Grenze International Journal of Engineering and Technology, January Issue

Ransomware Detection using Explainable AI For Endpoint Systems

Ms. Anu Prabhakar[1], Ms. Suganya R[2], Janani V[3], Narthika K[4], Thenmozhi H[5] and Kavin S[6]
[1-2]Assistant Professor, Department of Computer Science and Engineering, Coimbatore Institute of Technology, Coimbatore, India
Email: anuprabhakar@cit.edu.in
[3-6]Student, Department of Computer Science and Engineering, Coimbatore Institute of Technology, Coimbatore, India

***Abstract*— Ransomware is a rapidly evolving cyber threat that encrypts critical files and disrupts system availability, often bypassing traditional signature-based antivirus solutions and static machine learning models trained on controlled datasets. To address these limitations, an Explainable AI approach is employed to detect ransomware by monitoring endpoint behavior and distinguishing between normal and abnormal activities. Machine learning model is employed to classify five ransomware families Phobos, WannaCry, Hive, GandCrab, and LockBit using behavioral analysis. Key indicators such as unusual file encryption, rapid file modifications, suspicious process execution, and anomalous network activity are analyzed to identify malicious patterns. Explainable AI helps make the system's decisions easier to understand by showing which behaviors caused it to detect ransomware. A safe simulation environment supports the testing of ransomware-like behaviors without risk to actual systems.**

***Index Terms*— Ransomware Detection, Explainable AI, Endpoint Security, Machine Learning, Cybersecurity, Behavioral Analysis.**

I. INTRODUCTION

Ransomware has emerged as one of the most disruptive forms of cybercrime, targeting individuals, enterprises, and critical infrastructures worldwide. Its rapid evolution, combined with increasingly sophisticated evasion strategies, poses significant challenges to traditional signature-based and heuristic-based security solutions. Modern ransomware variants actively modify process workflows, encrypt critical files, and exploit system vulnerabilities while remaining undetected until substantial damage has occurred. As a result, the need for advanced, adaptive, and transparent detection mechanisms has become essential for securing endpoint environments.Machine learning–based detection systems have shown great promise in identifying malicious behaviors before file encryption begins. However, many existing models function as "black boxes," providing limited visibility into how decisions are made. Such opacity reduces trust, complicates incident investigation, and restricts the adoption of AI-driven security solutions in real-world enterprises. The cybersecurity domain increasingly demands models that not only perform accurately but also justify their predictions with interpretable evidence.Explainable Artificial Intelligence (XAI) addresses this requirement by enabling security systems to provide clear, human-understandable reasoning for their outputs. Integrating XAI with behavioral ransomware detection enhances analyst trust, improves forensic analysis, and supports faster and more informed mitigation decisions.

Grenze ID: 01.GIJET.12.1.132_1

Behavioral features such as process activity, file operations, registry modifications, and system resource usage provide rich insights into early-stage ransomware attacks when analyzed through interpretable machine learning techniques.Developing an effective ransomware detection system therefore requires a combination of robust behavioral analysis and transparent decision-making. An XAI- driven approach strengthens endpoint protection by reducing false positives, increasing threat visibility, and supporting accountability in automated security workflows. Such a system contributes to building resilient cyber-defense mechanisms capable of adapting to evolving ransomware threats.

