Information Security Education Journal (ISEJ)

DLINE JOURNALS

Print ISSN: 2349-8161
Online ISSN: 2349-817X

ISEJ 2026:13 (1)
https://doi.org/10.6025/isej/2026/13/1/1-19

Interpretable Cyber Threat Detection Using SHAP-Based Explainable AI in Classroom Data Security Systems

Dussadee Seewungkum
Faculty of Humanities
Srinakharinwirot University
Bangkok, Thailand
dussadee@g.swu.ac.th

ABSTRACT

This study presents an interpretable cyber threat detection framework using SHAP-based Explainable AI (XAI) for classroom data security systems. Addressing vulnerabilities in educational digital ecosystems, the research analyzes a simulated dataset containing diverse threat categories including insider threats, phishing, unauthorized access, data breaches, and malware to develop a transparent, predictive intrusion detection model. The methodology employs tree-based classifiers enhanced with SHAP values to quantify feature contributions, enabling both global and local interpretability of threat predictions.

Key findings reveal that insider threats (n=218) and phishing attacks (n=204) dominate the threat landscape, underscoring human-centric vulnerabilities over purely technical exploits. Threat severity distribution is relatively balanced across low, medium, and high categories, supporting robust multi-class classification. Analysis of access levels indicates administrative and standard user accounts represent primary attack surfaces, highlighting the need for stringent privilege management. SHAP interpretability results identify system response time as the most influential predictive feature, demonstrating that behavioral and performance anomalies are critical indicators of malicious activity.

The study advocates for adaptive, behavior centric cybersecurity frameworks integrating User Behavior Analytics, Role-Based Access Control, data centric protection strategies, and continuous user awareness training. By bridging predictive accuracy and model transparency, the SHAP enhanced approach empowers security analysts to understand the rationale behind detections, reduce false positives, and respond effectively to evolving threats. Ultimately, this research contributes a comprehensive, interpretable methodology for safeguarding smart educational environments, emphasizing that effective cybersecurity requires combining technical defenses with human-factor considerations and explainable artificial intelligence.

Keywords: Explainable AI (XAI), SHAP (Shapley Additive exPlanations), Cyber threat detection, Intrusion Detection Systems (IDS), Smart classroom security, Insider threats, Machine learning classifiers, User behavior analytics, Educational data protection

Received: 18 September 2025, Revised 29 November 2025, Accepted 8 December 2025

1. Introduction

Conventional classroom management systems are still used by educational institutions in many parts of the world. These systems frequently have inefficiencies in energy use, operational control, and security management. There is still a widespread prevalence of challenges, including manual attendance monitoring, unauthorised access, and inefficient resource utilisation. In these situations, the adoption of sophisticated technologies has been hindered by inadequate infrastructure and limited access to scalable, cost-effective smart solutions [1]. With the research on cost-effective models, these deficiencies are addressed by offering a smart classroom management system that is both adaptive and practical, tailored to meet the operational and security requirements of local educational institutions. A robust Internet of Things infrastructure that links sensors, devices, and management systems is essential to the foundation of any smart classroom.

2. Review of Earlier Studies

2.1 The Cyber Threat Landscape

With the rapid growth of digital connectivity and computational power, unauthorized network intrusions, commonly referred to as network attacks, have become increasingly frequent and sophisticated. These cyberattacks target vulnerabilities in network services to gain unauthorized access or disrupt operations, which ultimately endanger the operations of the entire Information Technology infrastructure [2, 3]. Overall, the surge in network-based threats has led to a notable rise in both the scale and complexity of cyber incidents worldwide [3, 4]. With minimal regulation governing the development and deployment of IoT devices, these security and privacy breaches are inevitable [5, 6].

2.2 The Necessity of Explainable AI (XAI) in Intrusion Detection

As cyber threats become increasingly sophisticated, the development of transparent, explainable artificial intelligence (XAI)-based intrusion detection systems (IDS) is a necessity in cybersecurity. The lack of transparency in traditional black-box machine learning models is mainly responsible for their limited applicability in high-risk environments [7]. Explainable artificial intelligence (XAI) has significantly enhanced intrusion detection systems (IDSs) by providing accountability, understanding, and integrity in decision-making. Traditional IDS models, particularly those based on deep learning, frequently function as “black boxes”, complicating security analysts’ understanding of their predictions.

Shapley additive explanations (SHAPs) have emerged as a powerful XAI approach, offering feature attribution scores that elucidate the influence of different input features on model predictions. Employing SHAP in intrusion detection systems helps researchers and practitioners identify significant attack patterns, reduce false positives, and enhance model robustness. Research, including [8], demonstrates that SHAP provides equitable and

consistent feature importance ratings, making it an ideal method for enhancing the interpretability of IDS. Moreover, SHAP facilitates feature selection and optimisation, thereby enhancing generalisation and efficiency in IDSs constructed using FL, where privacy preservation is paramount [9]. SHAP-based explainability succeeds in simplifying the system, improving its classification performance, and delivering stronger interpretability to address weaknesses in current IDS technologies. [10].

2.3 Model Architectures and Performance Implementations

Utilizing SHAP in FL-based IDS provides confidentiality security insights while safeguarding raw data, hence maintaining user privacy and delivering actionable knowledge for cyber defense. A pipeline incorporating Random Forest and XGBoost classifiers, with SHAP, enables a feedback-driven, closed loop mechanism for dynamic model retraining and adaptation to evolving threat landscapes. [11]. The Random Forest and neural networks achieved the highest performance, particularly when trained on a fusion of raw and SHAP features, attaining best accuracy and high sensitivity. In parallel, SHAP-based anomaly detection using cosine and correlation distance metrics proved effective, offering a scalable and interpretable detection solution. [12].

