explainable Artificial Intelligence (XAI), Malware Detection, Zero-Day Attacks, Machine Learning, Deep Learning, SHAP, LIME, Random Forest, Decision Tree, Logistic Regression, Multilayer Perceptron (MLP), Neural Networks, Feature Interpretability, Cybersecurity, AI-Driven Threat Detection, Transparent Security Systems.

1. INTRODUCTION

In today's digital environment, malware has become one of the most persistent and sophisticated threats to computing systems. Traditional signature-based antivirus tools can no longer keep pace with rapidly evolving and obfuscated malware variants. These systems mainly rely on predefined signatures, making them ineffective against newly generated or zero-day threats. As a result, modern cybersecurity requires intelligent and adaptable detection mechanisms capable of identifying unknown threats with high accuracy. Machine Learning (ML) and Deep Learning (DL) models have emerged as powerful alternatives for malware detection due to their ability to learn complex patterns from large datasets.

However, despite their accuracy, these models operate as "black boxes," providing no clarity behind their decisions. This lack of transparency makes it difficult for security analysts to interpret predictions, validate model reliability, or understand system behavior—limiting their practical adoption in cybersecurity workflows. To address this challenge, Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME have gained importance. These frameworks make AI models interpretable by highlighting key features that influence a classification decision. By integrating explainability with ML-based malware detection, it becomes possible to combine accuracy with trust, improving both system reliability and user confidence. Explain X-Mal Detect is a hybrid AI-driven malware detection system designed to classify files as malicious or benign while providing clear interpretability for each prediction. The system utilizes multiple ML/DL models—including Random Forest, Decision Tree, Logistic Regression, MLP, and Deep Learning—to analyse static features of dataset samples. When a file is identified as malware, the system automatically deletes it to prevent potential harm. Through XAI visualizations, users gain insights into why a file was flagged, bridging the gap between high-performance detection and transparency. Overall, Explain X-Mal Detect enhances cyber security by offering accurate detection, human understandable explanations, and automated defensive actions. The system contributes toward building intelligent, interpretable, and practical malware detection solutions suitable for modern threat landscapes.

1.1 Limitations of Traditional Malware Detection Systems

Traditional malware detection systems mainly rely on signature-based techniques, where known malware patterns are stored in a database and matched against incoming files. While effective against previously identified threats, these systems fail when dealing with new, polymorphic, or zero-day malware, which continuously evolve to evade detection. Frequent signature updates are required, and even minor changes in malicious code can bypass traditional defenses. Additionally, conventional systems lack adaptability and struggle to scale against the rapidly growing volume of malware. This results in delayed detection, increased false negatives, and higher vulnerability to sophisticated cyberattacks, making traditional approaches insufficient for modern cybersecurity needs.

1.2 for Explainable AI in Modern Malware Detection

Modern malware detection systems increasingly use Machine Learning and Deep Learning models to improve accuracy and handle complex threat patterns. However, many of these models function as black boxes, providing predictions without explaining the reasoning behind them. This lack of transparency reduces trust among security analysts and makes forensic investigation and decision validation difficult. Explainable Artificial Intelligence (XAI) addresses this challenge by offering human-interpretable explanations for model decisions. Techniques such as SHAP and LIME reveal feature-level contributions, helping analysts understand why a file is classified as malicious or benign. By integrating XAI with AI-driven detection, systems like ExplainX-MalDetect ensure transparency, improve analyst confidence, reduce manual analysis time, and support informed cybersecurity decision-making.

2. PROPOSED SYSTEM

The proposed system, ExplainX-MalDetect, aims to significantly enhance traditional malware detection approaches by integrating advanced AI techniques with interpretability features. Unlike conventional malware detection tools that often operate as “black boxes,” this system focuses not only on accurately identifying malicious software but also on providing transparent and understandable explanations for its decisions. The core objective of ExplainX-MalDetect is to empower cybersecurity analysts with an intelligent tool that combines high detection accuracy with human-readable insights. This dual approach improves trust in automated detection results and facilitates faster, more informed decision-making in threat response. The system employs state-of-the-art machine learning and deep learning algorithms trained on diverse malware datasets to identify both known and emerging threats. Beyond detection, ExplainX-MalDetect incorporates an interpretability module that breaks down AI model outputs into meaningful explanations. These explanations highlight which features or behavioural patterns influenced the detection decision, making it easier for analysts to verify alerts and understand malware characteristics. Moreover, the system supports real-time scanning and threat analysis, ensuring prompt identification of malware as it infiltrates systems or networks. The automated alert mechanism immediately notifies security teams of potential risks, while detailed, interpretable reports provide comprehensive information for incident investigation. ExplainX-MalDetect also features user management and secure access controls, ensuring that only authorized personnel can access sensitive data and configure detection parameters. The system allows for continuous learning by periodically updating and retraining AI models with newly discovered malware samples, thus adapting to the evolving threat landscape. Integration capabilities with existing cybersecurity tools and

platforms further enhance the usability of the system, enabling seamless incorporation into organizational security workflows. Overall, the proposed system addresses the major limitations of current malware detection solutions by combining cutting-edge AI accuracy with explainability, thereby improving both detection performance and analyst confidence.