II. Literature Review

Ransomware detection has been widely studied through static analysis, dynamic monitoring, and machine learning ap- proaches. Recent research on ransomware detection spans static analysis, dynamic monitoring, and machine learning– driven approaches, with growing emphasis on behavioral features, adversarial robustness, and explainability to overcome the limitations of traditional signature-based defenses. Alvi and Jalil (2025) introduced XRGuard, a model-agnostic framework that leverages Event Tracing for Windows (ETW) logs with machine learning classifiers, achieving high detection accuracy and improved interpretability through SHAP explanations, though real-time monitoring introduced
runtime overhead [1]. Jeong et al. (2025) developed MalFormer, a Vision Transformer–based technique that converts malware binaries into image representations, offering strong resilience against obfuscation but requiring substantial computational resources [2]. Yan et al. (2025) surveyed adversarial attack and defense strategies in malware classification, revealing the susceptibility of ML detectors to adversarial manipulation and the effectiveness of ensemble- based and adversarially trained models [3]. Ankalaki et al. (2025) compared classical machine learning with generative models such as GANs and VAEs for cyberattack prediction, finding improved generalization capabilities but at the cost of heavy data requirements and high computational demands [4]. Gebrehans et al. (2025) further demonstrated the utility of GANs for malware behavior simulation, though they highlighted challenges related to instability and limited fidelity in the generated adversarial samples [5]. Explainable AI (XAI) continues to gain prominence in cybersecurity research, with Manthena et al. (2025) emphasizing the role of SHAP, LIME, and saliency-based approaches in improving transparency despite their computational overhead [6]. Zhu et al. (2024) proposed a dynamic-analysis–based explanation framework that produced human-readable insights but faced scalability limitations when applied to large datasets or continuous monitoring [7]. Additionally, Gwak (2023) applied SHAP to ETW-based malware classifiers to diagnose misclassifications and reveal dataset inconsistencies, demonstrating how interpretability can assist in debugging ML systems [8]. Manthena et al. (2023) evaluated hybrid static–dynamic deep learning approaches, showing strong detection accuracy but noting deployment complexity and increased resource requirements [9]. Collectively, these studies demonstrate significant advancements in ML- and XAI-based ransomware detection, yet persistent challenges related to computational overhead, sandbox dependency, and limited real-world scalability highlight the need for a lightweight, interpretable, and endpoint-driven ransomware detection framework.

III. Proposed System

The proposed system is designed as a complete ransomware detection and classification framework that operates by capturing and analyzing behavioral patterns generated in a controlled simulation environment. Multiple ransomware families—WannaCry, LockBit, Hive, GandCrab, and Phobos—are safely emulated to produce realistic malicious behaviors such as file encryption, deletion, renaming, and abnormal file or network access, while parallel simulations of normal system activities create a reliable benign baseline. All events are logged with detailed attributes including EventID, ProcessID, ParentProcessID, ProcessName, CommandLine, Destination IP and Port, Timestamp, and Activity Label, forming a structured dataset suitable for multiclass classification. The system preprocesses this raw data by cleaning invalid records, standardizing timestamps, filling missing values, and transforming categorical fields into numerical form, while also aggregating event-level information into interval-based behavioral features such as counts of file, network, and process events, unique file extensions accessed, suspicious keyword matches, average command-line lengths, time gaps between events, and indicators of ransom note creation. A Random Forest classifier is then trained using these feature-rich records to differentiate benign activities from multiple ransomware families, leveraging the strength of ensemble learning to achieve high accuracy and generalization to unseen behaviors. To enhance system transparency, Explainable AI (XAI) using SHAP is integrated to show how specific behavioral features

influenced each prediction, enabling clear interpretation of detected threats. Finally, a real-time alert module presents immediate on-screen notifications with details such as threat type, affected processes, entropy levels, and recommended user actions. Together, these components form a robust, interpretable, and deployable ransomware defense system capable of real-time detection, explanation, and user-side threat awareness.

