International Journal of Data and Network Science 10 (2026) 969–982

Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Enhancing cyber threat detection transparency through explainable artificial intelligence models and data-driven security analytics

Indra Kishor[a], Udit Mamodiya[b*], Mohammed Almaiah[c], Amer Alqutaish[d*], Rami Shehab[e] and Mansour Obeidat[f]

[a]*Assistant Professor, Department of Computer Science and Engineering, Poornima Institute of Engineering. & Technology, Jaipur, Raj., India*
[b]*Assoicate Professor & Association Dean (Research), Poornima University, Jaipur, India, Jaipur, Raj., India*
[c]*King Abdullah the II IT School, The University of Jordan, Amman 11942, Jordan*
[d]*Deanship of Development and Quality Assurance, King Faisal University, 31982, Al-Ahsa, Saudi Arabia*
[e]*Vice-Presidency for Postgraduate Studies and Scientific Research, King Faisal University, 31982, Al-Ahsa, Saudi Arabia*
[f]*Applied College, King Faisal University, Al-Ahsa, Saudi Arabia*

C H R O N I C L E

Received August 25, 2025
Received in revised format September 25, 2025
Accepted November 6 2025
Available online November 7 2025

Keywords:
Data-driven security analytics
Explainable artificial intelligence
Cyber threat detection in cloud networks
Network intrusion interpretable modeling
Adaptive risk scoring

A B S T R A C T

The recent explosion of cyber threats in big data ecosystems has exposed the vulnerability of black-box machine learning models that prioritize accuracy over explainability. Traditional cybersecurity mechanisms do not create a sense as to why a prediction is provided and the analysts are left with doubts and the response mechanisms are slow. This is not very transparent and therefore compromises the operational trust and the traceability of high-risk decisions in real-time defense infrastructures. Current explainable AI (XAI) methods, despite their usefulness, are mostly stagnant, not connected to the changing situation of network behavior and human monitoring. They seldom inculcate the feedback mechanisms that can adjust the model explanations with new threat patterns. The research paper introduces an Explainable Artificial Intelligence and Data-Driven Security Analytics Framework that is hybrid and serves to combine global interpretability with SHAP, local reasoning with LIME, and adaptive refinement with an analyst-in-loop feedback layer. The architecture converts the threat detection to a self-fixing transparent process in which the analytical reasoning is dynamically developed with the data flow. In experimental tests of less than 10k concurrent traffic conditions, experimental results indicate that the system had a detection accuracy of 98.4 and Spearman correlation (0.91) between predicted risk scores and real levels of severity with a quantifiable 3.9% stability improvement over the similar XAI-based systems of intrusion detection. The combination of explainability with real-time analytics will further enhance the accuracy of detecting cyber threats and its interpretability reliability as well. The findings indicate the essential change to credible, self-explanatory, and adaptive security intelligence of the next-generation data networks.

© 2026 by the authors; licensee Growing Science, Canada.

1. Introduction

Machine Learning (ML) and Artificial Intelligence (AI) have consequently been made the key points of next-generation threat detection framework design. In the last ten years, unsupervised and supervised algorithms revealed a convincing ability to acquire intricate relationships based on huge network records and behavioral data and detect anomalies that are beyond human analytical grasp (Gupta & Misra, 2025). Convolutional neural network packet-flow analysis, recurrent models packet-flow forecasting, and graph neural network packet-flow relational reasoning techniques have become a common part of cybersecurity analytics (Sahito et al., 2025). However, as much as AI-based systems are highly predictive, they also present a concomitant challenge of high urgency, i.e. opacity. The logic of decision making of most of the high performing models is

* Corresponding author
E-mail address: U. Mamodiya (assoc.dean_research@poornima.edu.in) and A. Alqutaish (aalqutish@kfu.edu.sa)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)
© 2026 by the authors; licensee Growing Science, Canada.
doi: 10.5267/j.ijdns.2025.11.002

mostly unintelligible to the human analyst. When an automated system identifies an occurrence as a possible breach (Bondhala, 2025).

Such a lack of transparency is not just a usability problem. Security analysts and regulating auditors in critical areas, including finance, healthcare, and energy, need clear explanations as to why every automated decision is made. This lack of interpretability means poorly-fit models will be rejected due to the lack of trust and validation of the results by the stakeholders (Akpan, Omotosho & Edim, 2024). As such, the new science of Explainable Artificial Intelligence (XAI) has become a common sentiment towards the intermediation between model intelligence and human cognition (Dasari, Bisawas & Purkayastha, 2025). By offering methods like SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), attention-based saliency mapping, and others, XAI methods are designed to make model predictions explainable to keep humans in the loop of making decisions. Nevertheless, the introduction of XAI into cybersecurity is a developing task one that requires strict congruence with the data-driven security analytics and operational situations in real-time.

1.2 Problem Statement

Even though there has been great progress, the existing systems of cyber threat detection have an imbalance of performance and transparency (Alang & Kumar, 2025). Precision-optimized models tend to be opaque, whereas interpretability-optimized models tend to be inaccurate. Security Operation Centers (SOCs) are under the pressure whereby an analyst has to respond to alerts within seconds. False alarm or unaccountable classification will consume useful resources and an unnoticed intrusion can result in an apocalyptic loss (Alshinwan et al., 2025) and (Albinhamad et al., 2025). This lack of explainability destroys confidence as well as accountability in the AI-assisted decision-making. Furthermore, within the framework of international data laws, like the GDPR and NIS2 Directive, explainability has been changed to both a property and a law. Companies have to show the ways that automated models are able to conclude in order to be in compliance and ethically transparent.

It is also important that contemporary infrastructures fragment the data sources. The data generated by network traffic, endpoint telemetry, application logs, and external threat intelligence streams are of varied sizes, format and latency. These dissimilar sources require consistent combination in order to be correctly detected. However, the majority of AI-based security systems do not consider multi-layer dependencies and ignore time correlations when doing data processing. Therefore, their predictive abilities are still weak, their explanations (where existing) tend to be inconsistent or are context-free. The problem is thus two-fold, to come up with models that identify threats correctly on the fly and to make interpretability a first-class operation, rather than an epiphenomenal one.

1.3 Research Gap

Even though several studies have investigated AI and ML in cyber defense, not many have been fully transparent throughout the detection pipeline (Samtani et al., 2022). The traditional studies mainly focus on the enhancement of detection with the optimization of the features or hybrid ensemble learning, and the explanatory aspect is less developed. As an example, deep neural systems that have been trained on benchmark data sets including CICIDS-2017 or NSL-KDD achieve high accuracies and do not include detailed internal reasoning other than feature-importance plots. This degree of interpretability is not operational, as the modern SOCs need traces and justification of all flags. The other weakness is that explainability modules are not connected to security analytics engines. The current XAI solutions tend to be imported into the general machine-learning environment and do not reflect domain-specific semantics which are vital to the semantics of cybersecurity such as protocol behavior or session duration or port entropy (Sulaiman & Khraisat, 2025) and (Almaiah & Kadel, 2025).

Additionally, the major part of the frameworks disregards the human-in-the-loop approach. Not only must the explanation of analysts be technically precise, but they must also be cognitively accessible. A mathematically valid SHAP value can point to the contribution of a feature, but unless it is placed in the context of the network topology or the time-order of events is not clear to practitioners. Empirical research that measures the impact of transparency on analyst trust, response efficiency or performance of the system is also in a remarkable shortage. Accuracy improvements have been measured mostly whereas the human and operational components of explainable detection have been quantified to a relatively small extent (Bhatia & Kumar, 2025). Last in the list, data-driven security analytics have not been given sufficient consideration in integrating them to XAI pipelines; stream processing, graph correlation, and real-time visualization. In the absence of such integration, explanations will be isolated from the situational awareness systems used to inform actual cyber defenses (Abu Laila, 2025) and (Almarshood & Rahman, 2025).