The first joint utilisation of CNNs with both SHAP and LIME for XAI-IDS tasks yields high predictive performance, with higher accuracy, precision, recall, and F1-score. [13]. To enhance transparency and interpretability, studies incorporate Explainable AI (XAI) methods, specifically LIME (Local Interpretable Model agnostic Explanations) and SHAP (SHapley Additive exPlanations). This approach allowed security analysts to understand the importance of both local decision-making and global features, increasing trust in the IDS predictions and aiding in the identification of false positives and complex threats such as zero-day attacks. [14]. Qi Yan [15] used a hybrid explainable artificial intelligence (XAI) framework that integrates Extreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM) for behaviour classification, enhanced by Shapley Additive exPlanations (SHAP) and Local Interpretable Model Agnostic Explanations (LIME) for global and local interpretability. [15]. Additionally, a Malleable Support Vector Machine (MSVM) has been found to be more effective at detecting malicious activities and improving the security of educational content. [16].

3. Research Gaps and Limitations

Despite advancements, the implementation of XAI has been isolated from the model rather than integrated with it, and the experimental scope remains narrow. At present, the most urgent limitation is the lack of an architecture that can ensure privacy, explainability, and effective efficiency. [17].

It is an important mission for educational institutions to have a well-structured plan to prepare their students, starting as early as possible, for emerging technologies such as IoT [18]. Canbaz [19] provided guidelines to set up an in house cybersecurity lab, along with a toolset that is specifically designed and implemented as an open source smart home traffic measurement, visualization, and analysis. Hearon provided guidelines for setting up an in house cybersecurity lab, along with a toolset specifically designed and implemented to measure, visualise, and analyse open source smart home traffic. [20]. Kurni [21] demonstrated IoT implementations in classrooms by emphasising teamwork and privacy protection measures, educating users about security issues, conducting security audits in accordance with regulations, carefully choosing vendors, and encrypting data effectively. Luburi [22] developed Security Design Analysis by utilising case study analysis and the hybrid flipped classroom approach.

In the next section, 4, we outline the detailed methods, dataset, and data processing, followed by an extensive description of results and final inferences in section 5. Finally, we conclude the work with a summary.

4. Methodology

This section presents the methodological framework adopted for analyzing and modeling cybersecurity threats in a smart classroom environment. It encompasses dataset description, preprocessing steps, feature engineering, model development, and evaluation strategies.

4.1 Dataset Description

This study utilizes the *Classroom Data Security Threats Dataset* [24], a simulated dataset designed to represent cybersecurity incidents in a digital educational ecosystem. The dataset models realistic scenarios in which sensitive student information and system integrity are protected against various cyber threats.

It includes multiple categories of threats, such as unauthorized access, data breaches, phishing attacks, malware/spyware, and insider threats, thereby capturing both external attacks and internal vulnerabilities. Each record represents an individual incident, providing detailed information about the nature of the threat, system response, and user interaction context.

The dataset is stored in CSV format (*english_classroom_security_threats.csv*, ~90 KB) and contains multiple features describing threat characteristics, system performance, and user privileges. Its structure supports both descriptive analytics and predictive modeling tasks.

4.2 Feature Description

Table 1 summarizes the key features used in this study, along with their descriptions and data types.

Feature Name	**Description**	**Type**
Threat_Type	Category of cyber threat (e.g., phishing, insider threat, malware)	Categorical
Threat_Detected	Indicates whether the threat was detected (1 = Yes, 0 = No)	Binary
Threat_Severity	Severity level of the threat (Low, Medium, High)	Categorical
Response_Time	System response time (e.g., “120 ms”)	Text/Numeric
Response_Time_ms	Processed numeric response time in milliseconds	Numerical
Access_Level	User privilege level (Admin, User, Guest)	Categorical
Data_Encryption	Indicates whether data was encrypted (Yes/No)	Binary
Phishing_Attempt	Presence of phishing attempt (1 = Yes, 0 = No)	Binary
Malware_Detected	Presence of malware/spyware activity (1 = Yes, 0 = No)	Binary
Insider_Threat	Indicator of insider threat activity (1 = Yes, 0 = No)	Binary

Data_Breach	Indicator of data breach occurrence (1 = Yes, 0 = No)	Binary

Table 1. Summary of Dataset Features

4.3 Data Preprocessing

Prior to analysis and model development, several preprocessing steps were performed to ensure data quality and consistency:

• Data Cleaning: Missing or inconsistent entries were examined and handled appropriately to maintain dataset integrity.

• Feature Transformation: The *Response_Time* attribute, originally in string format (e.g., "212 ms"), was converted into a numerical variable (*Response_Time_ms*) to enable quantitative analysis.

• Categorical Encoding: Categorical variables such as *Threat_Type*, *Threat_Severity*, and *Access_Level* were encoded using appropriate techniques (e.g., label encoding or one-hot encoding) for compatibility with machine learning models.

• Binary Conversion: Boolean features (e.g., phishing attempts, malware detection) were standardized into binary format (0/1).

• Feature Selection: Non-informative or identifier-based attributes were excluded to prevent bias and overfitting.