IV. SYSTEM ARCHITECTURE

The workflow begins with the Dataset Collection phase, where ransomware activities from families such as WannaCry, LockBit, Hive, GrandCrab, and Phobos are generated in a controlled simulation environment. Instead of using public repositories, realistic ransomware behaviors—like file encryption, renaming, deletion, and abnormal access patterns are recreated safely without causing real system harm. Alongside these malicious events, normal user operations such as file creation, document editing, and software execution are also recorded. Each captured event includes detailed attributes like event ID, process ID, parent process, event category, and labels, creating a comprehensive dataset that represents both benign and malicious system activities. In the Simulation and Event Collection phase, the system continuously monitors and records dynamic behaviors, capturing process-level activities, file operations, and network communications. These collected logs represent real-time ransomware execution traces and serve as the raw input for model development. This phase ensures diversity and completeness in behavioral data, which is crucial for training a robust detection model. Preprocessing phase focuses on transforming the collected raw data into a structured, analysis-ready format. Timestamps are standardized, missing values are handled, and irrelevant or du-plicate entries are removed. The data is then grouped by host and time intervals to generate aggregated behavioral features such as total event counts, file and network operations, unique process counts, suspicious command-line patterns, and ransom note creation. Categorical features like protocol and event category are encoded numerically, and the resulting dataset is clean, balanced, and labeled for use in machine learning.

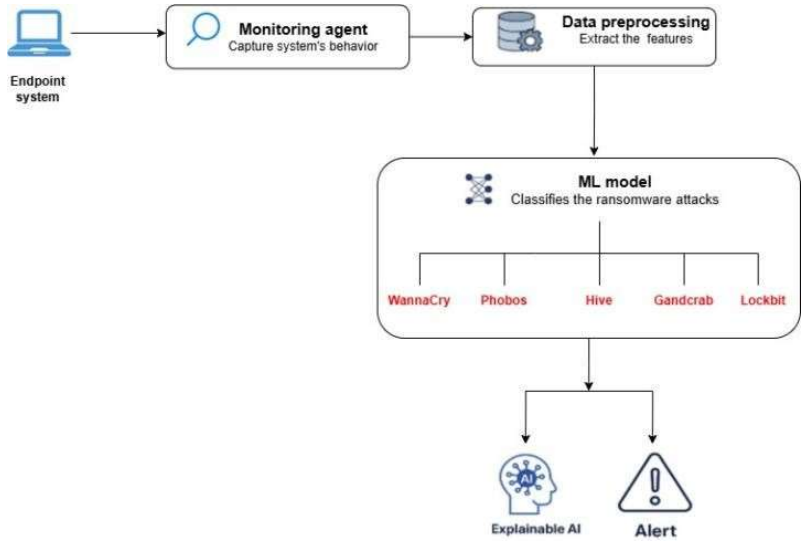

Figure 1. System Architecture

In the Model Training phase, a Random Forest classifier is employed to perform multiclass classification, distinguishing be-tween normal activity and the five ransomware families. The model learns behavioral patterns based on the engineered fea-tures, and its performance is evaluated using accuracy, precision, recall, and F1-score metrics. Feature importance analysis identifies the most significant indicators contributing to ransomware detection.Finally, in the alert and explainability phase, the trained model is deployed for real-time monitoring. As new system activities occur, they are analyzed instantly to determine if any behavior resembles known ransomware patterns. When suspicious activity is detected, an alert system notifies the user with critical details such as threat score, affected files, and suggested immediate actions. Simultaneously, Explainable AI (using SHAP) provides a transparent visualization showing which features influenced the model's decision. This final stage ensures that the system not only detects threats accurately but also explains its reasoning clearly, enabling users to take prompt and informed actions.

III. IMPLEMENTATION

A. Dataset Collection

A controlled simulation environment was created to replicate the behavior of multiple ransomware families such as WannaCry, LockBit, Hive, GandCrab, and Phobos. Malicious actions including file encryption, deletion, renaming, and abnormal access patterns were safely emulated without affecting the system. Parallelly, normal system activities like file creation, modification, and standard application operations were recorded to establish a benign baseline. Each event was captured with detailed attributes such as EventID, ProcessID, ParentProcessID, ProcessName, CommandLine, EventCategory, Destination IP, Port, Timestamp, and a label identifying the activity as malicious or benign. The resulting dataset provided a structured, multi-class labeled foundation for training and evaluating the detection model.