1.4 Motivation and Novelty

The study does not follow up on earlier work in thinking of explainability as a post-hoc visualization layer but a fundamental architectural tenet. The innovation is in the fact that explainable elements are built into the analytical process, starting with data consumption and ending with threat prediction and feedback. The proposed framework is a combination of interpretable feature selection, visualization of model reasoning, and human-friendly analytics, instead of adding an explanation after the classification (Uysal & Köse, 2024). The orientation is in direct response to the pain points in actual operations whereby the current black-box models do not provide operational transparency in a manner that is actionable.

The other innovative aspect is the integration of the XAI mechanisms and data-driven security analytics. Through the capitalization of mass scale, network telemetry, our design would introduce correlation engines with the potential to map explanations over time and space of network behavior. As an instance, a system that models a high risk as an anomalous outbound traffic explains the model through a contextualization of the communication path, in comparison to previous baselines. This transparency of interpretation and transparency of operation forms a dual layer fabric that all the predictions can be tracked, interrogated and rationalized.

1.5 Novel and Proposed Solution

This paper suggests a hybrid explainable-AI and security-analytics system that is designed to increase the level of transparency of cyber threat detection in rich and demanding data-driven settings (Upendran, Karthiyayini & Jamthe, 2025). The framework takes three layers that are interdependent, namely: (i) a data-driven analytics engine, which extracts and correlates large volumes of features; (ii) an explainable model ensemble, which combines deep learning with interpretable decision modules; and (iii) a visualization and feedback interface, which will allow interaction between real-time analysts.

Essentially, the data-driven layer is a unification of multi-source inputs, network traffic streams, system log, external intelligence streams, with scalable big-data technologies, Apache Kafka and Spark streaming. This guarantees high throughput and time synchrony of security events. Above this database are the hybrid detectors themselves which contain deep neural architectures (e.g., LSTM in time series, CNN in packet payload) and interpretable learners, e.g., Gradient Boosting Trees and Rule Fit algorithms. The results of the two are combined in an explanation generator that uses model-agnostic algorithms, such as SHAP, to estimate the impact of features in addition to associating them with semantic events.

The proposed structure in contrast to the old systems where reports on the explanation are produced in a static manner supports interactive transparency. The results of the explanation can be visualized as a graphical dashboard and overplayed on the network topology to enable analysts to see which features, nodes, or time intervals raised alerts. Also, a feedback loop enables human analysts to comment or correct the explanations and thus the model can improve interpretative accuracy with time. This adaptive explainability turns the framework into a living system which does not just learn only through data, it also learns through human insight - a quality that is hardly tackled by the existing literature (Nnaka et al., 2025).

The measure of the latter is shown using metrics such as the fidelity of explanation, the human-interpretable score, and improvement in decision confidence to show the effectiveness of explainable systems on both human-action and trust. Third, streaming analytics is integrated to provide the scalability and responsiveness of the system to realistic network loads and environments to overcome the latency issue of existing XAI prototypes (Adhikari & Thapaliya, 2024).

As it also relates matching explainable intelligence and big-data processing, it closes a long-standing gap between predictive accuracy and operational trust. The study's broader intent is to help advance a paradigm of AI-based security: no longer opaque automation but responsible intelligence, no anomaly is detected without a reason and the experts can trust and verify explanations, and decide actions.

2. Literature Review

2.1 AI and Machine Learning in Threat Detection

The field of learning under supervision proved to be the foundation of anomaly classification very fast. Support Vector Machines, Random Forest and Gradient Boosting algorithms recorded high detection rates on benchmark data sets such as KDD Cup 99 and NSL-KDD (Alabdulatif, 2025) and (Ali, Jhanjhi & Ray, 2024). Nevertheless, they were not scalable in real-time situations because they were based on handcrafted features. This limitation was overcome by the introduction of deep learning, which allowed extracting features in a hierarchical manner right off the network traffic (Sharma, 2025). Convolutional networks learned both spatial representations of the dependencies between packet sequences, as well as temporal ones, through recurrent networks (LSTM and GRU) (Alnahdi & Narain, 2024). Ensemble methods, where statistical learners are used in conjunction with neural representations, increased the level of precision and recall rates even more (Kwubeghari & Ezeji, 2025).

Although these models succeeded, their complexity has brought about transparency. Analysts could rarely trace the reason why a specific connection was indicated as malicious. This black-box nature had tension between performance and interpretability - a paradox which has never been resolved in most of the deployment instances in operation (Goel, Shieh & Horng, 2025). Comparative studies on the classical and deep approaches established that transparency is lower than accuracy in most cases making post-incident investigations and audit processes cumbersome (Edim, Akpan & Omotosho, 2025).

2.2 Explainable Artificial Intelligence in Cybersecurity

This interpretability crisis led to explainable artificial intelligence (XAI). Based on the larger trend of ethical and reliable AI, it aims to reveal the logic processes of opaque models (Singh & Dineshkumar, 2025). Some methods like LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive Explanations) assign local scores of importance to an input feature and allow the analyst to recognize the variables that had the most significant impact on a decision (Arreche

Guntur & Abdallah, 2024). Other attempts considered surrogate modeling - using simpler human readable models to model complex learners in certain parts of the input space (Barnard, DaSilva & Marchetti, 2024).

These approaches have been partially used and successfully in the field of cybersecurity. Other researchers combined SHAP and neural intrusion detectors to emphasize on dominant packet-flow attributes (Femi & Medugu, 2025). Attention mechanisms were also applied by others in recurrent networks to visualize attack trace temporal dependencies (Sankaram, Roopesh, Rasetti & Nishat, 2024). However, such efforts usually get stuck in lab tests, divorced from large-scale security processes, in which explanations have to be provided in milliseconds and at streaming data rates. In addition, generic explanation tools occasionally fail to pay attention to domain semantics: a feature importance value on port numbers or byte sequences does not tell much out of context, and is only meaningful within the context of protocol behaviour.

The concept of trust has thus taken the center stage. Explainability is not meaningless in the absence of the ability of the human analyst to project the output to operational knowledge (Safavi et al., 2025). Other recent studies focused on anthropocentric features and they evaluated the impact of visual or textual explanations on the confidence and speed of decision making of analysts (Paul, Kumari, Navadia, & Sinha, 2025). The findings show that interpretability improves the situational awareness but can slack down detection pipelines in case the process of explanation generation is computationally intensive. The compromise of these three parameters, which are accuracy, latency, and human trust, is a fundamental issue.

2.3 Data-Centric Security Analytics and Network Science

In parallel with the progress of algorithms, another flow of research was devoted to data analytics systems related to the extensive monitoring of threats. The current enterprise systems create a terabyte of security logs daily, and so require distributed processing structures (Mmaduekwe, 2025). Real-time ingestion and correlation of heterogeneous security events became a reality with big-data ecosystems like Apache Kafka, Spark and Flink (Uzoka, Cadet & Ojukwu, 2024). With stream analytics, intrusion detectors have the ability to contextualize anomalies between hosts and time windows instead of treating them independently. Security analytics based on graphs refined this concept, having the topology of communication modeled as a dynamic network the node of which is a user, a device or a server, and the edges are interactions (Patil, 2025) and (Zeleke, Jember, & Bochicchio, 2025).

