Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

An Explainable AI Framework For Cyber Threat Detection and Risk Assessment in Intelligent Networks

Nithin Reddy Gadicharla[1] , Abhilash Koka[2]

[1]Database Administrator Independent Researcher Celina, USA. , email; nithin.gadicharla@ieee.org
[2]Technical Manager Cognizant Technology Solutions Charlotte, USA. , email: abhilashkoka4@gmail.com

Abstract
This study aims to develop an explainable artificial intelligence framework for cyber-threat detection and risk assessment in intelligent networks. The framework combines machine-learning classification with SHAP-based interpretation to improve the transparency of automated cybersecurity decisions. The objective was to evaluate and compare Logistic Regression, Random Forest, XGBoost, and Multilayer Perceptron models for distinguishing benign and malicious network traffic, identify the most influential network features using SHAP analysis, and assess detected threats using a model-derived risk score. The analysis was conducted using labeled network-flow data from the CIC-IDS2017 dataset. Among the evaluated approaches, XGBoost demonstrated the strongest overall classification performance. SHAP analysis identified several network-flow characteristics that substantially influenced the model's predictions, providing insight into its decision-making process. Threat-specific analysis further demonstrated variation in detection performance and risk classification across different attack categories, with most categories receiving high model-derived risk levels and selected categories receiving medium risk classifications. The findings demonstrate that integrating machine-learning detection with explainability and risk assessment can provide a more informative approach to intelligent network security. The proposed framework combines predictive capability with feature-level interpretation and threat-oriented risk information, potentially supporting transparent cybersecurity analysis and informed security decision-making.

Keywords: Artificial intelligence, Cybersecurity, Explainable AI, Intrusion detection, Risk assessment.

1. Introduction

The rapid expansion of intelligent networks, cloud platforms, Internet of Things (IoT) devices, 5G infrastructures, and connected cyber-physical systems has increased the complexity of modern digital environments. At the same time, cyber threats are becoming more diverse and increasingly difficult to distinguish from legitimate network activity (Almheiri et al., 2025). Conventional security mechanisms that depend mainly on predefined signatures may therefore struggle to recognize previously unseen or rapidly changing attacks. Artificial intelligence (AI) and machine learning (ML) provide an opportunity to analyse large volumes of network-traffic data and identify patterns associated with malicious behaviour. Recent research has consequently explored AI-enabled threat detection as a means of improving the speed, adaptability, and scalability of cybersecurity operations (Dhanushkodi and Thejas, 2024). However, high predictive performance alone is insufficient for security-critical environments. Security analysts also need to understand why a system has identified an activity as threatening and how that prediction should influence subsequent risk-related decisions.

Intelligent networks generate heterogeneous and continuously changing traffic, creating substantial challenges for reliable threat identification. Network intrusion detection systems (IDSs) have consequently evolved from conventional rule-based mechanisms toward data-driven approaches capable of learning characteristics associated with malicious traffic (Nwakanma et al., 2023). Research in IoT, industrial systems, connected vehicles, transportation networks, and 5G environments demonstrates the growing application of AI and deep learning for detecting network attacks and abnormal behaviour (Khan et al., 2021; Albashayreh et al., 2025). Nevertheless, intelligent detection systems can produce complex decision boundaries that are difficult for

security personnel to interpret. This issue becomes particularly important when network decisions involve potentially serious consequences, such as blocking traffic, prioritising incidents, or allocating security resources (Oseni et al., 2022). Thus, effective cyber-threat detection requires not only accurate classification but also meaningful interpretation of model decisions.

AI and ML techniques have become important components of contemporary cybersecurity because they can learn relationships within network data without requiring every threat pattern to be manually encoded (Alshudukhi et al., 2025). Supervised learning methods can classify network flows according to known threat labels, while ensemble and neural-network approaches can capture nonlinear relationships among traffic characteristics (Nukala et al., 2026). Their application has been investigated across conventional enterprise networks as well as specialised environments such as industrial IoT and software-defined networking (Prasad et al., 2025). However, complex models may function as black boxes, making their outputs difficult to justify to analysts and other decision-makers. This limitation has encouraged the integration of explainable artificial intelligence (XAI) with cyber-threat detection. XAI provides mechanisms for examining the contribution of individual features to model predictions and can therefore support more transparent interpretation of AI-assisted security decisions (Capuano et al., 2022; Alketbi and Mehmood, 2025).