B. Dataset Preprocessing

The preprocessing stage involved cleaning, transforming, and structuring the raw simulation data. Relevant attributes such as process details, file operations, and network activity were extracted, while timestamps were standardized and missing values addressed. Events were grouped by time intervals and hosts to generate aggregated behavioral metrics, including counts of file, process, and network events, unique file extensions, suspicious keyword frequency, and ransom note indicators. Categorical variables were encoded numerically, resulting in a feature-rich, labeled dataset suitable for multiclass ransomware classification and behavioral analysis.

C. Random Forest Model

A Random Forest algorithm was implemented as the supervised learning model to classify ransomware behaviors. The model utilizes both numerical features (e.g., file operation counts, process durations, entropy) and categorical features (e.g.,
process names, file types). Data was divided into training and testing subsets to evaluate accuracy and generalization. The ensemble of decision trees predicts whether a behavior is benign or belongs to a specific ransomware family. Feature importance analysis highlights the most influential behavioral indicators, ensuring robust and interpretable ransomware detection suitable for deployment in real-time systems.

D. Explainable AI and Alert Mechanism

To enhance transparency, Explainable AI (XAI) techniques such as SHAP were integrated to interpret model predictions. This enables visualization of feature contributions, helping users understand why specific processes or file actions are flagged as suspicious. Upon detecting potentially harmful activity, the real-time alert system notifies users with details such as threat score, affected files, process information, entropy, and triggered detection rules. The system also provides immediate mitigation recommendations, including network disconnection, log inspection, and security contact instructions. This integration of explainable AI with alerting ensures actionable insights and trustworthy ransomware detection.

IV. RESULTS AND DISCUSSION

The experimental evaluation of the proposed ransomware detection framework was carried out to thoroughly assess its performance, detection capability, and interpretability across diverse ransomware families and normal system operations. To emulate real-world environments, a controlled execution setup was created where typical benign activities—such as file creation, file renaming, directory modification, and routine software operations—were executed alongside malicious ransomware behaviors like rapid file encryption, bulk file deletion, unauthorized registry updates, and abnormal access to sensitive directories. Logs generated from both benign and malicious executions were pre-processed, labeled, and structured into separate training and testing datasets. These datasets covered multiple prominent ransomware families, including Phobos, WannaCry, Hive, GandCrab, and LockBit, ensuring broad behavioral variability and robustness in model evaluation.The Random Forest classifier demonstrated consistently high performance across all evaluation metrics, achieving strong accuracy, precision, recall, and F1-scores for both benign and malicious classes. This indicates that the model not only successfully identified ransomware behaviors but also minimized false positives by correctly differentiating them from harmless system events. The ensemble-based nature of the Random Forest algorithm enabled effective handling of heterogeneous behavioral features, allowing it to generalize well across unseen ransomware patterns.To enhance transparency and trust in the decision-making process, Explainable AI (XAI) techniques—specifically SHAP (SHapley Additive exPlanations)—were integrated. SHAP visualizations

provided clear insights into how individual behavioral features influenced model predictions. For example, features such as unusually high file entropy, rapid modification rates, and irregular registry access contributed strongly to classifying a process as malicious. The interpretability framework helped in validating the correctness of the model's decisions by showing why certain processes were flagged, as illustrated below