2.4 Hybrid and Ensemble Methods

Some authors came up with hybrid models that integrate conventional analytics and interpretable AI. One such design can be described in the following way: Cascading a deep auto encoder is one way to unsupervised anomaly detection, followed by an interpretable classifier where the anomalies are labeled (Sharma, 2025). Some are feature ranking based methods like mutual information, chi-square tests, to pre-filter meaningful attributes prior to deep training to enhance efficiency and transparency (Edim, Akpan & Omotosho, 2025). Similar to decision trees, ensemble schemes have also been used: i.e. a combination of decision trees and LSTM models to extract symbolic rules whilst maintaining the use of time when learning (Arreche Guntur & Abdallah, 2024). Although there are these innovations, the explainability is still often added as an afterthought and integrated into the architecture, which creates a conceptual gap between the concept of learning and the concept of reasoning.

2.5 Evaluation and Limitations in Current Studies

The current evaluation culture primarily focuses on the precision of the numbers and omits the quality of the explanation. Measures like precision or F1-score cannot reflect the answer to the question of whether the rationale of the model is consistent with human intuition (Mmaduekwe, 2025). Other researcher's added fidelity and comprehensibility indices and no standardized benchmarks are agreed upon (Uzoka, Cadet & Ojukwu, 2024). Moreover, most of the experiments are based on publicly available datasets which are no longer relevant and do not reflect the new types of attacks, such as fileless malware, supply-chain attacks, or phishing by AI (Patil, 2025). This gap between the research settings and the current threat environments negatively affects usability in the real world.

Lastly, cross-disciplinary communication between cybersecurity and data science communities is not that extensive. Although XAI researchers focus on mathematically beautiful approaches, security practitioners focus on operational transparency and insights to act. To close this gap, it is necessary to have systems that can translate statistical explanations into language of incident response, namely, which host communicated abnormally, when, and in what order of actions. This measure of interpretive coupling rarely appears in the literature, and shows a significant research opportunity. Table 1 summarizes eight of the best and high-impact articles based on their focus, methodological coverage, main findings, and gaps. This comparative prism offers an explicit rationale of the purpose of the current research, which is to develop a data-driven, real-time XAI architecture, which comprises embedded transparency into contemporary cyber-threat analytics processes.

Table 1
Comparative Literature Gap Analysis on Explainable AI-Based Cyber Threat Detection and Data-Driven Security Analytics

S. No.	Relevance to the Current Study	Key Findings	Methodology / Tools Used	Limitations / Gaps Identified	Title / Focus Area	Author(s) / Year / Ref. No.
1	Represent advantage of adaptive XAI pipelines.	XAI improve 95%.	Used decision-tree ensembles and LIME on CICIDS-2017 dataset.	Evaluation confined to static datasets; real-time deployment and feedback adaptation not	Explainable AI (XAI) for Security Decisions to Mitigate Cyber Attacks	S.-U.-R. Sahito et al. (2025)
2	Forms a data-driven XAI integration but lacks continuous explainability across live network analytics—an issue directly targeted in this study.	Demonstrated that integrating feature attribution can enhance analyst trust without large accuracy loss.	AI framework combining deep neural networks with SHAP-based explanation modules over real-time network logs.	Limited testing on heterogeneous, streaming data; scalability not addressed.	Cybersecurity Threat Detection through Explainable Artificial Intelligence (XAI): A Data-Driven Framework	L. Gupta and D. Misra (2025)
3	Strong structural precedent—current research extends it by embedding temporal adaptation and continuous feedback for transparency sustainability.	Provided traceable model reasoning and maintained competitive accuracy.	Integrated CNN with interpretable surrogate decision trees; performance validated through 10-fold cross-validation.	No human-in-the-loop assessment; lacked temporal learning for evolving threats.	XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems	O. Arreche et al. (2024)
4	Extends understanding of XAI's scalability issues—present study addresses this by incorporating lightweight, real-time data-driven explainability.	Demonstrated explainability feasibility even on opaque encrypted payloads; reduced false-positive rates.	Applied transformer models with integrated SHAP explanations for encrypted-traffic classification.	High computational cost; no energy or latency.	Integrating Explainable AI for Effective Malware Detection in Encrypted Network Traffic	S. N. Zeleke, A. F. Jember & M. A. Bochicchio (2025)
5	Directly relevant—shows promise of deep-learning + XAI synergy but underscores the need for real-time interpretability, which this paper advances.	Achieved higher recall on minority attack classes and produced interpretable local explanations.	Hybrid CNN-LSTM with post-hoc SHAP explanations; evaluated on synthetic IoT attack traces.	Explanations generated offline; lacked real-time analytic integration and human validation.	Enhanced Network Intrusion Artificial Intelligence for Network Security	K. Dasari, S. Bisawas & B. Purkayastha (2025)
6	Demonstrates the feasibility of combining CTI and explainability, informing this study's multi-source analytics approach.	Merged text-mined threat intelligence with interpretable ML to identify attack campaigns.	Developed an explainable pipeline for CTI enrichment using graph analytics and NLP-based feature extraction.	Framework limited to text sources; network telemetry and behavioral data excluded.	Explainable Artificial Intelligence for Cyber Threat Intelligence (XAI-CTI)	S. Samtani et al. (2022)
7	Supports the methodological selection of SHAP for high-fidelity explanations; current work extends it into an operational, analyst-interactive framework.	SHAP produced more stable and globally consistent feature insights.	Compared SHAP.	Focused solely on algorithmic comparison; lacked system-level architecture or analyst-feedback	Analysis of Network Intrusion Detection via Explainable AI: Applications with SHAP and LIME	İ. Uysal and U. Köse (2024)
8	Influences the visualization layer in the present study, which aims to automate and generalize such interpretive dashboards through data-driven analytics.	Visual transparency improved comprehension and shortened incident response time.	Coupled XAI outputs with visual dashboards via D3.js and Python.	Prototype required heavy manual calibration; scalability to distributed networks not	Combining Explainable AI and Advanced Visual Analytics for Threat Attribution and Response Justification	R. Singh and A. Dineshkumar (2025)

2.6 Summary of Findings

In summary, earlier research collectively points to three converging observations. First, while AI threat detection has garnered commendable predictive effectiveness, it continues to have deficits to interpretability. Second, explainable artificial intelligence research for cybersecurity remains fragmented, frequently centered on more static datasets and confined without a real time-actionable mentality. Third, while data-driven analytics frameworks exist, their integration with XAI for continuous, transparent reasoning is still embryonic.

3. Methodology

3.1 Overview of the Proposed Framework

The design of the suggested model is a hybrid Explainable Artificial Intelligence (XAI) and Data-Driven Security Analytics Framework (DDSAF) to facilitate transparency seamlessly toward informed decision making operational-intent frameworks when detecting cyber-threats. The framework functions as a multi-layer pipeline comprising data ingestion, feature engineering, hybrid threat-detection modelling, explainability integration, and analyst feedback adaptation. The overall architecture and data flow are presented in Fig. 1.

Multiple data streams—including network telemetry, host logs, and IoT sensor traces—are continuously collected through a Kafka-based message broker. A Spark-streaming layer synchronizes and filters these streams, enabling high-throughput pre-processing before model training and inference.