Explainability is particularly valuable in cybersecurity because detection outcomes frequently require investigation, validation, and human intervention (Gupta and Misra, 2025). An unexplained prediction may indicate that an attack has been detected but does not reveal which characteristics of the network activity contributed to that conclusion (Islam et al., 2025). XAI techniques address this limitation by providing interpretable information about model behaviour and feature contributions (Nalinipriya et al., 2025). Previous studies have examined XAI for intrusion detection, cyber-risk assessment, insider-threat detection, and security decision-making, demonstrating its potential to improve transparency and analyst confidence (Arreche et al., 2024; Prasad et al., 2026). SHAP-based approaches are particularly useful because they can quantify the contribution of individual input features to predictions. However, explainability should extend beyond simply ranking important features (Moustafa et al., 2023). In practical cybersecurity applications, explanations can be connected with threat severity and risk assessment to provide more actionable information for security decision-making.

Although existing research has established the usefulness of AI for intrusion and cyber-threat detection, important limitations remain. A substantial portion of the literature concentrates on improving classification performance, whereas comparatively less attention is given to connecting model predictions, interpretable explanations, and systematic risk assessment within one framework (Mohale and Obagbuwa, 2025). Recent studies have increasingly investigated explainability, including XAI-enabled intrusion detection and cyber-risk decision-making (Ashfaq and Chowdhury, 2023). Nevertheless, there remains a need for an integrated approach that evaluates multiple AI models, identifies the factors driving the strongest classifier's decisions, and translates threat-specific predictions into understandable risk categories (Mohitkar and Lakshmi, 2025). Furthermore, cybersecurity environments require methods that can support both technical detection and practical interpretation (Haque et al., 2024). Addressing this gap motivates the development of an explainable framework combining machine-learning detection, SHAP-based interpretation, and model-derived cyber-risk assessment.

The study aims to evaluate and compare machine-learning models for cyber-threat detection and to identify the key network features influencing the predictions of the best-performing model using SHAP-based explainability. In addition, it seeks to assess detected cyber threats using a model-derived risk score and integrate threat detection, explainability, and risk assessment into a unified framework for supporting transparent and informed security decision-making in intelligent networks.

2. Materials and Methods

2.1 Research Framework

The proposed study follows an integrated framework combining machine-learning-based cyber-threat detection, explainable artificial intelligence, and threat-oriented risk assessment. The workflow begins with network-flow data obtained from the CIC-IDS2017 dataset, followed by data cleaning, numerical conversion, missing-value treatment, and feature preparation. A balanced analytical sample containing benign and malicious traffic is then used to train and evaluate four classification models: Logistic Regression, Random Forest, XGBoost, and Multilayer Perceptron. The best-performing model is subsequently examined using SHAP

analysis to identify influential traffic characteristics. Finally, threat-category detection results and model-derived prediction probabilities are incorporated into a study-specific risk assessment procedure.

2.2 Dataset and Data Sources

The experimental analysis uses the CIC-IDS2017 dataset developed by the Canadian Institute for Cybersecurity. The dataset contains labeled network-flow records representing benign communication and several common attack scenarios, including denial-of-service, distributed denial-of-service, port scanning, brute-force activity, web attacks, botnet traffic, infiltration, and Heartbleed. The machine-learning CSV files were used as the primary data source because they provide flow-level attributes suitable for supervised classification. The eight available traffic files were combined to create a unified analytical dataset. For model development, 100,000 benign observations and approximately 100,000 malicious observations were sampled using a fixed random seed, producing a near-balanced dataset for subsequent analysis.

2.3 Data Preprocessing and Feature Preparation

Before model development, the network-flow records underwent several preprocessing operations to improve consistency and reduce potential computational problems. Feature columns were converted to numeric representations where required, while infinite values were treated as missing observations. Missing numerical values were subsequently replaced using median imputation. Features containing no variation were removed because constant predictors provide no discriminatory information for classification. The resulting dataset contained 70 usable predictor variables. The target labels were transformed into two classes, namely benign and malicious traffic. The processed observations were divided into training and testing subsets using an 80:20 partition, with the same fixed random state maintained throughout the experimental workflow.

2.4 AI-Based Cyber Threat Detection

Four supervised learning approaches were used to distinguish malicious from benign network traffic. Logistic Regression served as the linear baseline, while Random Forest and XGBoost represented ensemble tree-based methods. A Multilayer Perceptron (MLP) represented a neural-network classifier. Logistic Regression used standardized predictors, whereas the tree-based models used prepared numerical features directly. Random Forest used 200 trees, and XGBoost used 250 estimators, a maximum depth of 6, a learning rate of 0.08, and 0.9 subsampling and column-sampling ratios. The MLP used two hidden layers with 64 and 32 neurons, a batch size of 256, and early stopping. A fixed random state of 42 was used where applicable.