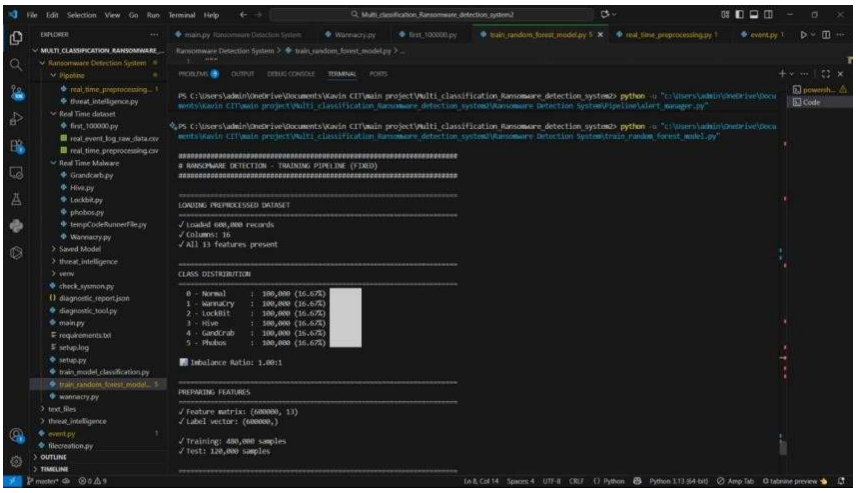

Figure 2 Random Forest Model for Training and Validation

Random Forest model serves as the core machine-learning classifier for detecting and categorizing ransomware behavior in the proposed system. It is an ensemble-based algorithm that constructs multiple decision trees during training and combines their outputs to produce a more accurate and robust prediction. Each tree learns from different subsets of the data, enabling the model to capture diverse behavioral patterns without overfitting

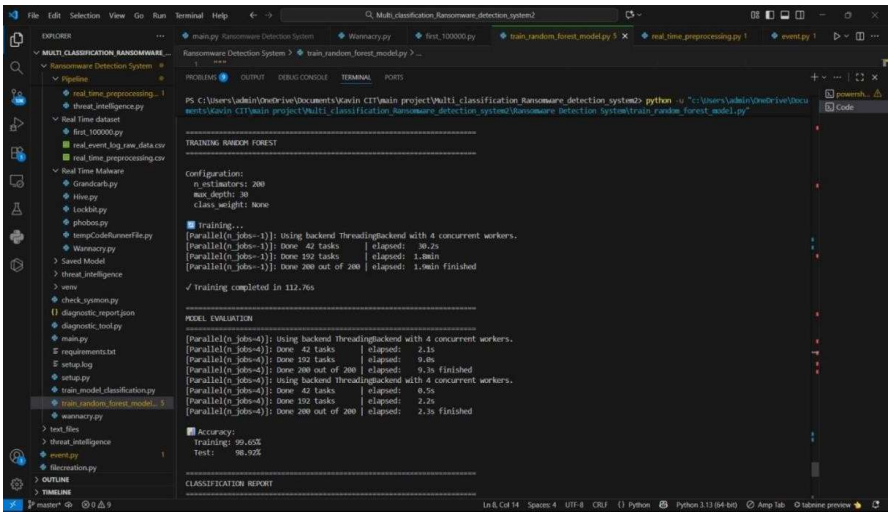

Figure 3 Classification accuracy

Fig 2,3 During the training phase, the model is fed with preprocessed behavioral logs consisting of features related to process activity, file operations, registry interactions, and system events. By learning from these patterns, the Random Forest develops the ability to distinguish between benign system activity and ransomware-specific behaviors. The model's multi-tree architecture ensures high stability, even when confronted with noisy, imbalanced, or high-dimensional data.For validation, the dataset is split into training and testing sets to evaluate the model's generalization capability. The Random Forest produces key performance measures such as accuracy, precision, recall, and F1-score, which help assess how effectively the model identifies different ransomware families like Phobos, WannaCry, Hive, GandCrab, and LockBit.

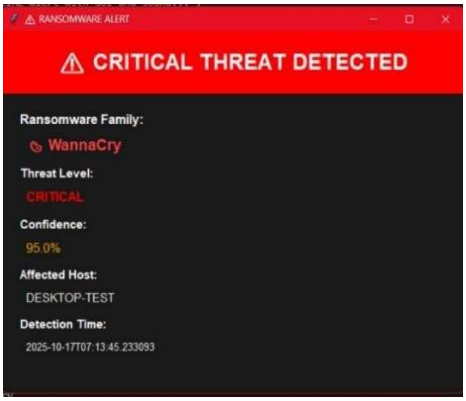

Figure.4 Malware classification and alert

Figure 4 illustrates the real-time ransomware alert window generated by the proposed detection system immediately after malicious activity is classified. In this instance, the system identifies the WannaCry ransomware family, one of the most destructive strains known for rapid file encryption and network propagation. The alert interface is designed with a bold red header titled "CRITICAL THREAT DETECTED", ensuring high visibility and conveying the urgency of the situation to system administrators or end-users.The alert window displays essential threat information, including the ransomware family name, threat severity, model confidence score, and the affected host or workstation.