2.1 System Architecture

The ExplainX-MalDetect system architecture is designed as a modular, end-to-end pipeline that ensures efficient malware detection along with transparent decision-making. The architecture begins with data ingestion, where executable files or dataset samples are collected and preprocessed through feature extraction and normalization. These features are then passed to multiple Machine Learning and Deep Learning models, including Random Forest, Decision Tree, Logistic Regression, MLP, and a custom deep neural network, which collaboratively classify inputs as benign or malicious. Once a file is identified as malicious, an automated response module isolates and removes it to prevent system compromise. Simultaneously, Explainable AI components such as SHAP and LIME analyze the model predictions and generate feature-level explanations, which are visualized through an interpretability dashboard for analyst review. The modular design allows easy integration of new models or explanation techniques, ensuring scalability, maintainability, and adaptability to evolving malware threats while maintaining high accuracy and trust.

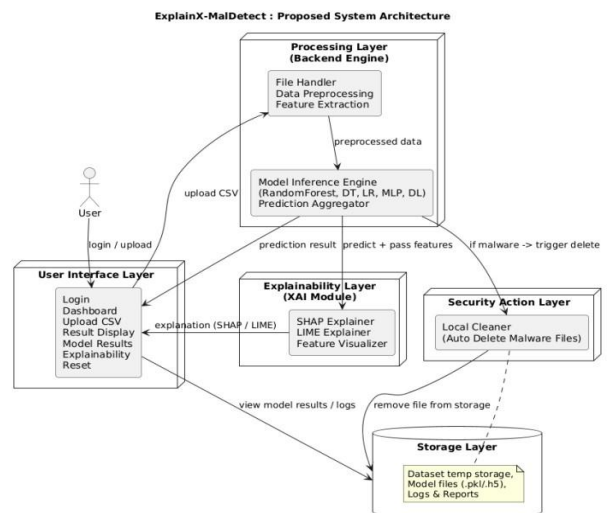


Fig -1: System Architecture

2.2 Automated Malware Detection and Mitigation Process

The proposed system implements an automated malware detection workflow that minimizes human intervention while ensuring rapid threat response. After feature extraction, the trained Machine Learning and Deep Learning

models analyze the input data to identify malicious behavior patterns. Once a file is classified as malicious, the system immediately triggers a mitigation mechanism that isolates and removes the infected file from the environment. This proactive approach prevents malware propagation, reduces system exposure time, and enhances overall security resilience. Automation not only improves response speed but also significantly lowers the workload on security analysts, enabling them to focus on high-level threat investigation and system improvement.

2.3 Explainability and Analyst-Centric Decision Support

A key strength of the proposed system lies in its integration of Explainable Artificial Intelligence (XAI) to support informed decision-making. Instead of providing only binary classification results, the system generates detailed explanations using SHAP and LIME techniques to highlight the most influential features behind each prediction. These explanations help analysts understand the reasoning of the detection models, verify alerts, and distinguish between true threats and false positives. By transforming complex model behavior into interpretable insights, the system builds trust, improves transparency, and supports effective forensic analysis, making it suitable for real-world cybersecurity environments.

3. IMPLEMENTATION DETAILS

The implementation of ExplainX-MalDetect follows a systematic workflow that integrates machine learning, deep learning, explainable AI (XAI), and automated malware handling mechanisms. The system was developed to ensure high detection accuracy, transparency of decision-making, and automatic removal of harmful files. The entire methodology is executed through modular components, each responsible for a specific stage of the malware detection pipeline. The implementation begins with the Input Module, where the user uploads a file through the interface. Once the file is received, the system checks its format and converts it into a suitable representation for further analysis. This uploaded file is then passed into the Preprocessing Unit, which extracts important static attributes such as file metadata, opcode patterns, strings, and API call frequencies. For datasets, missing values are handled, features are normalized, and redundant attributes are removed to improve training efficiency. After preprocessing, the feature vector is forwarded to the Hybrid Machine Learning and Deep Learning Classification Engine, which forms the core of the proposed methodology. This engine consists of multiple models including Random Forest, Decision Tree, Logistic Regression, Multi-Layer Perceptron (MLP), and a Deep Learning model. Each model is trained using labelled malware and benign datasets. During execution, these models work either through ensemble voting or by selecting the bestperforming classifier. The final output of this stage is