2.5 Explainable AI Framework

Explainability was incorporated after model comparison to investigate how the selected classifier reached its predictions. XGBoost was chosen for interpretation because it achieved the strongest overall classification performance in the experiment. SHAP was applied to a randomly selected subset of 5,000 observations from the test data using a tree-based explanation procedure. Mean absolute SHAP values were calculated to rank features according to their overall contribution to model predictions. Mean SHAP values were also examined to determine the average direction of influence. Positive contributions indicate movement toward the malicious class, whereas negative contributions indicate movement away from that prediction. This procedure provides both feature-importance ranking and directional interpretation.

2.6 Cyber Risk Assessment and Performance Evaluation

Threat-specific risk assessment was performed using the predicted malicious probabilities generated by the XGBoost classifier. The risk score was calculated as: Risk Score (%) = Mean predicted probability of malicious activity × 100. For each threat category, the average predicted malicious probability was converted to a percentage-based risk score. The study classified scores below 50 as low risk, scores from 50 to below 80 as medium risk, and scores of 80 or above as high risk. Detection performance was assessed using accuracy, precision, recall, F1-score, false-positive rate, AUROC, and AUPRC. Threat-category analysis additionally reported the number of test instances, correctly detected instances, and detection rate. These measures jointly evaluate classification effectiveness, discrimination, and threat-specific performance.

3. Results

3.1 Comparative Evaluation of Cyber-Threat Detection Models

The predictive performance of four machine-learning models for cyber-threat classification. XGBoost produced the strongest overall results, achieving 99.90% accuracy and an F1-score of 99.90%, together with 99.94% recall. Random Forest also performed exceptionally well, with 99.85% accuracy and an F1-score of 99.85%, as shown in Table 1. The MLP model achieved 99.26% accuracy, while Logistic Regression showed comparatively lower performance across the evaluated measures. The false-positive rates were particularly low for the ensemble models, indicating effective discrimination between benign and malicious network traffic. Overall, XGBoost provided the most balanced detection performance among the evaluated approaches.

Table 1. Comparative Performance of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FPR (%)	AUROC	AUPRC
Logistic Regression	94.77	94.00	95.67	94.83	6.12	0.982	0.980
Random Forest	99.85	99.87	99.83	99.85	0.13	1.000	0.999
XGBoost	99.90	99.86	99.94	99.90	0.14	1.000	1.000
MLP	99.26	98.99	99.53	99.26	1.02	1.000	1.000

Source: Authors' analysis based on the CIC-IDS2017 dataset (Sharafaldin et al., 2018).