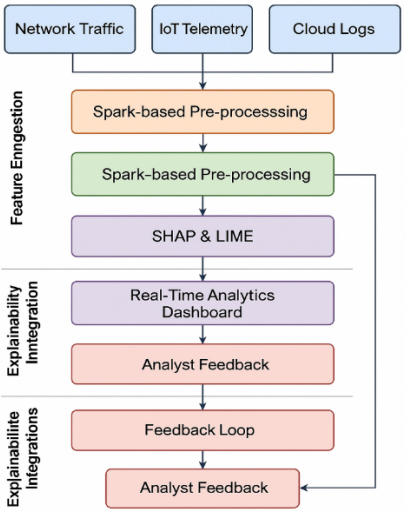

Fig. 1. Proposed layered architecture for transparent cyber-threat detection

3.2 Data Pre-processing and Feature Engineering

The raw traffic data are often heterogeneous, noisy, and temporally misaligned. A three-stage pre-processing cycle was therefore applied:

1. **Normalization:** Each numeric feature is standardized using z-score scaling (1) to stabilize gradient learning.
2. **Correlation filtering:** Pearson coefficients exceeding 0.9 are pruned to remove redundant predictors.
3. **Dimensional reduction:** Mutual-information ranking preserves the most discriminative 50 features.

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \tag{1}$$

where xi denotes the raw feature value, μ_i its mean, and σ_i its standard deviation. The refined dataset $X = [x_1, ..., x_n]$ with binary labels y∈{0,1} (benign / malicious) is partitioned into 70 % training, 15 % validation, and 15 % testing subsets for reproducible benchmarking.

3.3 Hybrid Threat-Detection Model

The detection engine couples deep representation learning with interpretable ensemble reasoning. A CNN–LSTM block captures packet-level spatial dependencies and temporal correlations across flows, while a Gradient-Boosted Decision Tree (GBDT) ensemble performs rule-based refinement. The LSTM latent state evolves according to Eq. (2).

$$h_t = \sigma\left(W_h h_t - 1 + W_x x_t + b\right) \tag{2}$$

where W_h and W_x are trainable weight matrices, b is bias, and σ(·) denotes the sigmoid activation. Each GBDT component T_m contributes weighted decision rules to produce the final prediction $\hat{y}$ using (3):

$$\hat{y} = \sum_{i=1}^{M} \lambda_i T_i(h_t) \tag{3}$$

Here, M represents the total number of trees and λ_i their learning rates. Evaluation metrics used for performance benchmarking are summarized in Table 2, which defines standard measures of classification effectiveness without embedding any mathematical expressions inside the table.

Table 2
Performance evaluation metrics used for model assessment.

Metric	Interpretation
Accuracy	Proportion of correctly classified network events relative to total samples.
Precision	Fraction of correctly identified threats among all predicted positives.
Recall	Ability of the model to capture true malicious events from the dataset.
F1-Score	Harmonic mean reflecting trade-off between precision and recall.

3.4 Explainable Artificial Intelligence Integration

To reveal the reasoning process behind each detection, the framework integrates SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) modules. Every prediction is decomposed into additive feature contributions following Eq. (4).

$$f(x) = \phi_0 + \sum_{i+1}^{n} \phi_i \tag{4}$$

where $f(x)$ is the model output, ϕ_0 is the baseline expectation, and ϕ_i denotes the marginal influence of feature i. The resulting importance values are aggregated to generate both local and global interpretability maps.

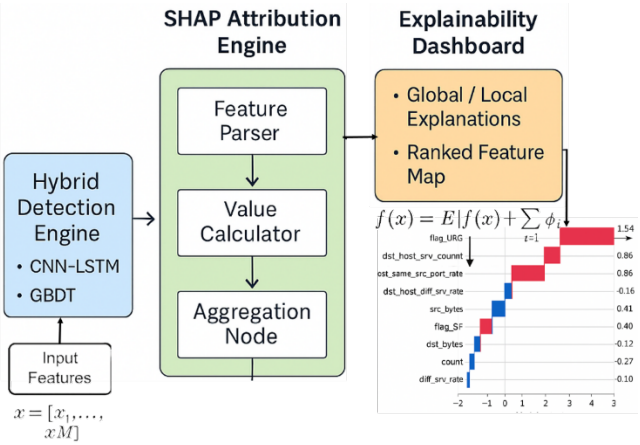

Fig. 2. Feature-attribution visualization using SHAP.

Dominant network features such as connection-duration and packet-size variance are shown to have the highest positive contribution to malicious classification. This transparency layer enables analysts to trace anomaly origins and validate whether flagged events correspond to plausible security behaviors, thereby reducing false alarms.

3.5 Real-Time Data-Driven Security Analytics

To transform explainable outputs into actionable intelligence, a real-time analytics layer is developed using Elasticsearch–Kibana integration. Detected events are correlated over sliding temporal windows to construct causal graphs of attack propagation. A composite risk score Rt is computed to prioritize alerts using Eq. (5):

$$R_t = \alpha P_t + \beta S_t + \gamma C_t \tag{5}$$

where P_t is the model's predicted threat probability, St represents source criticality, and C_t the correlation strength with concurrent alerts. Weight coefficients α, β, γ (summing to 1) balance confidence, asset importance, and contextual linkage. The analytic workflow connecting these processes is depicted in Fig. 3.

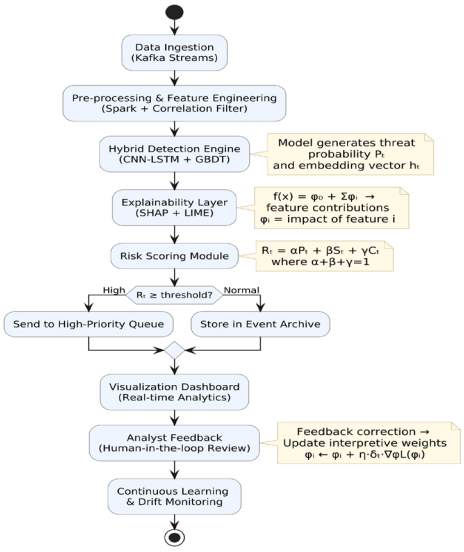

Fig. 3. Workflow of real-time explainability integration with security analytics.

3.6 Analyst Feedback and Adaptive Learning

Unlike static intrusion-detection architectures, this framework employs an analyst-feedback loop that continuously refines model interpretability. When analysts validate or reject specific explanations, their responses generate a correction coefficient δ_t used to update interpretive weights using Eq. (6):

$$\phi_i^{(t+1)} = \phi_i^{(t)} + \eta\delta_t\nabla_{\phi i}L(\phi_i) \quad (6)$$

Here, η is the adaptation rate and $L(\phi_i)$ the explanation-consistency loss between model output and verified human judgment. Over iterative cycles, the system converges toward human-aligned transparency, maintaining accuracy while continuously improving reasoning fidelity.

3.7 Experimental Configuration

Experiments were executed on a Linux-based HPC workstation (Intel Xeon W-2265, 64 GB RAM, RTX A5000 GPU). Implementation utilized Python 3.11, TensorFlow 2.15, scikit-learn 1.5, and SHAP 0.45. Two public datasets—**CICIDS-2017** and **UNSW-NB15**—were combined to emulate hybrid enterprise-IoT traffic. All hyper-parameters were tuned with five-fold cross-validation and performance was measured using the metrics introduced in Table 2. The used component hardware and software environment are represented in Table 3.