These preprocessing steps ensured that the dataset was structured and suitable for both statistical analysis and predictive modeling.

4.4 Exploratory Data Analysis (EDA)

An exploratory analysis was conducted to understand the distribution and relationships among variables. This included:

• Frequency analysis of different threat types

• Distribution of threat severity levels

• Analysis of specific threat indicators (phishing, malware, insider threats, data breaches)

• Examination of threat occurrence across access levels

• Temporal trend analysis of threat events

The insights derived from EDA informed feature relevance and guided the design of the predictive modeling framework.

4.5 Predictive Modeling Approach

A supervised machine learning approach was adopted to predict the binary outcome variable *Threat_Detected*. A tree-based classification model was employed due to its ability to capture nonlinear relationships and interactions among features.

Key aspects of the modeling approach include:

- Input Features: All relevant non-identifier features, including behavioral, system, and threat indicators
- Target Variable: *Threat_Detected* (1 = threat detected, 0 = not detected)
- Model Type: Tree-based classifier (e.g., Random Forest / Gradient Boosting)
- Training Strategy: Dataset split into training and testing subsets to evaluate generalization performance

Tree-based models are particularly suitable for cybersecurity applications due to their robustness, interpretability, and ability to handle heterogeneous data types.

4.6 Explainable AI (XAI) Integration

To enhance model transparency, SHapley Additive exPlanations (SHAP) were employed as an Explainable AI technique. SHAP provides both global and local interpretability by quantifying each feature's contribution to the model's predictions.
The following SHAP analyses were conducted:

- Global Interpretation: Summary (beeswarm) plot to identify key features influencing predictions
- Feature Importance: Bar plot of mean absolute SHAP values
- Feature Effect Analysis: Dependence plot for examining nonlinear relationships
- Local Explanation: Waterfall plot for interpreting individual predictions

This integration ensures that the model is not only accurate but also interpretable, which is essential for cybersecurity decision-making.

This structured approach ensures a comprehensive analysis of cybersecurity threats, combining statistical insights with machine learning and explainable AI techniques.

5. Results and Discussion

5.1 Distribution of Cyber Threat Types

Frequency of Different *Threat_Type*

Threat Type	Count
Insider Threat	218
Phishing	204
Unauthorized Access	203
Data Breach	193
Malware/Spyware	182

Table 2. Distribution of Cyber Threat Types

Distribution of identified cyber threat types within the dataset, showing the frequency of each category, including insider threats, phishing, unauthorized access, data breaches, and malware/spyware incidents. The table highlights the relative prevalence of different threat vectors in the analyzed environment.

The distribution of cyber threat categories (Table 2) reveals distinct patterns in the occurrence of security incidents within the classroom environment. Among the identified threats, *insider threats* constitute the largest proportion (n = 218), followed by *phishing attacks* (n = 204), *unauthorized access incidents* (n = 203), *data breaches* (n = 193), and *malware/spyware detections* (n = 182).

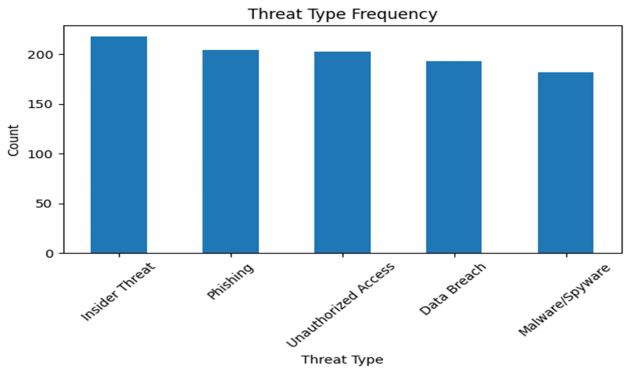

Figure 1. Threat Type Distribution

Figure 1 shows the comparative frequency of different cyber threats, highlighting the dominance of insider threats and phishing attacks.

This distribution highlights the predominance of human-centric and internally driven threats over purely technical attacks. The high frequency of insider threats suggests that vulnerabilities associated with user behavior, access misuse, and privilege exploitation are critical concerns in educational digital ecosystems. Similarly, the prevalence of phishing attacks underscores users' susceptibility to social engineering, which often serves as an entry point for more severe compromises.

5.2 Threat Severity Distribution

The dataset exhibits a relatively balanced distribution across threat severity levels, with low-severity incidents accounting for 348 cases, medium-severity for 336, and high-severity for 316.

Severity	Count
Low	348
Medium	336
High	316

Table 3. Threat Severity Distribution

Distribution of cyber threats (Table 3) across severity levels (low, medium, and high), indicating the number of incidents in each category. This table demonstrates the dataset's balanced nature, supporting robust multi-class classification and severity analysis.

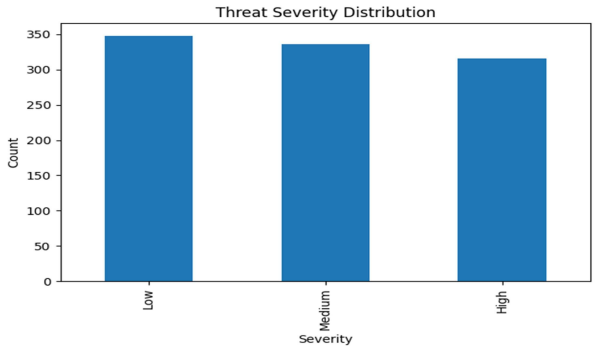

Figure 2. Threat Severity Distribution

Figure 2 illustrates the balanced distribution across low, medium, and high severity levels, supporting robust modeling.