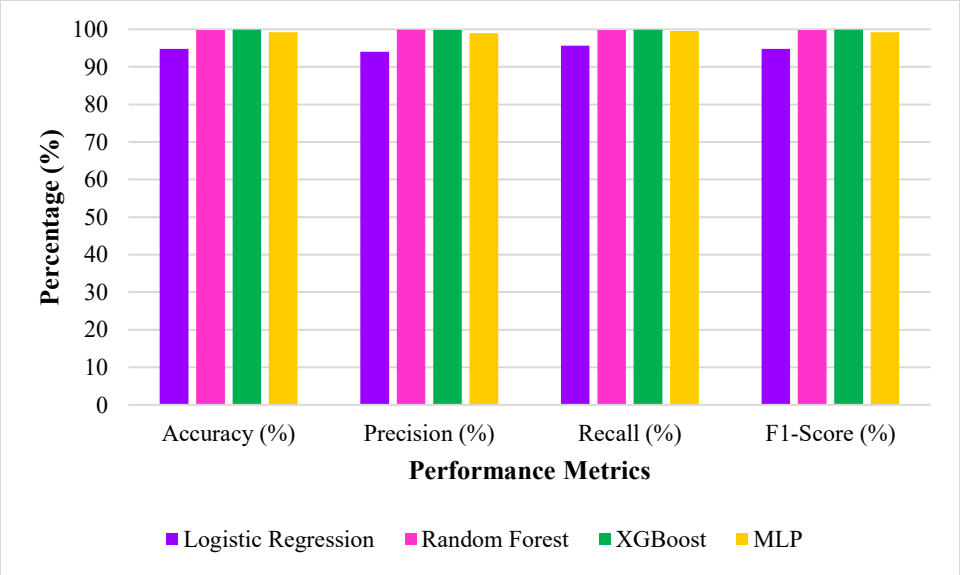

Figure 1. Evaluation of Machine-Learning Models for Cyber-Threat Detection

The comparative results demonstrate clear differences in the predictive behaviour of the four evaluated machine-learning models. Logistic Regression produced the lowest performance across the principal measures, whereas Random Forest, XGBoost, and MLP achieved substantially stronger results. XGBoost provided the highest overall accuracy and recall, while its precision and F1-score remained comparable with those of Random Forest, as shown in Figure 1. The false-positive rate was also very low for the three higher-performing models, indicating limited incorrect identification of benign traffic as malicious. Overall, the results show that ensemble and neural-network approaches performed more effectively than the linear baseline for the evaluated cyber-threat detection task.

3.2 Explainable AI-Based Feature Importance

The ten most influential features identified through SHAP analysis of the XGBoost model. Destination Port recorded the highest mean absolute SHAP value, indicating the greatest contribution to model predictions among the examined variables, as shown in Table 2. Bwd Packet Length Std ranked second, followed by Init_Win_bytes_forward and Init_Win_bytes_backward. Several features showed positive mean SHAP values, whereas others had negative mean contributions, demonstrating that their effects on predictions varied

according to the observed network traffic patterns. The results provide an interpretable view of the model's decision process by identifying the network characteristics that contributed most strongly to cyber-threat classification.

Table 2. SHAP-Based Feature Importance for the XGBoost Model

Rank	Feature	Mean Absolute SHAP	Mean SHAP	Influence
1	Destination Port	1.7642	0.0369	Positive
2	Bwd Packet Length Std	1.6695	0.8028	Positive
3	Init_Win_bytes_forward	1.3193	-0.4664	Negative
4	Init_Win_bytes_backward	0.9060	-0.0380	Negative
5	Average Packet Size	0.8919	0.0270	Positive
6	Bwd Packet Length Min	0.6657	-0.1144	Negative
7	min_seg_size_forward	0.4778	0.0109	Positive
8	Bwd Packet Length Mean	0.4468	-0.1434	Negative
9	Fwd IAT Total	0.4406	-0.0358	Negative
10	Bwd Header Length	0.3636	-0.1021	Negative

Source: Authors' SHAP analysis based on the CIC-IDS2017 dataset; SHAP methodology follows Lundberg and Lee (2017).

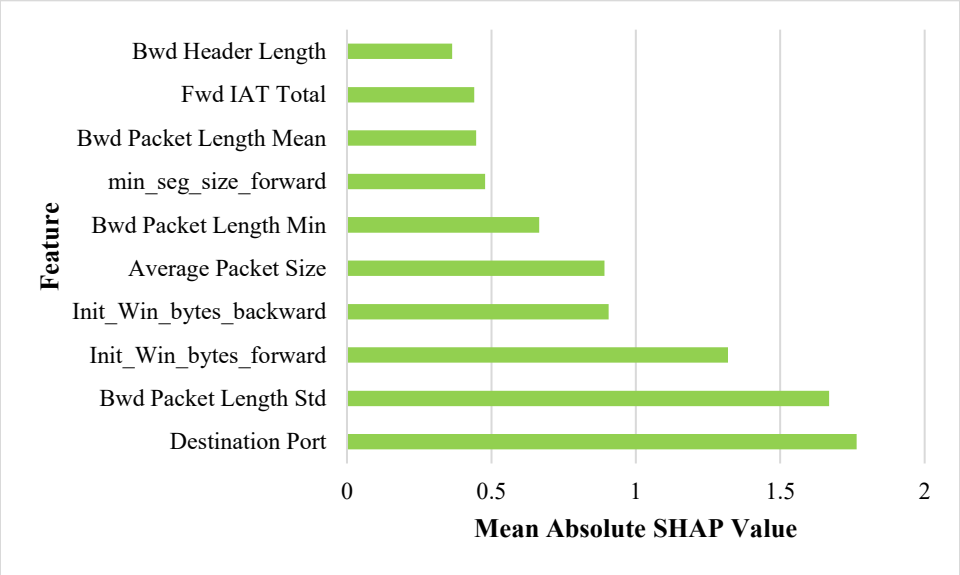

Figure 2. Interpretation of Feature Contributions Using SHAP Analysis

The SHAP analysis identifies the network-flow characteristics that contributed most strongly to the XGBoost model's predictions. Destination Port showed the highest overall contribution, followed by Bwd Packet Length Std and Init_Win_bytes_forward. Init_Win_bytes_backward and Average Packet Size also demonstrated relatively substantial influence, as shown in Figure 2. Other influential variables included Bwd Packet Length Min, min_seg_size_forward, Bwd Packet Length Mean, Fwd IAT Total, and Bwd Header Length. The ranking indicates that transport-level and packet-flow characteristics played an important role in distinguishing malicious from benign traffic. These findings improve model transparency by showing which network attributes were most relevant to the classifier's decision-making process.