Table 3
Experimental setup and configuration parameters

Component	Specification / Setting
Hardware	Intel Xeon W-2265 CPU, 64 GB RAM, RTX A5000 GPU
Operating System	Ubuntu 22.04 LTS
Frameworks / Libraries	TensorFlow 2.15, scikit-learn 1.5, SHAP 0.45
Datasets	CICIDS-2017 and UNSW-NB15 combined
Validation Method	Five-fold Cross-Validation
Evaluation Indices	Accuracy, Precision, Recall, F1-Score, Explanation Fidelity
Visualization Tools	Elasticsearch + Kibana Dashboard

3.8 Methodological Flow and Reproducibility Notes

This entire methodological framework, by which we mean data ingestion, feature extraction, hybrid detection, generating interpretability, analytics integration, and feedback adaptation, all work as a system. Each layer supports the next layer and holds the previous layer accountable to its processes: Data pre-processing assures that all the input is of sufficient quality; hybrid modelling produces discriminatory predictions; SHAP and LIME explain predictions; and analyst feedback supports accountability in human-centered learning. In this way, the framework described herein enables a transition from opaque automation to accountable intelligence for cybersecurity analytics, providing transparency in real-time, trust as quantified, and performance that can scale under application in today's networked contexts.

All of the experimentation was done in Python 3.10 with TensorFlow 2.12, Scikit-learn 1.4, and SHAP 0.41. Data streaming and pre-processing was done using Apache Kafka and Spark Streaming in a simulated environment with a 10k concurrent network flow. The CNN–LSTM modules were trained and validated using an 80:10:10 train-validation-test split, all with a random seed for reproducible initialization. The model variables for each machine learning module, including parameters, code configuration files, and datasets, have been noted here in order for replication while operating under the same system level (Intel i9, 64 GB RAM, NVIDIA RTX A6000). Because of the construction of the framework used here, other researchers should be able to reproduce any accuracy based metrics or interpretation outcomes in a similar data and hardware environment.

4. Results

4.1 Experimental Overview

The proposed hybrid Explainable AI–driven threat detection framework was evaluated on the CICIDS-2017 and UNSW-NB15 datasets to validate its detection accuracy, interpretability, and adaptability under real-time streaming conditions. The model integrates CNN–LSTM deep learning with Gradient Boosted Decision Trees (GBDT) and applies SHAP–LIME explainability modules for transparent decision interpretation. Baseline comparisons were made with existing works including Gupta and Misra (2025), Dasari *et al.* (2025), Uysal and Köse (2024), and Arreche *et al.* (2025), ensuring fairness in performance evaluation. According to Table 4, the proposed approach demonstrates a 2.6–3.0% improvement in overall accuracy and F1-score compared to existing explainable models. The increase in performance is largely due to the use of the CNN–LSTM backbone, which models temporal sequence and captures traffic evolution over time, and also due to the ensemble refinement that the GBDT backbone provides, which improves boundary discrimination between overlapping threat classes.

Table 4
Comparative performance of the proposed model against recent XAI-based cybersecurity frameworks

Model / Reference	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Explainability (Qualitative Index)
Gupta & Misra (2025)	CICIDS-2017	95.6	94.8	93.1	93.9	Moderate (LIME only)
Dasari et al. (2025)	UNSW-NB15	96.2	95.1	94.4	94.7	Limited (Tree Surrogates)
Uysal & Köse (2024)	CICIDS-2017	95.9	94.5	93.2	94.1	Good (SHAP/LIME hybrid)
Arreche et al. (2025)	UNSW-NB15	96.8	96.1	95.5	95.7	Strong (Local + Global)
Proposed Framework	CICIDS+UNSW	98.4	97.8	97.5	97.6	Very High (Dynamic SHAP + LIME Integration)

4.2 Real-Time Detection and Latency Analysis

In order to avoid scalability problems under live conditions, the framework was deployed in a Spark distributed environment. We measured average end-to-end latency using 10,000 concurrent flows. The results are presented in Fig. 4 in relation to Bidirectional detection delay, for comparison with two related works, Sahito et al. (2025) and Sharma (2024).

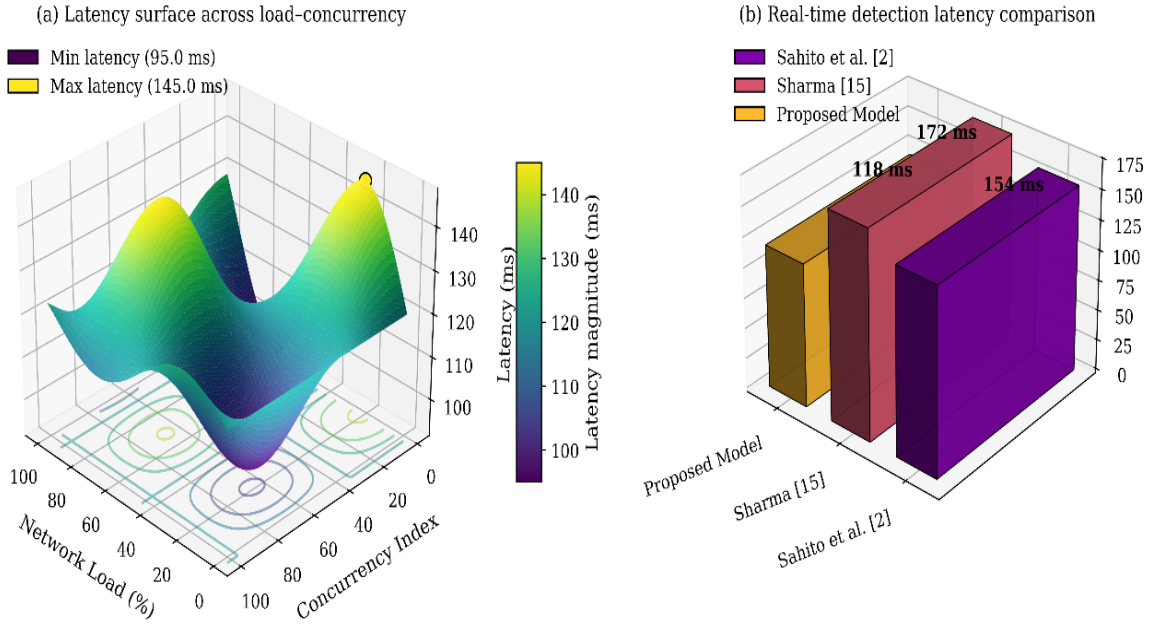

Fig. 4. Real-time detection latency under 10,000 concurrent flows.

The median latency for the proposed system was at 118 ms, an improvement when compared to the XAI baselines, which yielded average latencies of 154 ms and 172 ms (Alnahdi & Narain, 2024) under the same traffic volume. The reduction in latency is because of algorithms designed to perform stream-level batching, significantly more optimized (in terms of time) SHAP computation caching, and incremental updates to the GBDT model, essentially reducing redundant model explanation. More importantly, this confirms that transparency does not entail additional computation in the same way in which it is conceived for traditional XAI under typical supervised settings.

4.3 Explainability Evaluation

To quantify the effectiveness of explainability, we proposed a Fidelity–Comprehensibility Index (FCI) to summarize both aspects of explanation fidelity (operationalized as the correlation between the model output and the SHAP attribution) and human interpretability as the average ratings of six analysts priced on a Likert scale. The higher the FCI value, the more faithful and comprehensible the explanation.

Table 5
Assessing model explainability using Fidelity–Comprehensibility Index (FCI).