This near-uniform distribution indicates that the dataset is well-suited for multi-class classification and severity prediction tasks, avoiding class imbalance issues commonly observed in cybersecurity datasets. From a practical perspective, the presence of a significant proportion of high-severity threats (31.6%) is noteworthy, as it reflects a realistic threat landscape in which critical incidents are not rare.

Furthermore, the comparable distribution of medium and high severity threats suggests that escalation pathways from minor anomalies to critical breaches may exist, warranting further temporal and causal investigation.

5.3 Analysis of Specific Threat Indicators

Binary threat indicators provide deeper insights into the prevalence of specific attack vectors. Insider threat flags were observed in 218 instances, making them the most prominent individual threat indicator. Phishing attempts were recorded in 204 cases, followed by data breaches (193) and malware detections (182).

Count of Specific Threat Indicators

Phishing Attempts

- Yes (1): 204
- No (0): 796

Malware Detections

- Yes (1): 182
- No (0): 818

Insider Threats

- Yes (1): 218

• No (0): 782

Data Breaches

• Yes (1): 193

• No (0): 807

Table 4. Frequency of Specific Threat Indicators

Binary distribution of key threat indicators, including phishing attempts, malware detections, insider threats, and data breaches. The table presents counts of presence (1) and absence (0), providing insight into the occurrence of specific attack vectors. (Table 4)

These findings reinforce the observation that behavioral and social engineering threats outweigh purely technical exploits in this dataset. The relatively low rate of malware detections may indicate either effective endpoint protection mechanisms or a shift in attacker strategies toward less-detectable methods, such as credential theft and privilege misuse.

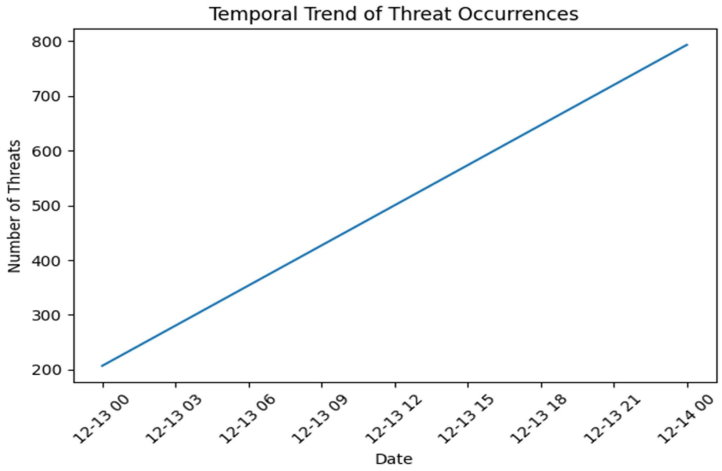

Figure 3. Temporal Trend of Threat Occurrences

Figure 3 depicts how threats evolve over time, useful for identifying peak attack periods and temporal patterns.

The presence of 193 data breach incidents is particularly significant, as it reflects tangible consequences of security failures, potentially involving unauthorized data exposure or exfiltration. This emphasizes the need for robust data protection and monitoring mechanisms within digital learning environments.

5.4 Access Level and Threat Occurrence

An analysis of threat occurrence across different access levels reveals important security implications. Administrative accounts recorded 176 threat instances, slightly higher than standard user accounts (180) and guest accounts (150). While the difference between admin and user accounts is marginal, both categories exhibit more detected threats than guest users.

Access Level	vs Threat	Occurrence
Admin	160	176
Guest	160	150
User	174	180

Table 5. Access Level vs Threat Occurrence

Cross-tabulation of user access levels (admin, user, and guest) against threat detection outcomes. The table illustrates the relationship between privilege levels and the frequency of detected threats, highlighting potential vulnerabilities associated with higher access privileges.

This trend suggests that privileged and semi-privileged accounts constitute critical attack surfaces, likely due to their broader access rights and system-control capabilities. The elevated threat exposure among administrative accounts is particularly concerning, as compromises at this level can lead to widespread system impact.

Interestingly, the relatively high number of threats associated with standard user accounts suggests that attackers may be leveraging lateral movement, initially targeting less-protected accounts before escalating privileges. Guest accounts, although less frequently targeted, still demonstrate a non-negligible level of threat activity, highlighting that no access tier is entirely secure.

5.5 Discussion

Collectively, the results indicate that the cybersecurity landscape within the analyzed environment is heavily influenced by user behavior, access privileges, and social engineering vulnerabilities. The dominance of insider threats and phishing attacks suggests that technical defences alone are insufficient without complementary behavioural monitoring and user awareness mechanisms.

The balanced severity distribution further implies that the system experiences a wide spectrum of threat intensities, from minor anomalies to critical incidents. This reinforces the importance of adaptive threat detection systems capable of handling multi-level risk scenarios.

Moreover, the relationship between access level and threat occurrence underscores the necessity of implementing role-based access control (RBAC), adhering to the principle of least privilege, and continuously monitoring high-privilege accounts. The findings also suggest potential benefits from integrating anomaly detection and user behavior analytics (UBA) to proactively identify suspicious activities.