3.3 Threat Detection and Risk Assessment

The threat-specific detection outcomes and the corresponding risk assessment are generated by the proposed framework. DoS, PortScan, Brute Force, and Web Attacks achieved complete detection in the evaluated test instances, while DDoS showed a marginally lower detection rate as shown in Table 3. Botnet detection remained high despite a small number of missed instances. Infiltration produced the lowest detection rate among the reported categories and was assigned a medium risk level. Heartbleed also received a medium risk classification. In contrast, the remaining threat categories were classified as high risk, reflecting their comparatively high model-derived risk scores and strong detection performance.

Table 3. Threat Detection and Risk Assessment by Threat Category

Threat Category	Test Instances	Detected	Detection Rate (%)	Risk Score	Risk Level
DoS	9,132	9,132	100.00	99.93	High
PortScan	5,719	5,719	100.00	99.88	High
DDoS	4,499	4,497	99.96	99.93	High
Brute Force	467	467	100.00	99.87	High
Botnet	77	72	93.51	86.67	High
Web Attacks	70	70	100.00	98.89	High
Infiltration	8	5	62.50	66.75	Medium
Heartbleed	5	4	80.00	61.12	Medium

Source: Authors' analysis based on the CIC-IDS2017 dataset.

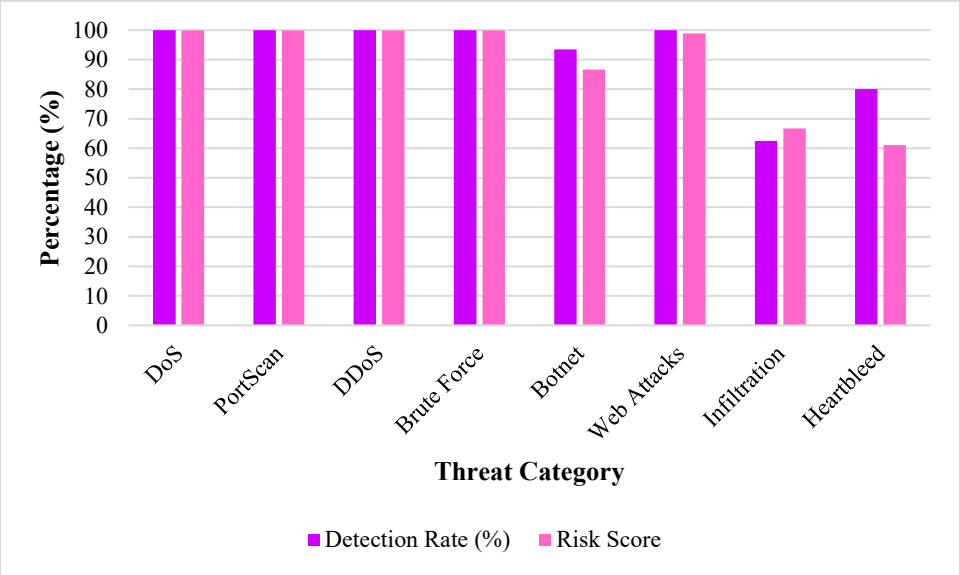

Figure 3. Threat Detection and Risk Assessment by Threat Category

The relationship between threat detection performance and the corresponding model-derived risk scores across the evaluated attack categories. DoS, PortScan, DDoS, Brute Force, and Web Attacks showed detection rates close to or at 100%, accompanied by high risk scores as shown in Figure 3. Botnet activity also demonstrated strong detection performance and a comparatively high-risk score. In contrast, Infiltration recorded the lowest detection rate and a medium risk score. Heartbleed showed a higher detection rate than Infiltration but remained within the medium-risk category. Overall, the results indicate that detection effectiveness and model-derived risk estimates varied across threat categories.