The alert interface displays essential information such as the ransomware family, threat level, confidence percentage, and affected host as shown in Fig.5

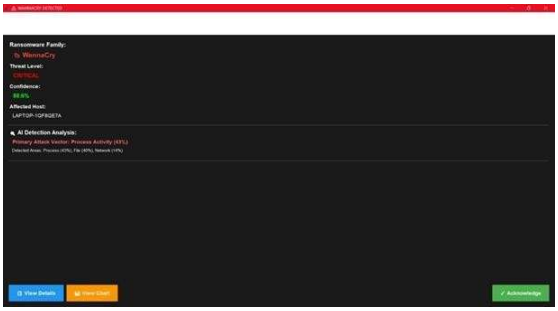

Figure.5 Ransomware type and threat level

When the user selects "View Chart," the SHAP-based AI Detection Analysis shown in Fig.4visualizes the attack area distribution, category impact ranking, and top 10 feature contributions that influenced the model's prediction. From the visualization, it is observed that process activity and file activity have the highest impact, indicating that the ransomware primarily affects system processes and file operations.

Figure.6 Chart representing the important features

Fig 6 shows the AI Detection Analysis for the WannaCry ransomware, generated using SHAP-based Explainable AI. The pie chart on the left displays the attack area distribution, highlighting that process activity contributes the highest portion to the detection, followed by file and network activities. The bar chart on the top right ranks the category impact, showing how significantly each behavior type influenced the model's decision. The lower chart presents the Top 10 SHAP feature contributions, indicating which specific system events either increased or decreased the likelihood of classifying the activity as WannaCry. Features with positive SHAP values strongly pushed the prediction toward ransomware classification, while negative values indicated normal behavior.

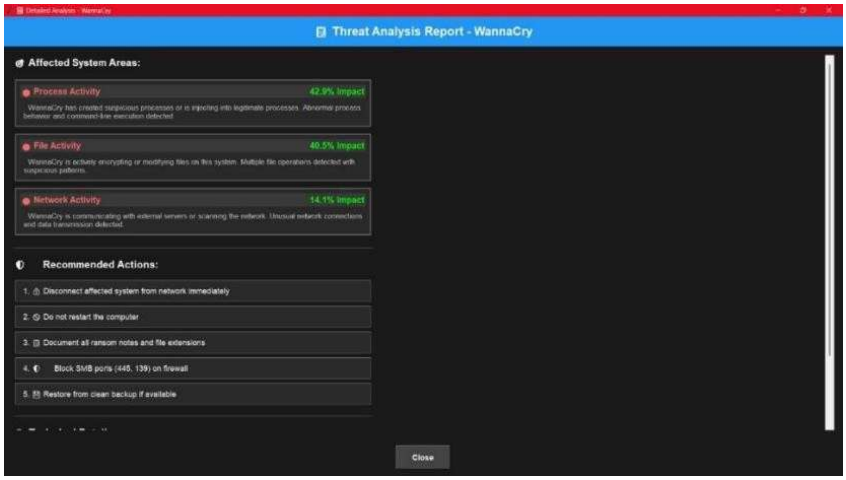

Figure.7 Threat analysis report

Fig 7 shows the Threat Analysis Report for WannaCry, highlighting the system areas affected during the attack. The model identifies abnormal process activity (42.9%), file activity (40.5%), and network activity (14.1%), indicating strong ransomware behavior. Each section explains what suspicious actions were detected, such as rapid file encryption and unusual process execution. The report also provides essential recommended actions, including disconnecting the system, avoiding restart, documenting ransom notes, blocking SMB ports, and restoring from a clean backup.