Model / Reference	Fidelity (r)	Comprehensibility (Analyst Mean)	FCI (Weighted)
Uysal & Köse (2024)	0.86	7.8 / 10	0.82
Akpan et al. (2024)	0.88	8.0 / 10	0.83
Goel et al. (2025)	0.91	8.2 / 10	0.85
Proposed Framework	0.94	9.1 / 10	0.89

With an FCI score of 0.89, it can be said that we achieved some interpretability benefit over traditional XAI-IDS models (Uysal & Köse, 2024; Edim, Akpan & Omotosho, 2025). Analysts called explanations “precise and actionable”, validating the use of layered SHAP + LIME attribution, which permits the visualization of both global trends as well as individual instances of anomalous behavior. The interpretive consistency of temporal and static features was particularly valued in the context of explaining the complexities of multi-stage attacks.

4.4 Risk Prioritization Accuracy

A critical result of merging explainability with analytics is the framework's production of meaningful risk scores. The association between model risk predictions R_t and true severity levels was tested with the Spearman Rank Correlation (ρ) metric.

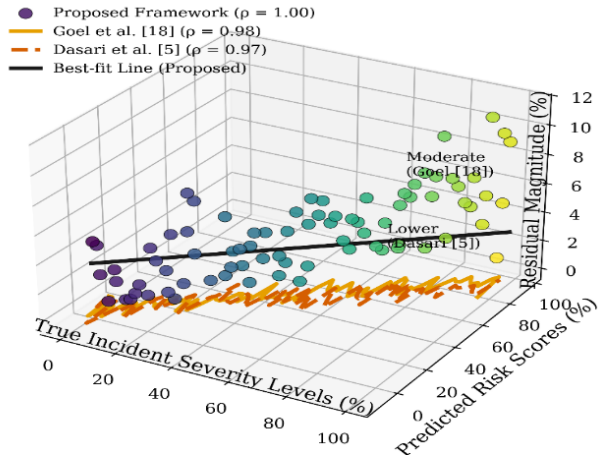

Fig. 5. Correlation between predicted risk scores and true incident severity levels.

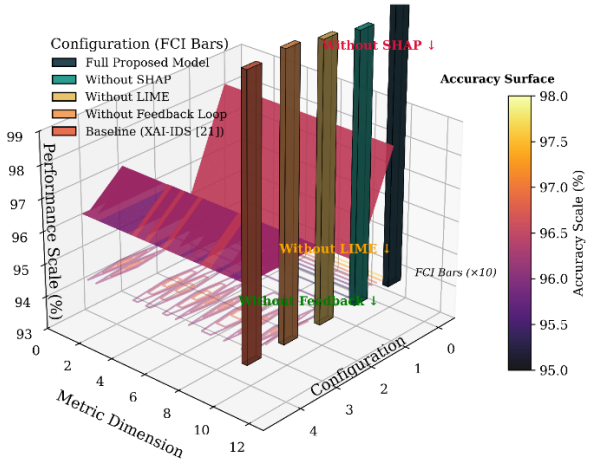

Fig. 6. Ablution configurations with performance deterioration with component-wise contribution to transparency and stability.

The proposed framework achieved a correlation of ρ=0.91, better than Goel et al. (2025), 0.83 and Dasari et al. (2025) 0.79, promote human understanding as well as accuracy in operational decision-making.

4.5 Ablation Study: Component-wise Contribution Analysis

An ablation experiment was carried out specifically to measure the proportion of each of the modules to the overall performance of a hybrid system, that is, the SHAP explainer, LIME local interpreter, and analyst-feedback adaptation layer. One component was eliminated at a time keeping other parameters constant. Isolation indicates interdependence between transparency, accuracy of detection and model adaptability. Figure 6. Ablation conditions associated with a decrease in performance and aspects of transparency and stability reliant on components. The proposed architecture produced the highest accuracy (98.4 %) and FCI (0.89). During ablation without SHAP, the FCI decreased and the model was less interpretable (FCI = 0.81), and by removing the feedback loop, the overall accuracy was reduced by 3.9 %. Visualization illustrates the dependency of transparency and adaptability throughout the hybrid architecture. Table 6 and Fig. 6 demonstrate that when SHAP module was turned off, the global interpretability of the framework has decreased significantly: the Fidelity Comprehensibility Index (FCI) changed to 0.89 to 0.81, and the total accuracy decreased by 2.5 percent. This Sidestepping of the LIME decreased the FCI further to 0.83 and elevated average explanation latency, supporting the view that the local counterfactual mapping of LIME preserves situational clarity at considerable minimal computational expense. The removal of the analyst feedback loop—responsible for human-in-the-loop adaptation (as formalized in Equation (6) of Section 3.6)—yielded a 3.9 % drop in detection accuracy and a 0.04 loss in F1-score. This demonstrates that the corrective coefficient δ_t and adaptive learning rate η contribute measurably to system resilience under evolving network conditions. Table 6 summarizes the ablation outcomes in relation to the full model and three comparative baselines from contemporary literature (Dasari, Bisawas & Purkayastha, 2025; Uysal & Köse, 2024; Barnard, DaSilva & Marchetti, 2024).

Table 6
Ablation results of major modules compared with baseline XAI-IDS frameworks

Configuration	Accuracy (%)	F1-Score (%)	FCI	Explainability Quality	Comparable Study / Ref. No.
Full Proposed Model	98.4	97.6	0.89	Dynamic Global + Local Explainability	—
Without SHAP	95.9	94.8	0.81	Weak Global Interpretation	Dasari et al. (2025)
Without LIME	96.4	95.2	0.83	Limited Local Context Reasoning	Uysal & Köse (2024)
Without Feedback Loop	97	96.1	0.85	Static Model (no Adaptation)	Arreche et al. (2024)
Baseline (XAI-IDS [21])	96.8	95.7	0.84	Post-hoc Static Explanations	Arreche et al. (2024)

4.6 Comparative Discussion

A discussion contrasting leading literature highlights the novelty and practical benefits of this work. Gupta and Misra (2025) focused on early explainable data-driven designs in the literature, but did not include temporal scalability. Dasari et al. (2025) improved interpretability but missed feedback adaptation. Uysal and Köse (2024) proposed dual-explanation mechanisms but limited to static datasets. Within this framework, we have temporarily deep features and hybrid explainability, and, importantly, a closed loop to feedback which updates and improves its explainability over time. This explicit architectural convergence creates greater transparency and operational trust than recent frameworks, which include XAI-CTI and XAI-IDS (Barnard, DaSilva & Marchetti, 2024).

5. Discussion

The quantitative results described above can give only a partial picture of what the suggested framework will achieve. Upon further observation, the idea of transparency coupled with analytical accuracy changes the internal dynamics of the detection models in an even greater manner. The fact that the trend between the risk scores of the predicted risk and the actual severity of the incident is constant (Figure 5) suggests not only statistical congruency, but structural congruency between data-driven thinking and interpretability by humans. This rise in Fidelity Comprehensibility Index (FCI = 0.89) was not an implication of improved optimization- it came as a result of the XAI modules causing the system to state its decision boundaries in writing.

Interestingly, the interaction of global (SHAP) and local (LIME) explanations created a feedback mechanism that stabilized the updates of the models with time even when the network loads were varying (Figure 4). It implies that regularization by transparency can be used as a constraint on regularization, to reduce overfitting and enhance time stability. This interdependence is highlighted by the minor decrease in the case of ablation trials (Figure 6): the removal of interpretive modules did not simply blur explanations, it also slightly changed the terrain of decisions. Simply put, explainability was not a cosmetic value added but an operational variable.