4. Discussion

This study developed an integrated approach combining machine-learning-based cyber-threat detection, SHAP-based explanation, and model-derived risk assessment. Among the evaluated classifiers, XGBoost demonstrated the strongest overall predictive performance, while Random Forest and MLP also achieved high classification effectiveness. The SHAP analysis further identified the network-flow characteristics that

contributed most strongly to the XGBoost predictions, providing an interpretable perspective on the model's decisions. The risk assessment additionally showed that most evaluated threat categories received high risk scores, whereas Infiltration and Heartbleed were assigned medium risk levels. Together, these findings indicate that combining detection accuracy with explainability and risk interpretation can provide a more informative cybersecurity analysis than classification performance alone (Rjoub et al., 2023).

The comparative evaluation indicates that ensemble-based approaches were particularly effective for distinguishing malicious from benign network traffic. XGBoost achieved the highest accuracy, recall, and F1-score among the evaluated models, while maintaining a very low false-positive rate. Random Forest produced similarly strong results, suggesting that nonlinear ensemble methods can effectively capture relationships within network-flow characteristics. Logistic Regression performed less strongly, which may reflect the limitations of a linear decision boundary for complex traffic patterns. These observations are consistent with broader research showing the effectiveness of machine-learning and deep-learning approaches for intrusion detection in increasingly complex network environments. Nevertheless, strong benchmark performance should not automatically be interpreted as equivalent real-world deployment performance.

The SHAP analysis provides an additional layer of interpretation beyond conventional performance metrics. Destination Port and Bwd Packet Length Std were among the most influential characteristics in the XGBoost model, indicating that traffic-flow and transport-related properties contributed substantially to classification decisions. The presence of both positive and negative SHAP contributions in the underlying analysis demonstrates that individual network characteristics can influence predictions in different directions. Such information is valuable because security analysts can examine which characteristics are associated with model decisions rather than relying exclusively on the final classification label. Previous research has similarly identified explainability as an important requirement for trustworthy cybersecurity AI, particularly where complex models are used for intrusion detection (Zhang et al., 2022).

The threat-specific analysis showed high model-derived risk scores for DoS, PortScan, DDoS, Brute Force, Botnet, and Web Attacks. In contrast, Infiltration and Heartbleed received medium risk classifications. The lower detection rate observed for Infiltration is particularly relevant because only a small number of test instances were available for this category. Consequently, its risk classification should be interpreted cautiously rather than considered definitive evidence of lower real-world severity. Similarly, the risk score represents the study's model-derived measure based on predicted malicious probabilities rather than an externally validated cybersecurity risk standard. Integrating risk interpretation with explainability nevertheless provides a useful analytical bridge between automated classification and security-oriented decision-making.

The findings broadly agree with previous investigations demonstrating that AI-based IDSs can achieve strong predictive capability while XAI can improve the interpretability of complex security models. Sharma et al. (2024) examined explainable deep-learning approaches for intrusion detection in IoT networks, highlighting the value of combining detection capability with interpretability. Similarly, Shoukat et al. (2025) integrated explainable AI with deep learning to develop a more transparent threat-detection system for industrial networks. Wali et al. (2025) also demonstrated the relevance of combining explainability with machine-learning-based intrusion detection. The present study extends this general direction by combining comparative model evaluation, SHAP-based feature interpretation, and threat-specific risk assessment within a single analytical workflow.

The proposed framework has practical relevance for intelligent networks in which security systems must process substantial quantities of network-flow information. High-performing classifiers can assist with rapid identification of suspicious activity, while SHAP explanations can provide additional information about the characteristics associated with individual model decisions. This combination may support prioritisation of alerts and facilitate subsequent investigation. For example, identifying influential traffic characteristics can help analysts focus attention on particular flow properties instead of treating every detection as an unexplained warning. The approach is therefore consistent with the broader movement toward transparent and trustworthy AI-enabled cybersecurity systems.

For security analysts and network administrators, explainable predictions can make automated detection outputs easier to examine and validate. Rather than receiving only a benign or malicious classification, analysts can obtain information about the network features contributing to the prediction. Risk scores can further provide an additional basis for prioritising detected activities. However, these outputs should support rather than replace expert assessment. Analysts should consider network context, asset importance, attack consequences, and operational conditions before taking defensive action. This human-centred interpretation

is important because explainability is most valuable when it helps security personnel understand and appropriately respond to AI-generated alerts.

