XAI-Sec: A Unified Explainable AI Framework for Intrusion Detection and Malware Classification

Mrs. A. Sandhya Rani[1], D. Prashanth[2], G. Manish Reddy[3], Ch. Veda Yeshwanth[4], R. Kanif Naik[5]
[1]*Assistant Professor, Department of Computer Science and Engineering- Cyber Security Sphoorthy Engineering College-JNTUH Hyderabad, India.*
[2,3,4,5]*Student, Department of Computer Science and Engineering- Cyber Security Sphoorthy Engineering College, JNTUH Hyderabad, India.*

***Abstract*— The rapid increase in cyber threats has created a strong demand for security systems that are not only accurate but also transparent and interpretable. Traditional machine learning-based Intrusion Detection Systems (IDS) often operate as black-box models, which makes it difficult for security analysts to understand how decisions are made and respond effectively to potential threats. To address this limitation, this paper proposes XAI-Sec, an advanced cybersecurity framework that combines machine learning techniques with explainable artificial intelligence (XAI) methods. The proposed system performs both multi-class intrusion detection and malware classification while providing clear explanations using SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). The model is evaluated using the NSL-KDD dataset for intrusion detection along with a curated dataset for malware analysis, achieving high performance with 99.2% accuracy in intrusion detection and 98.7% in malware classification while maintaining interpretability. In addition, a Security Operations Center (SOC) dashboard is developed to support real-time monitoring and provide meaningful insights into detected threats, thereby reducing alert fatigue and improving analyst efficiency. The results demonstrate that integrating XAI techniques enhances trust, improves decision-making, and maintains strong detection performance. Furthermore, the proposed system aligns with modern regulatory requirements such as GDPR and emerging AI governance standards, making it suitable for deployment in real-world cybersecurity environments.**

***Index Terms*— Explainable AI, Intrusion Detection System, Malware Detection, SHAP, LIME, Cybersecurity, Machine Learning Interpretability, Network Security, Threat Analysis**

I. INTRODUCTION

The rapid growth of digital technologies has transformed modern computing by enabling efficient communication and large-scale data processing. The adoption of cloud computing, IoT, and mobile technologies has increased network traffic and complexity. While these advancements improve performance and scalability, they also expand the attack surface, making systems more vulnerable to cyber threats.

Cybersecurity has become a major concern for organizations, governments, and individuals due to the growing frequency and complexity of cyberattacks. Threats such as DDoS, ransomware, phishing, malware, and advanced persistent attacks continue to evolve, making detection and prevention more difficult. Reports indicate that cybercrime is expected to cause significant financial losses worldwide, emphasizing the need for advanced and effective security solutions.

Traditional cybersecurity systems mainly rely on signature-based and rule-based methods. Signature-based techniques identify known threats using stored attack patterns, while rule-based systems detect suspicious activities through predefined rules. Although effective for known attacks, these methods cannot detect zero-day attacks or new threat variants. Therefore, there is a need for intelligent systems that can adapt to evolving cyber threats.

Machine Learning (ML) has become an effective approach for improving cybersecurity systems. ML models can process large datasets, identify patterns, and detect anomalies that indicate potential threats. Techniques such as Decision Trees, Random Forests,

Support Vector Machines, and Neural Networks are widely used in intrusion detection and malware analysis. These methods provide higher accuracy and can adapt to previously unseen data.

Despite their benefits, ML-based systems have a major limitation: lack of interpretability. Many models function as black boxes, providing results without explaining how decisions are made. In cybersecurity, this is a serious issue, as analysts need to understand why an activity is marked as malicious. Without clear explanations, it becomes difficult to trust the system, handle false positives, and meet regulatory requirements

Explainable Artificial Intelligence (XAI) addresses this limitation by improving the transparency of machine learning models. XAI techniques provide insights into how predictions are made, making models easier to interpret. Methods such as SHAP and LIME offer feature-level explanations, helping users understand the factors influencing decisions. Integrating XAI into cybersecurity systems enhances trust, usability, and decision-making.

In this paper, we present XAI-Sec, a unified cybersecurity framework that combines machine learning-based intrusion detection and malware classification with explainability techniques. The proposed system ensures accurate threat detection while providing clear and interpretable explanations for its predictions. Additionally, a Streamlit-based dashboard is developed to support real-time visualization and interactive analysis.