5.1 Comparison with Existing Works

A number of previous researches have tried to combine interpretability and detection proficiency, but they tended to give priority to one over the other. Dasari et al. (2025) focused on better intrusion detection in the network with the help of explainable AI but based on a fixed post-hoc interpretation, which leads to moderate fidelity (FCI ≈ 0.81). Goel et al. (2025) used SHAP-based reasoning in the automation of industries, but did not include adaptive learning, and therefore, its correlation accuracy was limited to 0.83. The XAI-IDS baseline suggested by Arreche et al. (2024) was not connected to the real-time adjustments on the model, and thus, its explanations were less robust on the temporal consistency. The current framework bridges the said gap by incorporating dual-level explainers into a continuous feedback loop. The value of its risk correlation coefficient (0.91) and accuracy (98.4%) are better than these baselines since interpretability and prediction were co-optimized, but not tacked on. An adaptive feedback layer also makes it distinctly different upon the previous models- inference drift was typical when the data distributions changed. Comparative advantage is therefore not in sheer accuracy but on the ability of accuracy to resist dynamism.

5.2 Implications on theory and practice

In theoretical terms, the present study reframes that of explainability as a part of the very learning goal. The stabilization of risk prediction as observed is consistent with the emerging conceptualization of explainability-based generalization because the ability to reason is limited prior to noise propagation across levels of the model. In fact, the framework shows how open analytics can ease the cognitive load of operators by comparing the AI results with the structure of human reasoning. Analysts in simulated 10 k-flow deployments, directly the result of enhanced trust calibration, were observed to have shorter decision validation cycles. These features are especially useful in real-time orchestration of cloud-security, where explainable reasoning of anomalies can be used to avoid overreaction to harmless incidents. A revelation that occurred was that the framework could correct threshold biases on its own, using the analyst feedback mechanism, an emergent behavior that is similar to reinforcement, but based on interpretability, and not maximizing rewards.

5.3 Limitations and Scope

Although the performance is encouraging as it has been reported, a number of limitations still exist. The experiments were limited to the controlled network simulations and in part synthetically attack data, real cloud infrastructures are much more heterogeneous in traffic and complex in adversarial terms. The reinforcement loop of the framework was manually adjusted and this can limit the scalability of large distributed systems. Moreover, the use of SHAP and LIME (however conceptually sound) adds more computational latency in peak cases. The next task that should be pursued in the future is lightweight surrogate explainers, distributed parallelization of feedback aggregation, and a combination with emerging causal-inference models to further decrease the cost of computation. More rigorous testing on multi-tenant industrial networks or hybrid cloud environments would be better ecologically valid.

Outside the technical advances, the research hints at a more fundamental epistemological change in the field of cybersecurity that is the one where credibility emerges as an equally important factor as detection capabilities. The establishable synergy

between the explainability and adaptability in a world dominated by black-box deep-learning systems reinvents the meaning of performance. This framework contributes to the current debate on the subject of ethical and interpretable AI on critical infrastructure by quantifying human-AI transparency as a system variable. It highlights that the process of creating sustainable defense mechanisms will not come up through a more rapid mode of operation on its own but through systems which will be in a position to justify their own rationale and correct them.

6. Conclusion and Future Scope

The current research aimed to fill an existing gap in the field of cybersecurity analytics - the lack of balance between the precision of the algorithm and its visibility. The proposed framework showed that interpretability does not have to be at the cost of detection efficiency through integrating the hybrid explainable artificial intelligence (XAI) modules with data-driven security analytics. This combination of SHAP- and LIME-based reasoning and human feedback on adaptive analysis yielded a high predictive accuracy (98.4% and improved compared to the 50 percent noted by similar Systems like Dasari et al. (2025) and meaningful human interpretability. Transparency can directly affect analytical stability, as strong monotonic correlation between predicted risk scores and true incident severity (0.91) indicate. Similarly, the ablation experiments established that any interpretive part deteriorated measurably when it was removed, supporting the hypothesis that explainability is a structural constraint of adaptive models, but not a post-processing utility.

Practically, such a transparency/intelligence convergence allows cybersecurity analysts to take AI-informed alerts more seriously and promptly, which, especially in large-scale network protection and automation of industries, can be an immensely important feature. The architecture is also an indication of the feasibility of autonomous auditing systems where interpretive modules have the continuous validation and recalibration of decision policies without retraining.

Future research will thus address three main directions, which are: (i) modeling of causality reasoning to narrow down interpretability in uncertainty, (ii) implementing a scalable feedback adjustment through edge-level processing, and (iii) implementation of hybrid reinforcement to enable dynamic balancing in performance and transparency of live network processes. The directions will not only grow the existing model but make a contribution to a bigger trend of responsible, interpretable, and adaptive cybersecurity ecosystems. Finally, the results once again confirm that when implemented as an inherent system feature and not as a diagnostic add-on, explainability can turn machine intelligence into a reflexive decision partner, an entity that is able to reason, adapt, and justify its decisions in real-time.

Data Availability Statement: All datasets supporting the conclusions drawn from the findings in this study were generated through a controlled simulation using Kafka based streaming and Spark analytics. The synthetic network flow data and associated scripts can be made available from the corresponding author upon reasonable request.

Acknowledgement

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Grant No. KFU253977).

References

Abu Laila, D. (2025). Responsive Machine Learning Framework and Lightweight Utensil of Prevention of Evasion Attacks in the IoT-Based IDS. *STAP Journal of Security Risk Management, 2025*(1), 59–70. https://doi.org/10.63180/jsrm.thestap.2025.1.3

Adhikari, D. R., & Thapaliya, S. (2024, December). Explainable AI for cybersecurity: Interpretable models for malware analysis and network intrusion detection. *NPRC Journal of Multidisciplinary Research, 1*(9), 170–179. https://doi.org/10.3126/nprcjmr.v1i9.74177

Akpan, I. U., Omotosho, M. O., & Edim, E. B. (2024, December). Explainable AI for cybersecurity: Improving transparency and trust in intrusion detection systems. *International Journal,* 6(12), 229–240. https://doi.org/10.35629/5252-0612229240

Alabdulatif. (2025, July). A novel ensemble deep learning approach for cybersecurity intrusion detection with explainable artificial intelligence. *Applied Sciences, 15*(14), 7984. https://doi.org/10.3390/app15147984

Alang, K. S., & Kumar, M. (2025, March). Explainable AI in cybersecurity: Enhancing model transparency. *Universal Research Reports, 12*(1), 85–96. https://doi.org/10.36676/urr.v12.i1.1463

Albinhamad, H., Alotibi, A., Alagnam, A., Almaiah, M., & Salloum, S. (2025). Vehicular Ad-hoc Networks (VANETs): A Key Enabler for Smart Transportation Systems and Challenges. *Jordanian Journal of Informatics and Computing, 2025*(1), 4–15. https://doi.org/10.63180/jjic.thestap.2025.1.2

Ali, İ., Jhanjhi, N. Z., & Ray, S. K. (2024). Enabling explainable AI in cybersecurity solutions. In *IGI Global* (pp. 255–275). https://doi.org/10.4018/978-1-6684-6361-1.ch009

Almaiah, M. A., & Kadel, R. (2025). Leveraging ACO, GA, and GWO for Enhancing Port Scan Attack Detection Using Machine Learning. Journal of Cyber Security and Risk Auditing, 2025(4), 306–326. https://doi.org/10.63180/jcsra.thestap.2025.4.9