5.6 Implications for Cybersecurity Frameworks

The empirical findings from the dataset provide critical insights into the design and enhancement of modern cybersecurity frameworks, particularly in smart educational environments and digitally mediated learning systems. The observed predominance of insider threats and phishing attacks highlights the necessity of transitioning from traditional perimeter-based security models toward behavior centric and adaptive security architectures.

First, the high incidence of insider threats underscores the importance of integrating User and Entity Behavior Analytics (UEBA) into cybersecurity frameworks. Conventional rule based detection systems are often insufficient for identifying subtle deviations in legitimate user behavior. Therefore, incorporating machine-learning driven behavioural profiling can enable detection of anomalous access patterns, privilege misuse, and lateral movement within systems. Such approaches are particularly relevant in environments where users (students, faculty, administrators) interact dynamically with digital resources.

Second, the significant presence of phishing attacks indicates a persistent vulnerability to social engineering exploits, necessitating a dual-layered defense strategy. On the technical side, frameworks should incorporate advanced email filtering, URL inspection, and real-time threat intelligence integration. On the human side, continuous cybersecurity awareness training and simulated phishing exercises are essential to reduce susceptibility. This socio-technical approach ensures that both system-level defenses and user cognition are addressed.

Third, the analysis of access-level vulnerabilities reveals that administrative and user accounts constitute major attack surfaces. This finding strongly supports enforcing Role-Based Access Control (RBAC) and the Principle of Least Privilege (PoLP). Additionally, frameworks should incorporate Privileged Access Management (PAM) solutions to monitor, restrict, and audit high-level account activities. Real-time logging and alerting mechanisms for privileged actions can significantly reduce the risk of catastrophic system compromise.

Fourth, the presence of a considerable number of data breach incidents emphasizes the need for data-centric security models. This includes the adoption of encryption (both at rest and in transit), data loss prevention (DLP) systems, and continuous monitoring of data access. Frameworks should also integrate zero-trust architectures, in which every access request is continuously verified regardless of its origin, whether inside or outside the network.

Furthermore, the balanced distribution of threat severity levels suggests the feasibility of implementing AI-driven risk-scoring and severity-prediction models. Machine learning algorithms can be trained on such datasets to classify threats in real time, prioritize incident response, and optimize resource allocation. This capability is particularly valuable in environments with high volumes of heterogeneous security events.

Finally, the findings advocate for the adoption of a multi-layered defense strategy, combining:

- Behavioral analytics for insider threat detection
- Social engineering mitigation mechanisms
- Privileged access monitoring and control
- Data-centric protection strategies
- AI-based threat intelligence and predictive analytics

Such an integrated and adaptive cybersecurity framework can significantly enhance resilience, reduce detection

latency, and improve overall incident response effectiveness in evolving threat landscapes.

5.7 Explainable AI (XAI) Interpretation of Threat Detection Model

Building upon the statistical and exploratory analyses presented in Sections 5.1–5.6, this section introduces an Explainable AI (XAI) perspective to further interpret the predictive modeling outcomes. While earlier sections established the distribution of threat types, severity levels, and access-level vulnerabilities, they primarily provided descriptive and inferential insights. In contrast, this section focuses on model-driven interpretability, revealing how individual features contribute to the automated prediction of *Threat_Detected*.

The motivation for incorporating XAI arises directly from the findings in Section 5.5 (Integrated Discussion) and Section 5.6 (Implications for Cybersecurity Frameworks), which emphasize the need for intelligent, transparent, and adaptive threat detection systems. In this context, SHapley Additive exPlanations (SHAP) are employed to bridge the gap between predictive performance and interpretability, enabling both global and local understanding of model behavior.

5.7.1 Linking Feature Importance to Observed Threat Patterns

The SHAP summary plot (Fig. 1) provides a global interpretation of feature importance, complementing the empirical observations reported in Section 5.1 (Threat Type Distribution) and Section 5.3 (Specific Threat Indicators). The dominance of behavioural and interaction-driven threats, such as insider threats and phishing, identified earlier, is reflected in the model's reliance on features associated with system usage patterns.