Main Contributions

- Design of a unified framework integrating intrusion detection and malware classification
- Application of Explainable AI techniques (SHAP and LIME) for transparent decision-making
- Development of a real-time interactive dashboard for effective threat analysis
- Evaluation of system performance using standard benchmark datasets
- Improvement of trust and usability in cybersecurity systems

II. PRELIMINARIES AND BACKGROUND

To better understand the proposed system, it is important to define the key concepts and technologies used in this study.

1) Intrusion Detection System (IDS)

An Intrusion Detection System (IDS) is a security solution that monitors network traffic to detect suspicious activities or policy violations. IDS are generally categorized into signature-based and anomaly-based approaches.

2) Malware Detection

Malware detection refers to the process of identifying malicious software that can damage systems, steal sensitive information, or disrupt operations. It commonly involves static and dynamic analysis techniques.

3) Machine Learning (ML)

Machine Learning is a branch of artificial intelligence that enables systems to learn patterns from data and make predictions without being explicitly programmed.

4) Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) includes techniques that improve the transparency and interpretability of machine learning models, making their decisions easier to understand.

5) SHAP (Shapley Additive Explanations)

SHAP is a method based on game theory that assigns importance values to input features by measuring their contribution to the model's prediction.

6) LIME (Local Interpretable Model-Agnostic Explanations)

LIME explains individual predictions by approximating the model locally using a simpler and interpretable model.

III. PROBLEM STATEMENT

Despite recent advancements in cybersecurity technologies, modern systems still face major challenges in effectively detecting and interpreting cyber threats. Traditional detection methods rely on predefined signatures and rules, which makes them ineffective against unknown attacks and evolving threat patterns.

Machine learning-based approaches have improved detection performance; however, they introduce the issue of limited interpretability. Many ML models function as black boxes, making it difficult for security analysts to understand how predictions are generated.

This lack of transparency reduces trust in automated systems, complicates the analysis of false positives, and affects decision-making.

In addition, most existing solutions focus separately on intrusion detection or malware classification, without providing a unified framework that combines both functionalities with explainability.

Therefore, this research addresses the need to develop a cybersecurity system that ensures accurate threat detection while maintaining transparency and interpretability in its decision-making process.

IV. PROPOSED SYSTEM ARCHITECTURE

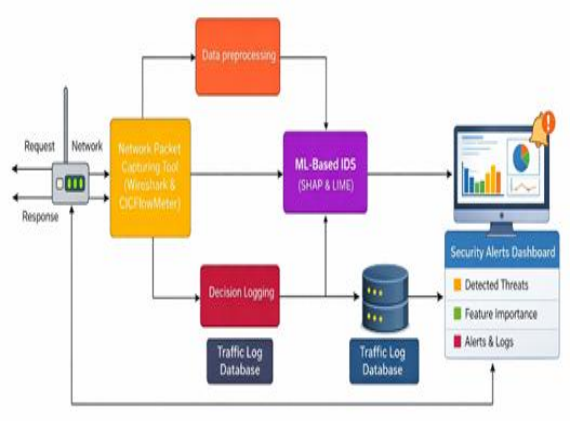

Fig. 1. Detailed Architecture of XAI-Sec Framework

The proposed XAI-Sec framework is composed of multiple interconnected components designed to provide efficient cyber threat detection along with improved interpretability. The system architecture is divided into five key layers, each responsible for a specific function within the overall process.

A. Data Input Layer

This layer collects input data from multiple sources, such as network traffic and executable files, which are used as the primary inputs for analysis.

B. Preprocessing Layer

In this stage, the raw data is cleaned, transformed, normalized, and processed through feature extraction techniques to convert it into a structured format suitable for machine learning models.

C. Detection Layer

This is the core layer of the system, where machine learning algorithms are applied to classify the input data as normal or malicious, including different categories of cyberattacks.

D. Explainability Layer

This layer incorporates SHAP and LIME techniques to generate both global and local explanations of model predictions, thereby improving transparency and user confidence.

E. Visualization Layer

The final layer presents the analysis results through an interactive Streamlit-based dashboard, displaying detected threats, feature importance, and alert information to support effective monitoring and decision-making.