A major strength of the study is its integration of three complementary analytical components: predictive detection, model explanation, and risk assessment. Evaluating four different machine-learning approaches provides a comparative basis for selecting the strongest classifier rather than relying on a single algorithm. The subsequent SHAP analysis adds transparency to the selected model, while threat-category analysis provides more detailed information than an overall binary classification result. Another strength is the use of multiple performance measures, including discrimination and error-related metrics. This multidimensional evaluation is consistent with recommendations for more reliable and interpretable cybersecurity AI systems.

Several limitations should be considered when interpreting the findings. First, the analysis relies on the CIC-IDS2017 benchmark dataset, which may not fully represent current operational network environments or emerging attack behaviours. Second, the classification task uses a binary benign-versus-malicious formulation, limiting detailed differentiation among individual attack subtypes during model training. Third, some threat categories contained relatively few test observations, particularly Infiltration and Heartbleed, which reduces confidence in category-specific estimates. Fourth, the risk score is a study-specific model-derived measure rather than a validated operational risk framework. Finally, the experimental results do not establish real-time deployment performance. These limitations should be addressed before applying the framework directly to production security environments.

Future research should evaluate the framework using contemporary and heterogeneous network datasets to determine whether the observed performance generalises beyond CIC-IDS2017. Multi-class attack identification could provide more detailed threat characterisation, while cross-dataset testing could examine robustness under changing traffic conditions. Future studies could also investigate real-time implementation and assess computational latency in operational environments. The risk component could be strengthened by incorporating contextual factors such as asset criticality, attack impact, exposure, and temporal threat information rather than relying solely on model confidence. Further investigation of XAI techniques and human-centred evaluation could also determine whether explanations genuinely improve analysts' decisions and reduce inappropriate responses to automated security alerts.

5. Conclusion

This study presented an explainable AI framework for cyber-threat detection and risk assessment in intelligent networks. Four machine-learning approaches were evaluated using labeled network-traffic data, followed by SHAP-based analysis of the strongest-performing model. A model-derived risk assessment was then used to examine threat categories according to their predicted security risk. XGBoost achieved the strongest overall detection performance, demonstrating effective discrimination between benign and malicious traffic. The SHAP analysis identified network-flow characteristics that contributed substantially to the model's predictions. Most evaluated threat categories received high risk scores, whereas Infiltration and Heartbleed were classified at a medium risk level. The main contribution is the integration of machine-learning detection, feature-level explanation, and threat-oriented risk assessment within one analytical framework. This combination provides information not only about whether traffic is potentially malicious but also about influential network characteristics and the associated model-derived risk level. The framework can support security analysts by making automated detection outputs more interpretable and facilitating the prioritisation of potentially important threats. Such information may assist subsequent investigation and security decision-making in intelligent network environments. Future work should examine the framework across additional datasets, multi-class attack scenarios, real-time environments, and contextual risk factors. Such extensions could improve generalisability, operational usefulness, and the reliability of explainable AI for cybersecurity applications.

References

1. Alketbi, K. S., & Mehmood, A. (2025). A comprehensive survey of explainable artificial intelligence techniques for malicious insider threat detection. IEEE Access, 13, 121772-121798.
2. Albashayreh, A., Al-Sharaeh, S., Tashtoush, Y., & Zahariev, P. (2025). Enhancing 5G Network Security: A Deep Learning Framework for Real-Time DDoS Detection and Explainable Threat Analysis. IEEE Access.