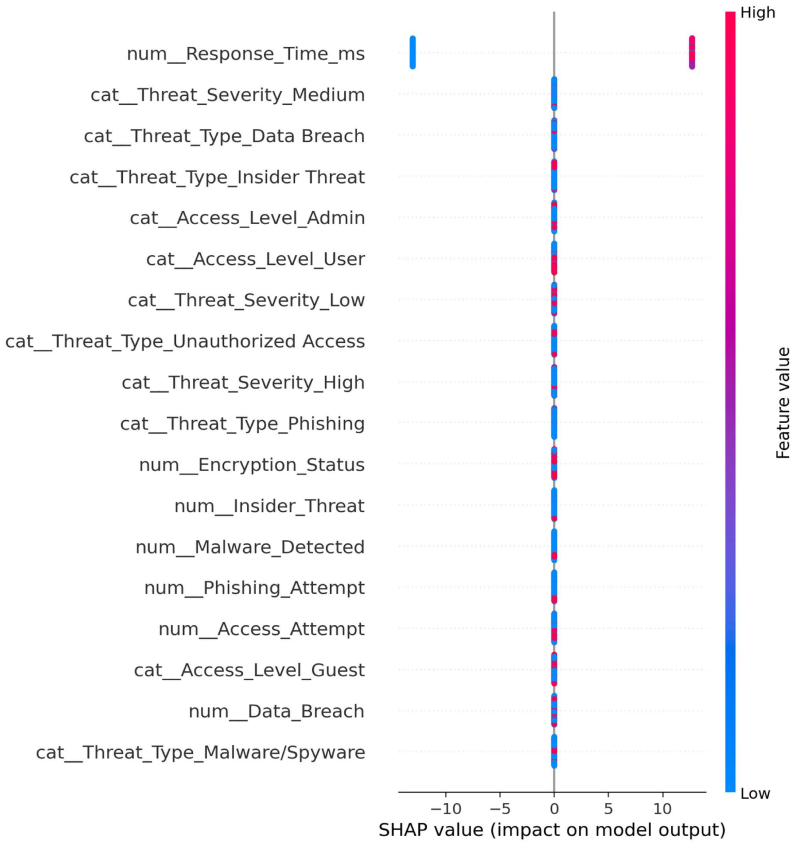

Figure 4. SHAP Summary (Beeswarm) Plot

SHAP summary (beeswarm) plot illustrating the global impact of features on the prediction of *Threat_Detected* (Figure 4*)*. Features are ranked by their mean absolute SHAP values, indicating overall importance. Each point represents an individual observation, with color encoding the feature value (low to high). The horizontal spread reflects the magnitude and direction of each feature's contribution to the model output.

In particular, *Response_Time_ms* emerges as the most influential feature in the SHAP analysis. This finding aligns with the temporal trends observed in Section 5.3, where variations in system activity over time were linked to threat occurrences. Increased response times may indicate abnormal system behaviour, such as resource contention, malicious execution, or delayed processing due to unauthorised activities.

Thus, the XAI results reinforce the earlier conclusion that cyber threats are not solely defined by categorical labels (e.g., phishing or malware), but are strongly associated with dynamic system behavior and performance anomalies.

Figure 5. SHAP Feature Importance (Bar) Plot

SHAP feature importance bar plot showing the mean absolute SHAP values for each feature (Figure 5). This aggregated view highlights the relative contributions of features to the model's predictive performance, providing a clear ranking of the most influential variables for threat detection outcomes.

5.7.2 Relationship with Threat Severity Distribution

The balanced distribution of threat severity levels discussed in Section 5.2 is further validated by the SHAP

feature importance bar plot. The presence of multiple contributing features with comparable importance suggests that the model effectively captures patterns across low, medium, and high severity threats.

Rather than being biased toward a specific severity class, the model leverages a diverse set of features to distinguish between varying threat intensities. This supports the earlier assertion that the dataset is well-suited for multi-class classification and severity prediction tasks, and demonstrates that the learned decision boundaries are not dominated by a single feature or class imbalance.

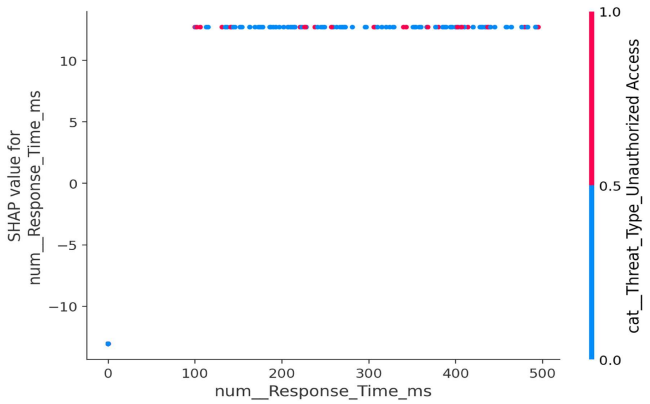

Figure 6. SHAP Dependence Plot for Response Time

SHAP dependence plot illustrating the relationship between *Response_Time_ms* and its contribution to the prediction of *Threat_Detected*. (Figure 6). The plot reveals how variations in response time influence the model output, capturing potential nonlinear effects and interaction patterns with other features.

5.8 Interpreting Access-Level Vulnerabilities through Feature Interactions

Section 5.4 (Access Level and Threat Occurrence) highlighted the elevated risk associated with administrative and user accounts. The SHAP dependence plot for *Response_Time_ms* (Fig. 6) provides additional insight into how such vulnerabilities manifest in the model's decision-making process.

The observed nonlinear relationship between response time and prediction output suggests that performance anomalies may be indicative of privilege misuse or unauthorized access activities, particularly in high-access accounts. Furthermore, the spread of SHAP values for similar response times indicates interaction effects with other features, potentially including access level, user behavior, or concurrent system events.

This supports the hypothesis proposed in Section 5.4 that attackers may exploit user accounts and subsequently escalate privileges, with such activities leaving detectable traces in system performance metrics.

SHAP waterfall plot explaining an individual high-risk prediction. The figure demonstrates how feature contributions incrementally shift the model output from the baseline (expected value) to the final predicted probability of threat detection, highlighting the most influential factors for that specific instance.

5.9 Alignment with Integrated Cybersecurity Insights

The local explanation provided by the SHAP waterfall plot (Fig. 7) complements the integrated discussion presented in Section 5.5. While Section 5.5 emphasized the combined influence of user behavior, social

engineering, and access privileges, the waterfall plot demonstrates how these factors are operationalized within the model for a specific prediction.

For a high-risk instance, the model aggregates contributions from multiple features, most notably *Response_Time_ms*, to arrive at the final threat probability. This illustrates the multifactor nature of cyber threat detection, where no single feature is solely responsible; rather, a combination of signals determines the outcome.