V. CONCLUSION

The ransomware detection system successfully demonstrates an effective, intelligent, and comprehensive approach to identifying modern ransomware threats. By utilizing machine learning techniques trained on behavioral and registry- based features, the system is capable of detecting both benign activities and a wide range of ransomware families including Phobos, WannaCry, Hive, GandCrab, and LockBit. This multi-class classification ability ensures that the solution does not merely identify suspicious activity but precisely recognizes the specific ransomware variant, improving accuracy and enabling targeted response strategies.The integration of Explainable AI (XAI) through SHAP values significantly enhances the trustworthiness of the model by providing clear explanations for each detection decision. This transparency allows cybersecurity analysts to better understand the underlying behaviors that influenced the classification, bridging the gap between complex machine learning models and human interpretability. Additionally, the real-time alert mechanism, delivered through a user-friendly dialog box, ensures immediate threat notification with clear details such as threat level, confidence score, and affected system. The system further strengthens security response by generating a detailed threat analysis report, highlighting impacted system areas and offering practical mitigation steps. This ensures that detection is complemented by actionable guidance, supporting faster containment and recovery.

ACKNOWLEDGMENT

The authors wish to thank Ms. Anu Prabhakar, Assistant Professor and Ms. Suganya R, Assistant Professor,Department of Computer Science and Engineering, Coimbatore Institute of Technology, for their valuable guidance and constant support throughout this project.

REFERENCES

[1] M. A. Alvi and Z. Jalil, "XRGuard: A Model-Agnostic Approach to Ransomware Detection Using Dynamic Analysis and Explainable AI," Department of Cyber Security, Faculty of Computing and Artificial Intelligence, IEEE, 2025.

[2] In-Woong Jeong, Han-Ju Lee, Gwang-Nam Kim, and Seok-Hwan Choi, "MalFormer: A Novel Vision Transformer Model for Robust Malware Analysis," Yonsei University, Republic of Korea, IEEE, 2025.

[3] Senming Yan, Jing Ren, Limin Sun, Wei Zhang, Wei Wang, and Quan Yu, "A Survey of Adversarial Attack and Defense Methods for Malware Classification in Cyber Security," IEEE, 2025.

[4] Shilpa Ankalaki, Geetabai S. Hukkeri, Aparna Rajesh Atmakuri, M. Pallavi, Tony Jan, and Ganesh R. Naik, "Cyber Attack Prediction: From Traditional Machine Learning to Generative Artificial Intelligence," IEEE, 2025.

[5] Ghebrebrhan Gebrehans, Naveed Ilyas, Khouloud Eledlebi, Willian Tessaro Lunardi, Martin Andreoni, Chan Yeob Yeun, and Ernesto Damiani, "Generative Adversarial Networks for Dynamic Malware Behavior: A Comprehensive Review, Categorization, and Analysis," IEEE, 2025.

[6] Harikha Manthena, Mahmoud Abdelsalam, Shaghayegh Shajarian, Jeffrey C. Kimmell, and Sajad Khorsandr, "Explainable Artificial Intelligence (XAI) for Malware Analysis: A Survey of Techniques, Applications, and Open Challenges," IEEE, 2025.

[7] H. Zhu, X. Chen, L. Wang, Z. Xu, and V. S. Sheng, "A Dynamic Analysis-Powered Explanation Framework for Malware Detection," IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 12, pp. 7483–7496, Dec. 2024.

[8] J. Y. Gwak, "Debugging Malware Classification Models Based on Event Tracing for Windows Logs and SHAP Explanations," Proceedings of CxAI-2023, 2023.

[9] H. Manthena et al., "Deep Learning Based Hybrid Analysis of Malware Detection and Classification: A Recent Review," 2023