3. Almheiri, S. J., Shah, A. A., Abbas, S., Ahmad, M., & Khan, M. A. (2025). Smart sustainable cyber security: modelling an interpretable and transparent threat detection with explainable artificial intelligence. Discover Sustainability, 6(1), 442.
4. Alshudukhi, K. S., Ali, S., Humayun, M., & Alruwaili, O. (2025). Next-generation lightweight explainable AI for cybersecurity: A review on transparency and real-time threat mitigation. Computer Modeling in Engineering & Sciences, 145(3), 3029-3085.
5. Arreche, O., Guntur, T., & Abdallah, M. (2024). Xai-ids: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems. Applied Sciences, 14(10), 4170.
6. Ashfaq, S., & Chowdhury, T. K. (2023). Explainable artificial intelligence (XAI) approaches for cyber risk assessment in financial services. American Journal of Interdisciplinary Studies, 4(03), 96-135.
7. Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable artificial intelligence in cybersecurity: A survey. IEEE Access, 10, 93575-93600.
8. Dhanushkodi, K., & Thejas, S. (2024). Ai enabled threat detection: Leveraging artificial intelligence for advanced security and cyber threat mitigation. IEEE access, 12, 173127-173136.
9. Gupta, L., & Misra, D. C. (2025). Cybersecurity Threat Detection Through Explainable Artificial Intelligence (XAI): A Data-Driven Framework. International Research Journal of MMC, 6(2), 119-131.
10. Haque, B. M. T., Rahman, M. A., Chowdhury Rubel, M. S. K., & Hossan, M. I. (2024). Explainable AI-driven cyber risk analytics and model reliability assessment for intelligent governance of U.S. critical infrastructure: An XGBoost and SHAP-based intrusion detection framework. Applied IT & Engineering, 2(1), 1–20. DOI: 10.25163/engineering.2110762.
11. Islam, S., Basheer, N., Papastergiou, S., Ciampi, M., & Silvestri, S. (2025). Intelligent dynamic cybersecurity risk management framework with explainability and interpretability of AI models for enhancing security and resilience of digital infrastructure. Journal of Reliable Intelligent Environments, 11(3), 12.
12. Khan, I. A., Moustafa, N., Pi, D., Sallam, K. M., Zomaya, A. Y., & Li, B. (2021). A new explainable deep learning framework for cyber threat discovery in industrial IoT networks. IEEE Internet of Things Journal, 9(13), 11604-11613.
13. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
14. Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Frontiers in Artificial Intelligence, 8, 1526221.
15. Mohitkar, C., & Lakshmi, D. (2025). Explainable AI for transparent Cyber-Risk assessment and Decision-Making. In Machine Intelligence Applications in Cyber-Risk Management (pp. 219-246). IGI Global Scientific Publishing.
16. Moustafa, N., Koroniotis, N., Keshk, M., Zomaya, A. Y., & Tari, Z. (2023). Explainable intrusion detection for cyber defences in the internet of things: Opportunities and solutions. IEEE Communications Surveys & Tutorials, 25(3), 1775-1807.
17. Nalinipriya, G., Rama Sree, S., Radhika, K., Laxmi Lydia, E., Karim, F. K., Ishak, M. K., & Mostafa, S. M. (2025). Leveraging explainable artificial intelligence for early detection and mitigation of cyber threat in large-scale network environments. Scientific Reports, 15(1), 24662.
18. Nukala, H., Reddy, G. D. H., & Asha, R. (2026). A Real-Time Threat Intelligence System for Comprehensive Cyber Attack Detection and Mitigation Using Explainable AI. In ITM Web of Conferences (Vol. 85, p. 02004). EDP Sciences.
19. Nwakanma, C. I., Ahakonye, L. A. C., Njoku, J. N., Odirichukwu, J. C., Okolie, S. A., Uzondu, C., ... & Kim, D. S. (2023). Explainable artificial intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: A review. Applied Sciences, 13(3), 1252.
20. Oseni, A., Moustafa, N., Creech, G., Sohrabi, N., Strelzoff, A., Tari, Z., & Linkov, I. (2022). An explainable deep learning framework for resilient intrusion detection in IoT-enabled transportation networks. IEEE Transactions on Intelligent Transportation Systems, 24(1), 1000-1014.
21. Prasad, P. W. C., Sayeed, M. S., Nguyen, D. M., Hutabarat, D. P., & Mohiuddin, G. M. (2026). Explainable AI: enhancing decision-making in the detection of cyber threats. Frontiers in Computer Science, 8, 1762332.
22. Prasad, P., Tahir, M., & Isoaho, J. (2025). Enhancing explainability of artificial intelligence for threat detection in SDN-based multicast systems. Procedia Computer Science, 257, 569-574.

23. Rjoub, G., Bentahar, J., Wahab, O. A., Mizouni, R., Song, A., Cohen, R., ... & Mourad, A. (2023). A survey on explainable artificial intelligence for cybersecurity. IEEE Transactions on Network and Service Management, 20(4), 5115-5140.
24. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP).
25. Sharma, B., Sharma, L., Lal, C., & Roy, S. (2024). Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach. Expert Systems with Applications, 238, 121751.
26. Shoukat, S., Gao, T., Javeed, D., Saeed, M. S., & Adil, M. (2025). Trust my IDS: An explainable AI integrated deep learning-based transparent threat detection system for industrial networks. Computers & Security, 149, 104191.
27. Wali, S., Farrukh, Y. A., & Khan, I. (2025). Explainable AI and random forest based reliable intrusion detection system. Computers & Security, 157, 104542.
28. Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEe Access, 10, 93104-93139.