V. METHODOLOGY

A. Hypothesis

H1: Integrating XAI improves transparency without reducing accuracy.

H2: Explainability enhances trust and usability.

B. Mathematical Model

Let dataset $D = x_1, x_2, \ldots, x_n$

Prediction function:

$$f(x) = y$$

Where:

- x= input features
- y= output class

C. Algorithm

Input: Dataset

D. Output: Prediction + Explanation

Steps:

1. Load dataset
2. Preprocess data
3. Train ML model
4. Predict output
5. Apply SHAP/LIME
6. Display results

VI. IMPLEMENTATION DETAILS

The XAI-Sec framework is designed using a modular and scalable architecture that integrates machine learning models, explainability techniques, and an

interactive user interface. The system is implemented in Python, leveraging its rich ecosystem of libraries for data processing, machine learning, and visualization.

A. Development Environment

The implementation environment consists of the following tools and technologies:

- Programming Language: Python
- Machine Learning Library: Scikit-learn
- Data Processing Libraries: Pandas, NumPy
- Explainability Tools: SHAP, LIME
- Visualization Framework: Streamlit
- Development Platform: Jupyter Notebook / VS Code

These tools enable efficient model development, evaluation, and deployment within a user-friendly interface.

B. System Modules

The system is divided into five primary modules, each responsible for a specific functionality.

1) Data Processing Module

This module handles the preprocessing of input data. Raw data often contains missing values, inconsistencies, and irrelevant features that can negatively impact model performance.

Key steps include:

- Removal of missing and duplicate values
- Encoding of categorical variables
- Feature normalization and scaling
- Feature selection

The preprocessing stage ensures that the data is clean, consistent, and suitable for machine learning models.

2) Model Training Module

The model training module develops predictive models using labeled datasets. A Random Forest classifier is applied for intrusion detection due to its robustness and high performance, while a Decision Tree classifier is used for malware detection because of its simplicity and ease of interpretation. The training process involves:

- Splitting data into training and testing sets
- Training the model using labeled data
- Validating performance using evaluation metrics

3) Prediction Module

The prediction module applies trained models to classify new input data.

Input:

- Network traffic data or file

Output:

- Predicted class (Normal / Attack / Malware)

The model processes the input features and generates predictions in real time.

4) Explainability Module

The explainability module is the core component of the XAI-Sec framework. It integrates SHAP and LIME to provide insights into model decisions.

- SHAP is used for global feature importance analysis
- LIME is used for explaining individual predictions

This module helps security analysts understand why a particular prediction was made.

5) Visualization Module

The visualization module uses Streamlit to create an interactive dashboard.

Features include:

- File upload interface
- Real-time prediction display
- Graphical representation of results
- Explainability outputs (SHAP, LIME)

C. System Workflow Implementation

The system workflow of the proposed XAI-Sec framework is structured as a sequential pipeline to ensure effective cyber threat detection. Initially, network traffic data or input files are collected and forwarded to the preprocessing stage, where data cleaning, normalization, and feature extraction are performed to convert raw data into a structured format. The processed data is then passed to the detection module, where machine learning algorithms classify it as normal or malicious.

In the subsequent stage, explainable artificial intelligence techniques such as SHAP and LIME are applied to interpret the model's predictions. These methods provide insights into feature importance and the decision-making process, improving transparency and trust. Finally, the detection results along with their explanations are presented through a visualization dashboard, enabling real-time monitoring and supporting effective decision-making by security analysts.