Such instance-level interpretability is critical for real-world cybersecurity operations, where analysts must understand *why* a particular alert was generated before taking action.

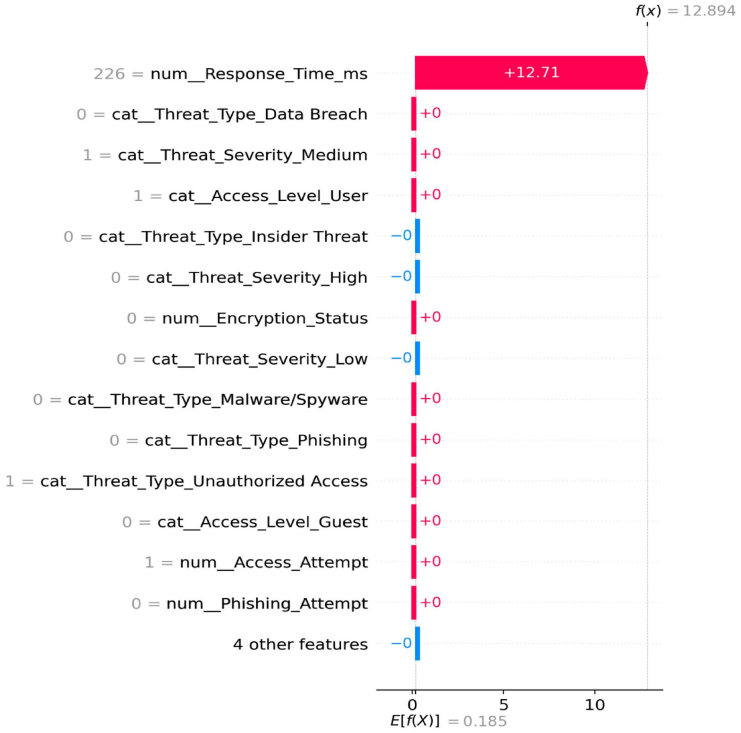

Figure 7. SHAP Waterfall Plot
(Single Instance Explanation)

5.10 Implications for Framework Design and Deployment

The XAI findings provide strong support for the framework recommendations outlined in Section 4.6. Specifically:

• The importance of behavioural and temporal features validates the integration of User Behaviour Analytics (UBA) and real-time monitoring systems.

• The interaction effects observed in SHAP analysis reinforce the need for context-aware and multi-dimensional threat detection models.

• The ability to explain individual predictions aligns with the requirement for transparent and accountable AI systems in cybersecurity frameworks.

Moreover, the consistency between empirical observations (Sections 5.1–5.4) and model-based explanations

(Section 5.8) demonstrates that the proposed approach achieves both analytical coherence and practical relevance.

5.11 Summary of Findings

This study provides a comprehensive analysis of cyber threat patterns within the examined dataset, yielding several important findings with both theoretical and practical implications.

Firstly, insider threats are the dominant attack category, indicating that internal vulnerabilities, whether intentional or accidental, pose a greater risk than external technical exploits. This finding reinforces the growing recognition that human factors are central to cybersecurity risk.

Secondly, the distribution of threat severity levels is relatively balanced across low, medium, and high categories. This uniformity enhances the dataset's suitability for developing and evaluating multi-class classification models and supports the design of robust threat detection systems capable of handling diverse risk levels.

Thirdly, phishing and social engineering attacks remain critical entry points for system compromise. Their high frequency highlights persistent gaps in user awareness and the effectiveness of deceptive attack techniques, emphasizing the need for continuous education and adaptive filtering mechanisms.

Fourthly, the analysis of access-level vulnerabilities reveals that both administrative and standard user accounts are primary targets for cyber threats. This suggests that attackers may employ strategies involving initial compromise of user accounts followed by privilege escalation, thereby exploiting hierarchical access structures.

Fifthly, the occurrence of numerous data breaches signifies real-world impact, extending beyond theoretical vulnerabilities to tangible consequences such as data exposure and system compromise. This underscores the importance of proactive monitoring and rapid incident response mechanisms.

In summary, the findings collectively demonstrate that:

- Cybersecurity risks are heavily influenced by user behavior and access privileges
- Social engineering continues to be a dominant attack vector
- Balanced threat severity supports advanced analytical modeling
- Privileged accounts represent critical points of vulnerability
- Data breaches highlight the need for strong data protection strategies

These insights contribute to a deeper understanding of contemporary cyber threat dynamics and provide a foundation for developing intelligent, adaptive, and human-aware cybersecurity solutions.

5.12 Concluding Integration

In summary, the Explainable AI analysis serves as a unifying layer that connects descriptive statistics, behavioural insights, and predictive modelling. While Sections 4.1–4.6 establish *what* patterns exist in the data, this section explains *how and why* these patterns influence automated threat detection.

The integration of SHAP-based interpretability confirms that:

- Behavioral anomalies and system performance indicators are key drivers of threat detection
- Threat severity and occurrence patterns are effectively captured by the model
- Access-level risks and user activities are reflected in feature interactions
- Transparent AI models can enhance both trust and operational effectiveness

This seamless linkage between statistical analysis and explainable modeling strengthens the overall contribution of the study, positioning it as a comprehensive and interpretable approach to cybersecurity analytics in smart environments.

References

[1] Rahman, M. M., Maruf, A. A., Hassan, M. I. J. et al. (2026). Next-generation learning environments: comprehensive implementation of IoT-enabled smart classrooms with advanced security system integration. *Wireless Netw* 32, 409–432.