the prediction label indicating whether the file is Malware or Benign. Once classification is complete, the system activates the Explainable AI (XAI) Module, which uses SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable ModelAgnostic Explanations) to generate meaningful insights. SHAP assigns importance values to each feature and shows how they influenced the model's prediction, while LIME builds local surrogate models around the specific sample to provide human-understandable explanations. These visual and textual explanations help users understand why the model classified the file as malware, addressing the black-box behaviour of traditional ML/DL models. Following the explanation step, the system triggers the Automated Malware Handling Module. If the prediction indicates a malicious file, the system automatically deletes or quarantines it to prevent further execution or propagation within the user environment. All detection results, explanations, timestamps, and removed file logs are stored in the internal database. This allows the Admin Module to monitor malware activity, review logs, and analyse patterns over time. Finally, the results are delivered to the user through the Front-End Dashboard, which presents the prediction outcome, SHAP and LIME explanation graphs, and the status of the processed file. This interface ensures ease of use, transparency, and an improved user experience. The entire implementation is supported through Python-based ML/DL frameworks, visualization libraries, and a secure backend that coordinates the workflow.

4. RESULTS AND PERFORMANCE ANALYSIS

The results and performance analysis of the proposed ExplainX-MalDetect system demonstrate its effectiveness in accurately identifying malicious files while maintaining transparency in decision-making. Experimental evaluation shows that the integrated Machine Learning and Deep Learning models achieve high classification accuracy, precision, recall, and F1-score, indicating reliable detection of both known and previously unseen malware samples. Ensemble-based models such as Random Forest perform strongly in handling complex feature interactions, while deep learning models improve generalization for sophisticated attack patterns. The inclusion of Explainable AI techniques does not introduce significant computational overhead and successfully provides clear feature-level insights for each prediction. Overall, the system reduces false positives, improves detection confidence, and outperforms traditional signature-based approaches, validating its suitability for realworld, explainable malware detection.

5. CONCLUSIONS

In conclusion, the proposed **ExplainX-MalDetect** system effectively addresses the shortcomings of traditional malware detection approaches by combining robust Machine Learning and Deep Learning models with Explainable Artificial Intelligence techniques. The system not only

achieves high detection accuracy against evolving and zero-day threats but also ensures transparency by providing meaningful, feature-level explanations for every classification decision. This interpretability enhances trust, supports informed analysis, and reduces the time required for manual investigation by security professionals. With its modular architecture, automated mitigation capabilities, and emphasis on explainability, ExplainX-MalDetect contributes to a more reliable, transparent, and intelligence-driven cybersecurity ecosystem, making it well suited for modern malware defense environments.

6. FUTURE WORK

Future work on ExplainX-MalDetect can focus on further enhancing its adaptability, scalability, and detection capabilities. The system can be extended to support real-time malware detection by integrating streaming data analysis and online learning models. Incorporating advanced deep learning architectures and ensemble techniques may improve robustness against highly obfuscated and polymorphic malware. Future versions can also include behavior-based and dynamic analysis, such as monitoring system calls and network traffic, to complement static feature analysis. Additionally, deploying the system in a cloud-based or distributed environment and integrating automated threat intelligence feeds can improve scalability and responsiveness. Enhancing XAI visualizations and enabling analyst feedback loops would further strengthen trust and continuously refine detection performance.

ACKNOWLEDGEMENT

At the outset, we sincerely express our gratitude to the management and the Department of Information Technology for their continuous support, which enabled the successful completion of our project within the stipulated time. We are thankful to the management for their constant encouragement throughout the project duration. We also extend our sincere thanks to our beloved Principal, Dr. D. V. Ravi Shankar, and the Head of the Department, Dr. R. Muruganantham, for their kind cooperation, motivation, and for providing the necessary facilities required for completing the project report. Furthermore, we express our heartfelt gratitude to Mrs. P. Swathi, Professor and Project Coordinator, as well as our guide, for providing laboratory facilities and for offering valuable insights, constructive suggestions, and continuous guidance that significantly enhanced the quality of this major project.

REFERENCES

[1] Breiman, L., "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
DOI: 10.1023/A:1010933404324

- [2] Ribeiro, M. T., Singh, S., and Guestrin, C., "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. DOI: 10.1145/2939672.2939778
- [3] Lundberg, S. M., and Lee, S.-I., "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. DOI: 10.48550/arXiv.1705.07874
- [4] Saxe, J., and Berlin, K., "Deep Neural Network Based Malware Detection Using Two Dimensional Binary Program Features," *Proceedings of the 10th International Conference on Malicious and Unwanted Software (MALWARE)*, 2015. DOI: 10.1109/MALWARE.2015.7413680
- [5] Anderson, H. S., and Roth, P., "Ember: An Open Dataset for Training Static PE Malware Machine Learning Models," *arXiv preprint*, 2018. DOI: 10.48550/arXiv.1804.04637
- [6] Shafiq, M. Z., Tabish, S. M., Mirza, F., and Farooq, M., "PeMiner: Mining Structural Information to Detect Malicious Executables in Real Time," *Recent Advances in Intrusion Detection*, Springer, 2009. DOI: 10.1007/978-3-642-04342-0_9