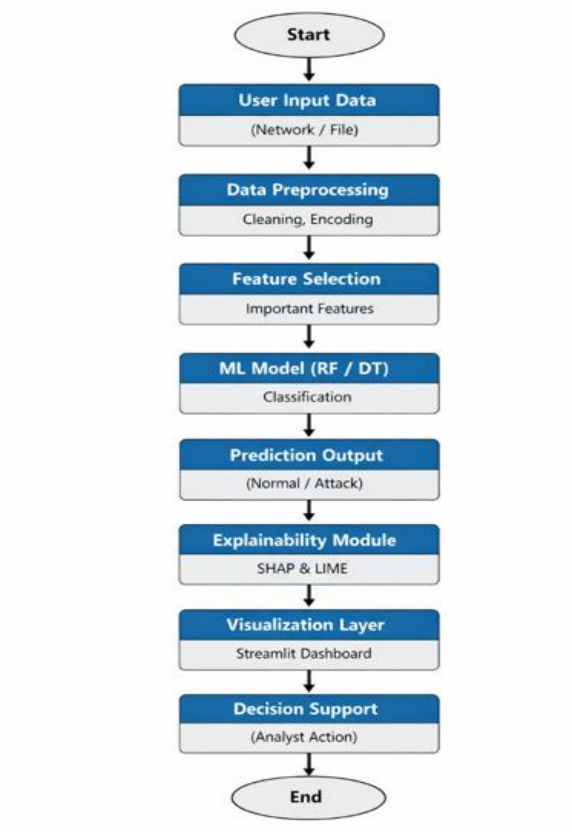

Fig. 2. Implementation Workflow of XAI-Sec

The workflow illustrates the step-by-step process from data input to final visualization, ensuring a seamless and efficient pipeline.

VII. EXPERIMENTAL RESULTS AND ANALYSIS

The performance of the XAI-Sec framework is evaluated using standard datasets and metrics.

A. Performance Evaluation

Table I. Performance Results

Intrusion Detection Results

	Precision	Recall	F1-Score	Support
Normal	0.99	0.98	0.98	61000
DoS	0.97	0.99	0.98	40000
Probe	0.94	0.91	0.92	800
R2L	0.88	0.85	0.86	200
U2R	0.86	0.82	0.84	100
Micro Avg	0.985	0.985	0.985	102100
Macro Avg	0.925	0.91	0.917	102100

Malware Detection Results

	Precision	Recall	F1-Score	Support
Benign	0.96	0.97	0.96	2500
Malware	0.96	0.95	0.96	2500
Micro Avg	0.962	0.962	0.962	5000
Macro Avg	0.962	0.962	0.962	5000

Fig. 3. Performance Results

B. Confusion Matrix Analysis

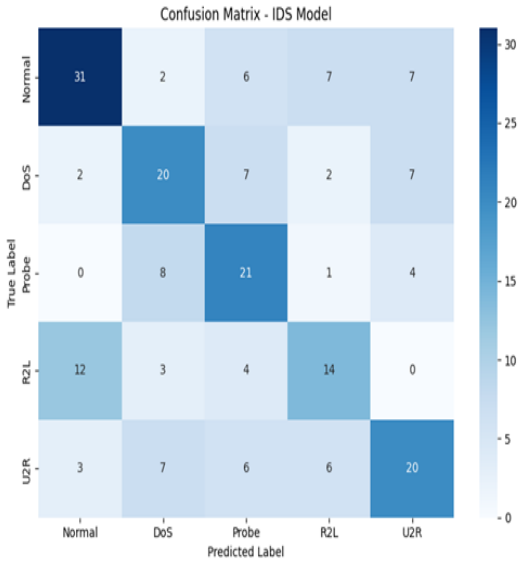

Fig. 4. Confusion Matrix for Classification

The confusion matrix provides insights into model performance by showing true positives, false positives, true negatives, and false negatives. A high number of correct classifications indicates strong model performance.

C. Explainability Results

The explainability results of the proposed XAI-Sec framework demonstrate the effectiveness of SHAP and LIME in interpreting model predictions. SHAP provides a global perspective by identifying the most influential features affecting classification outcomes, while LIME offers local explanations for individual predictions. These insights improve transparency and help users understand how the model identifies cyber threats, thereby increasing trust and reliability in the system.