[2] Alabdulatif, A. A. (2025). Novel Ensemble of Deep Learning Approach for Cybersecurity Intrusion Detection with Explainable Artificial Intelligence. *Appl. Sci.* 2025, 15, 7984.

[3] Perez, S. I., Criado, R. (2022). Increasing the Effectiveness of Network Intrusion Detection Systems (NIDSs) by Using Multiplex Networks and Visibility Graphs. *Mathematics* 2022, 11, 107.

[4] aimultiple.com (2025).100+ Network Security Statistics in 2025. AIMultiple. Available online: https://research.aimultiple.com/network-security-statistics/.

[5] Radanliev, et al. P (2019). Definition of Internet of Things (IoT) Cyber Risk Discussion on a Transformation Roadmap for Standardisation of Regulations Risk Maturity Strategy Design and Impact Assessment, ArXiv Prepr. ArXiv190312084.

[6] Zucchetto, D., Zanella, A, (2017). Uncoordinated access schemes for the IoT: approaches, regulations, and performance, *IEEE Commun. Mag.*, 55 (9) 48–54.

[7] Goyal, S. B., Thakur, Narina, Qadeer., Shaik, A. (2025). Bridging the Interpretability Gap: A SHAP-Enhanced Framework for Intrusion Detection in Cybersecurity, *International Journal of Management and Data Intelligence.* Vol. 1 (1).

[8] Lundberg, S. (2017). A unified approach to interpreting model predictions. arXiv 2017, arXiv:1705.07874.

[9] Dong, T., Li, S., Qiu, H., Lu, J. (2022). An interpretable federated learning-based network intrusion detection framework. arXiv:2201.03134.

[10] Assudani, Purshottam J., Pavan Kumar, N. V. S., Mohanambal, K., Chitra, R. (2025). Explainable Artificial Intelligence Driven Intrusion Detection System for Enhancing Reliability and Interpretability in IoT Based Network Security Solutions. *Journal of Intelligent Systems Internet of Things*, 2025, 17 (1). p. 219.

[11] Myakala, R. K., Jagirapu, A. R., Podduturi, V. R., Nagata, P. K., Lavudya S., Pedhapally, S. K. (2025). AI-Powered Intrusion Detection with SHAP Explainability and Feedback Loop: A Modular Pipeline for Cyber Threat Classification, *In: 2025 IEEE International Carnahan Conference on Security Technology (ICCST), San Antonio, TX, USA*.

[12] Petihakis, Georgios. (2025). Bridging security and interpretability in AI: a SHAP-centric framework for adversarial attack detection, Master's Thesis, 72 p. https://dione.lib.unipi.gr/xmlui/handle/unipi/17952.

[13] Salloum, S., Norozpour, S. (2025). XAI-IDS: A Transparent and Interpretable Framework for Robust Cybersecurity Using Explainable Artificial Intelligence. *SHIFRA*, 2025, 69-80.

[14] Alatawi, Mohammed Naif. (2025). Enhancing Intrusion Detection Systems with Advanced Machine Learning Techniques: An Ensemble and Explainable Artificial Intelligence (AI) *Approach, Security Privacy*, 8 (1) p. e496.

[15] Yan, Qi (2025). Explainable Machine Learning for Detecting Malicious Student Behavior in Campus Networks, *In*: ICCSIE '25: Proceedings of the 10th International Conference on Cyber Security and Information Engineering. P. 299 - 3 *Association for Computing Machinery. New York, NY, USA*.

[16] Jia, L. (2025). Research on English Classroom Data Security Protection Technology Based on Intelligent Algorithm. *International Journal of High Speed Electronics and Systems,* 2540465.

[17] Fatema, K., Dey, S. K., Anannya, M., Khan, R.T., Rashid, M. M., Su, C., Mazumder, R. (2025). Federated XAI IDS: An Explainable and Safeguarding Privacy Approach to Detect Intrusion Combining Federated Learning and SHAP. *Future Internet* 17, 234.

[19] Campanile, L., Iacono, M., Marulli, F., Mastroianni, M. (2020). Privacy Regulations Challenges on Data-centric and IoT Systems: A Case Study for Smart Vehicles., *In*: *IoTBDS*, 2020, p. 507–518.

[20] Canbaz, M. A., O Hearon, K., McKee, M., Hossain, M. N. (2021, July). IoT Privacy and Security in Teaching Institutions: Inside the Classroom and Beyond, *In*: 2021 ASEE Virtual Annual Conference Content Access, *Virtual Conference*. 10.18260/1-2—37407.

[21] Hearon, K. O., McKee, M., Hossain, N., & Canbaz, M. A. (2021). IoT privacy and security in teaching institutions: Inside the classroom and beyond. *In*: *American Society for Engineering Education*.

[22] Kurni, M., K. G., S. (2025). Secure IoT-Based Classroom. *In*: The Internet of Educational Things. *Springer,* Cham.

[23] Nikola, Luburiæ., Goran, Sladiæ., Jelena, Slivka., Branko, Milosavljeviæ. (2019). A Framework for Teaching Security Design Analysis Using Case Studies and the Hybrid Flipped Classroom, *ACM Transactions on Computing Education (*TOCE) 19 (3) Article No.21, 1 – 19.

[24] https://www.kaggle.com/datasets/ziya07/classroom-data-security-threats-dataset?resource=download