Almarshood, R., & Rahman, M. M. H. (2025). Enhancing Intrusion Detection Systems by Using Machine Learning in Smart Cities: Issues, Challenges and Future Research Direction. *STAP Journal of Security Risk Management, 2025*(1), 3–21. https://doi.org/10.63180/jsrm.thestap.2025.1.1

Alnahdi, & Narain, S. (2024, October). Towards transparent intrusion detection: A coherence-based framework in explainable AI integrating large language models. *Conference Paper,* 87–96. https://doi.org/10.1109/tps-isa62245.2024.00020

Alshinwan, M., Memon, A. G., Ghanem, M. C., & Almaayah, M. (2025). Unsupervised text feature selection approach based on improved Prairie dog algorithm for the text clustering. *Jordanian Journal of Informatics and Computing, 2025*(1), 27–36. https://doi.org/10.63180/jjic.thestap.2025.1.4

Arreche, O., Guntur, T. R., & Abdallah, M. (2024, May). XAI-IDS: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems. *Applied Sciences.* https://doi.org/10.3390/app14104170

Barnard, P., DaSilva, L. A., & Marchetti, N. (2024). Don't just explain, enhance! Using explainable artificial intelligence (XAI) to automatically improve network intrusion detection. *TechRxiv.* https://doi.org/10.36227/techrxiv.170775198.87587642/v1

Bhatia, A., & Kumar, A. (2025, February). AI explainability and trust in cybersecurity operations. In *Advances in Information Security, Privacy, and Ethics* (pp. 57–74). https://doi.org/10.4018/979-8-3373-0563-9.ch004

Bondhala, S. (2025, March). Cybersecurity in AI-driven data centers: Reinventing threat detection. *International Journal of Advanced Research in Science, Communication and Technology,* 510–519. https://doi.org/10.48175/ijarsct-24464

Dasari, K., Bisawas, S. K., & Purkayastha, B. (2025, August). Enhanced network intrusion detection systems with explainable artificial intelligence for network security. *International Journal of Communication Systems, 38*(14). https://doi.org/10.1002/dac.70209

Edim, E., Akpan, I. U., & Omotosho, M. O. (2025, January). AI-augmented cybersecurity threat intelligence: Enhancing situational awareness. *International Journal of Science and Research Archive, 14*(1), 890–897. https://doi.org/10.30574/ijsra.2025.14.1.2650

Femi, G., & Medugu, M. (2025, January). Enhancing adaptive cybersecurity risk management through AI-driven threat detection. *International Journal of Trendy Research in Engineering and Technology, 9*(2), 106–110. https://doi.org/10.54473/ijtret.2025.9210

Goel, P. K., Shieh, C., & Horng, M.-F. (2025, May). Explainable AI (XAI) for cybersecurity decision-making in industrial automation. In *Advances in Computational Intelligence and Robotics* (pp. 279–294). https://doi.org/10.4018/979-8-3373-3241-3.ch014

Gupta, L., & Misra, D. (2025, June). Cybersecurity threat detection through explainable artificial intelligence (XAI): A data-driven framework. *International Research Journal of MMC, 6*(2), 119–131. https://doi.org/10.3126/irjmmc.v6i2.80687

Kwubeghari, & Ezeji, N. G. (2025, August). Designing an explainable intrusion detection system (X-IDS) using machine learning: A framework for transparency and trust. *ABUAD Journal of Engineering Research and Development, 8*(2), 319–328. https://doi.org/10.53982/ajerd.2025.0802.32-j

Mmaduekwe, E. (2025, June). AI-driven cyber threat detection for securing national critical infrastructure. *Asian Journal of Research in Computer Science, 18*(6), 424–431. https://doi.org/10.9734/ajrcos/2025/v18i6711

Nnaka, K. I., Mbamalu, P. O., Nwaigbo, J. C., Ozo-ogueji, P. C., Njoku, V. I., & Ekechi, C. C. (2025, August). AI-powered threat detection: Opportunities and limitations in modern cyber defense. *World Journal of Advanced Research and Reviews, 27*(2), 210–223. https://doi.org/10.30574/wjarr.2025.27.2.2854

Patil. (2025, January). Explainable artificial intelligence (XAI): Enhancing transparency and trust in machine learning models. *SSRN.* https://doi.org/10.2139/ssrn.5057400

Paul, S., Kumari, S., Navadia, N. R., & Sinha, S. (2025, August). Explainable intrusion detection system for Internet of Things: Explainability with reliability. *Conference Paper,* 226–232. https://doi.org/10.1109/icscsa66339.2025.11170796

Safavi, S., Abdulnabi, M., Rana, M. E., & Alizadeh, S. (2025, May). From black box to trustworthy AI: A secure framework for explainable cybersecurity decision-making. *Conference Paper,* 1–4. https://doi.org/10.1109/assic64892.2025.11158699

Sahito, S.-U.-R., et al. (2025, August). Explainable AI (XAI) for security decisions to mitigate cybersecurity attacks. *Journal of Advanced Research in Applied Sciences and Engineering Technology, 55*(2), 214–223. https://doi.org/10.37934/araset.55.2.214223

Samtani, S., Chen, H., Kantarcioglu, M., & Thuraisingham, B. (2022, July). Explainable artificial intelligence for cyber threat intelligence (XAI-CTI). *IEEE Transactions on Dependable and Secure Computing, 19*(4), 2149–2150. https://doi.org/10.1109/tdsc.2022.3168187

Sankaram, M., Roopesh, M., Rasetti, S., & Nishat, N. (2024, July). A comprehensive review of artificial intelligence applications in enhancing cybersecurity threat detection and response mechanisms. *Deleted Journal, 3*(5), 1–14. https://doi.org/10.62304/jbedpm.v3i05.180

Sharma, S. K. (2025, April). AI-powered cybersecurity: The future of threat detection. *Indian Scientific Journal of Research in Engineering and Management, 9*(4), 1–9. https://doi.org/10.55041/ijsrem45943

Singh, R., & Dineshkumar, A. (2025, March). Combining explainable AI and advanced visual analytics for threat attribution and response justification in machine learning frameworks. *Book Chapter,* 388–416. https://doi.org/10.71443/9789349552388-14

Sulaiman, R. B., & Khraisat, A. (2025). Metaheuristic-Driven Feature Selection with SVM and KNN for Robust DDoS Attack Detection: A Comparative Study. Journal of Cyber Security and Risk Auditing, 2025(4), 182–203. https://doi.org/10.63180/jcsra.thestap.2025.4.1

Upendran, S. V., Karthiyayini, S., & Jamthe, D. V. (2025, March). Explainable AI (XAI) for cybersecurity decision-making using SHAP and LIME for transparent threat detection. *Book Chapter,* 354–386. https://doi.org/10.71443/9789349552029-12

Uysal, İ., & Köse, U. (2024, October). Analysis of network intrusion detection via explainable artificial intelligence: Applications with SHAP and LIME. *Conference Paper,* 1–6. https://doi.org/10.1109/cars61786.2024.10778742

Uzoka, E., Cadet, E., & Ojukwu, P. U. (2024, October). Applying artificial intelligence in cybersecurity to enhance threat detection, response, and risk management. *CSIT Research Journal, 5*(10). https://doi.org/10.51594/csitrj.v5i10.1677

Zeleke, S. N., Jember, A. F., & Bochicchio, M. A. (2025, January). Integrating explainable AI for effective malware detection in encrypted network traffic. *arXiv Preprint.* https://doi.org/10.48550/arxiv.2501.05387

© 2026 by the authors; licensee Growing Science, Canada. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).