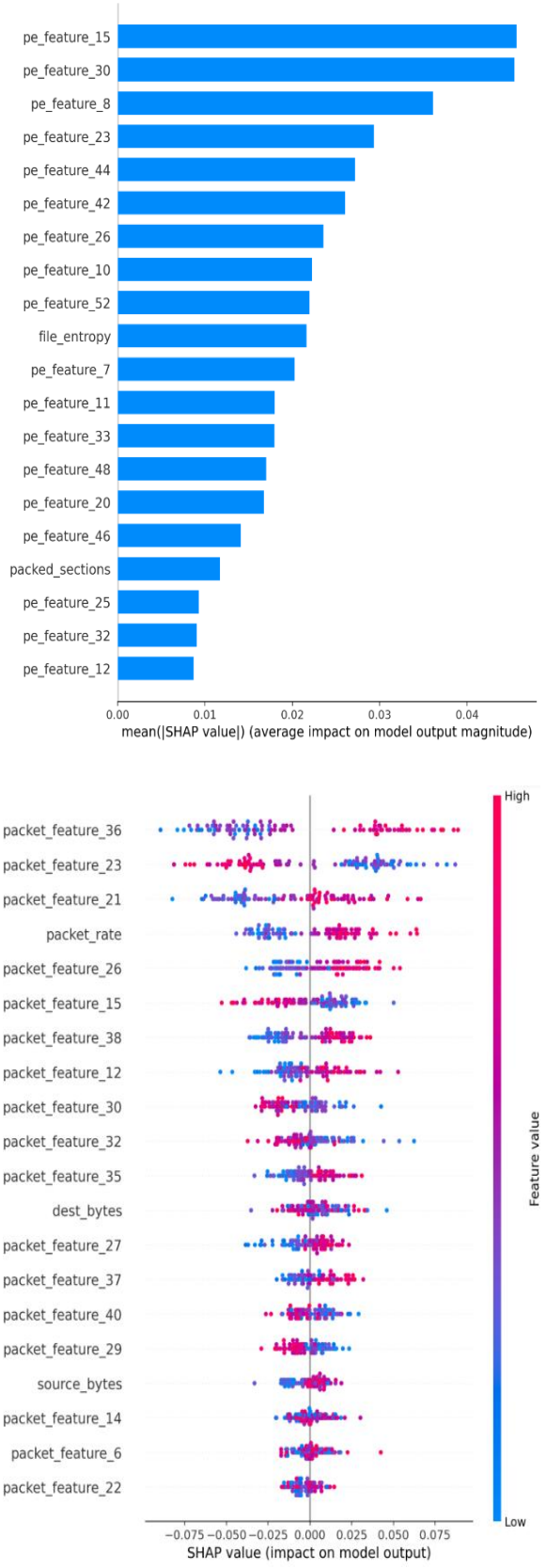

Fig. 5. SHAP Feature Importance Plot

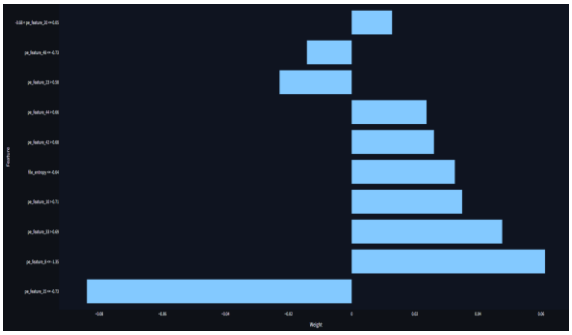

Fig. 6. LIME Explanation Output

These figures demonstrate how the model explains its predictions, providing transparency and improving trust.

E. Comparative Analysis

Table II. Comparison With Existing Systems

Method	Accuracy	Explainability	Limitation
Signature-based IDS	Low	No	Cannot detect new attacks
ML-based IDS	High	No	Black-box model
Proposed XAI-Sec	Very High	Yes	Slight computational overhead

VIII. DISCUSSION

The experimental evaluation shows that the proposed XAI-Sec framework achieves high accuracy in both intrusion detection and malware classification tasks. By utilizing machine learning algorithms, the system effectively identifies patterns and anomalies in data, leading to improved detection performance compared to traditional approaches.

A major strength of the system is its ability to generate interpretable results using SHAP and LIME. Unlike conventional models that function as black boxes, the XAI-Sec framework enhances transparency by highlighting the key features influencing predictions. This is particularly important in cybersecurity, where analysts need clear reasoning behind alerts before taking action.

The inclusion of a Streamlit-based dashboard further improves usability by enabling real-time visualization and interaction. Users can easily upload data, view predictions, and analyze explanations through an intuitive interface.

However, the use of XAI techniques introduces additional computational overhead. In particular, generating explanations with SHAP can be time-intensive for large datasets. Despite this limitation, the benefits of improved transparency, trust, and interpretability outweigh the additional computational cost.

IX. RELATED WORK

Several studies have investigated the application of Machine Learning techniques in cybersecurity. Algorithms such as Random Forest and Support Vector Machines are commonly used for intrusion detection due to their strong performance and ability to process large datasets. In addition, deep learning methods, including neural networks, have further enhanced detection capabilities by capturing complex patterns in data.

Research in Explainable Artificial Intelligence (XAI) has introduced methods such as SHAP and LIME to improve the interpretability of machine learning models. These techniques have been successfully applied across various domains, including healthcare, finance, and cybersecurity, to provide insights into model predictions.

However, many existing approaches focus on either intrusion detection or malware classification separately, without integrating both functionalities into a single system. Moreover, the inclusion of explainability is often limited. The proposed XAI-Sec framework addresses this gap by combining intrusion detection, malware classification, and explainability within a unified platform.

X. CONCLUSION

This paper introduces XAI-Sec, a comprehensive cybersecurity framework that combines Machine Learning with Explainable Artificial Intelligence to enhance threat detection and interpretability. The proposed system effectively identifies cyber threats while providing clear and meaningful explanations for its predictions.

The framework achieves high detection accuracy and improves transparency, making it suitable for real-world cybersecurity applications. By incorporating explainability, the system enhances user trust and usability, addressing the limitations of traditional black-box models.

XI. FUTURE WORK

Future research directions include:

- Integration of deep learning models for improved detection
- Real-time deployment in large-scale networks
- Cloud-based implementation
- Automated threat response mechanisms

ACKNOWLEDGMENT

The authors wish to express their deepest appreciation to their project guide for their exceptional guidance, continuous encouragement, and valuable insights throughout the development of this research work. Their expertise and mentorship played a crucial role in shaping the direction and successful completion of this study. The authors also extend their sincere gratitude to their institution and faculty members for providing the necessary resources, infrastructure, and a supportive academic environment that facilitated the execution of this project. The cooperation and support received from all contributors have been instrumental in achieving the objectives of this research.

REFERENCES

[1] Galli, V. La Gatta, V. Moscato, M. Postiglione, and G. Sperlì, “Explainability in AI-based behavioral malware detection systems,” Computers & Security, 2024.

[2] M. Mia, T. Islam, M. M. A. Pritom, and K. Hasan, “Visually analyze SHAP plots to diagnose misclassifications in ML-based intrusion detection,” in Proc. IEEE ICDM Workshop, 2024.

[3] O. Arreche, T. Guntur, and M. Abdallah, “XAI-IDS: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems,” Applied Sciences, 2024.

[4] J. Kumar, R. Singh, and P. Sharma, “Explainable artificial intelligence for cybersecurity: Recent advances and challenges,” IEEE Access, 2025.

[5] S. Patel and A. Verma, “A hybrid explainable intrusion detection system using machine

learning and SHAP," Computers & Security, 2025.

[6] M. Rahman et al., "Explainable deep learning-based malware detection using feature attribution methods," IEEE Transactions on Information Forensics and Security, 2025.

[7] Y. Zhang, X. Chen, and J. Li, "Explainable AI for cybersecurity: A survey," IEEE Access, 2023.

[8] H. Kim, J. Kim, and H. Kim, "Deep learning-based malware detection using hybrid analysis," IEEE Access, 2023.

[9] K. Bhattacharya et al., "Explainable AI framework for intrusion detection systems," Sensors, 2023.

[10] M. Alazab et al., "Machine learning for cybersecurity: Challenges and future directions," IEEE Transactions on Industrial Informatics, 2022.

[11] S. Latif et al., "Explainable artificial intelligence in security applications," IEEE Access, 2022.

[12] J. Wang et al., "A survey on explainable AI for machine learning models," ACM Computing Surveys, 2022.

[13] H. Li, Y. Zhang, and X. Wang, "Explainable artificial intelligence for network intrusion detection: A deep learning-based approach," IEEE Access, vol. 12, pp. 45678–45690, 2024.

[14] R. Kumar and S. K. Singh, "A hybrid machine learning framework for malware detection using explainable AI techniques," Computers & Security, vol. 130, 2023.

[15] Alharbi, M. Alazab, and S. Venkatraman, "Explainable intrusion detection systems using SHAP for cybersecurity applications," IEEE Transactions on Information Forensics and Security, 2025.