International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

EXPLAINABLE ARTIFICIAL INTELLIGENCE APPLICATIONS IN CYBERSECURITY: ENHANCING TRANSPARENCY IN INTRUSION DETECTION SYSTEMS

Asha Abraham Chandi |

Cyber Security Analyst

Pittsburgh, PA, USA |

asha.chandi6@gmail.com

Abstract

The increasing sophistication of cyberattacks and the rapid expansion of cloud, IoT, and distributed network environments have accelerated the adoption of Artificial Intelligence (AI) for intrusion detection and threat analysis. While AI-based Intrusion Detection Systems (IDS) offer superior accuracy and adaptability compared to traditional rule-based methods, they suffer from a critical limitation: the lack of transparency in their decision-making processes. This "black-box" nature reduces trust, complicates incident investigation, and hinders regulatory compliance. Explainable Artificial Intelligence (XAI) addresses these challenges by providing interpretable, human-understandable insights into how AI models detect anomalies, classify threats, and differentiate benign from malicious behaviour. This paper presents a comprehensive study of XAI applications in cybersecurity with a focus on enhancing the transparency of AI-driven IDS. It reviews existing literature, analyzes key XAI methods—including SHAP, LIME, surrogate models, and counterfactual reasoning—and introduces a structured XAI-enabled IDS framework integrating local and global interpretability, multi-modal data analysis, and analyst-centric visualization. Real-world applications such as threat detection, malware analysis, insider threat monitoring, fraud detection, root cause analysis, and compliance auditing are discussed to demonstrate XAI's practical impact. The paper concludes by highlighting open challenges and future research opportunities, emphasizing the need for real-time, scalable, privacy-preserving, and adversarial-resilient XAI solutions to enable trustworthy and operationally effective cyber defense systems.

Keywords. Explainable Artificial Intelligence (XAI), Intrusion Detection System (IDS), Cybersecurity, Threat Detection, Model Interpretability, SHAP, LIME, Deep Learning, Network Security, Security Operations Center (SOC), Explainability, Transparency, Malware Analysis, Insider Threats, Root Cause Analysis.

1. Introduction

The rapid digital transformation across industries has significantly expanded the scale and complexity of modern network infrastructures, thereby increasing exposure to sophisticated cyberattacks, advanced persistent threats, insider misuse, and zero-day vulnerabilities. As cyber adversaries continue to refine their evasion techniques, traditional signature-based and anomaly-based Intrusion Detection Systems (IDS) often fall short in detecting novel or polymorphic attacks, resulting in high false positives and poor adaptability [1]. To address these limitations, Artificial Intelligence (AI) and Machine Learning (ML) have emerged as powerful enablers of next-generation IDS due to their ability to learn complex patterns, model dynamic behaviors, and detect anomalies with high accuracy. However, the adoption of AI-

Received: August 08, 2025

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

driven IDS introduces a critical challenge: the opacity of black-box models. Deep learning and ensemble algorithms, despite their superior predictive performance, often fail to provide human-understandable rationale behind their decisions, making it difficult for analysts to trust alerts, investigate incidents, and comply with regulatory requirements regarding automated decision-making [2]. Explainable Artificial Intelligence (XAI) has therefore become a crucial component in the design of transparent and trustworthy security systems. XAI techniques such as SHAP, LIME, counterfactual reasoning, attention visualization, interpretable surrogate models, and concept-based explanations enable cybersecurity professionals to understand why a particular intrusion was detected, how feature contributions influence model behavior, and which patterns differentiate malicious activity from legitimate traffic. This interpretability is essential for building analyst confidence, reducing bias, enhancing visibility into AI decisions, and facilitating operational efficiency in Security Operations Centers (SOCs). Moreover, explainability plays a central role in regulatory compliance, as organizations must demonstrate accountability and justification for automated security decisions under frameworks such as GDPR, ISO-27001, NIST, and emerging AI governance policies [3].

With the increasing prevalence of encrypted traffic, cloud-native environments, IoT deployments, 5G networks, and edge computing, the need for transparent and context-aware intrusion detection mechanisms has become more critical. XAI-driven IDS allows security analysts to perform root cause analysis, identify attack vectors, reconstruct intrusion paths, and interpret threat patterns with improved clarity [4]. Furthermore, human-in-the-loop explainability enhances collaborative decision-making between algorithms and analysts, leading to faster response times and reduced false alarms. The integration of explainability also supports threat intelligence generation, enabling organizations to document attack behavior, share insights across security teams, and build more resilient defensive strategies. Despite significant progress, several challenges persist in operationalizing XAI within cybersecurity systems. These include balancing accuracy and interpretability, ensuring real-time explanations with minimal computational overhead, maintaining robustness against adversarial manipulations, and developing standardized evaluation frameworks for explainable IDS [5]. Additionally, current XAI techniques are often designed for general-purpose ML applications and may not directly address the domain-specific requirements of cybersecurity, where temporal dependencies, high-dimensional traffic logs, and continuous data streams must be handled efficiently.

This paper investigates the role of Explainable AI in enhancing the transparency, trust, and usability of AI-based Intrusion Detection Systems. It explores the conceptual foundations of XAI, reviews existing XAI-driven IDS approaches, and proposes a structured explainability-enabled IDS framework that integrates model interpretability at both local and global levels. Furthermore, the paper highlights critical applications of XAI across cybersecurity domains, including threat detection, fraud analysis, malware forensics, insider threat detection, and security auditing. By bridging the gap between high-performing AI models and human-understandable insights, this study emphasizes the importance of explainability as a core requirement for next-generation cybersecurity solutions.

2. Literature Review

Artificial Intelligence has significantly reshaped the cybersecurity landscape, enabling automated, intelligent, and adaptive intrusion detection capabilities. However, the integration of AI also introduces concerns related to model transparency, trustworthiness, and accountability. This section reviews foundational IDS concepts, the evolution of AI in intrusion

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

detection, the emergence of explainability requirements, and the current state of XAI-driven IDS research [6]. It also highlights limitations in existing methods and outlines opportunities for integrating explainability into security systems.

2.1 Overview of Intrusion Detection Systems (IDS)

Intrusion Detection Systems play a critical role in safeguarding networks by monitoring traffic, user activities, and system behaviour to identify security breaches or malicious activities. Signature-based IDS detect attacks by comparing network traffic patterns against predefined signatures, allowing accurate recognition of known threats but failing to detect emerging or polymorphic attacks [7]. In contrast, anomaly-based IDS establish behavioural profiles of normal network operations and identify deviations that may signify malicious intent. Although anomaly-based methods offer the advantage of detecting zero-day attacks, they often suffer from high false-positive rates due to the dynamic nature of network traffic. Hybrid IDS approaches combine the strengths of both signature-based and anomaly-based mechanisms, leveraging real-time behavioural learning alongside known threat patterns to deliver more accurate and resilient detection capabilities. Despite these advances, traditional IDS lack the adaptability needed to handle large-scale, high-velocity network environments and often provide insufficient context for human analysts to understand alerts [8].

2.2 Emergence of AI in IDS

AI techniques, particularly machine learning and deep learning models, have become pivotal in overcoming limitations of conventional IDS by enabling automated feature extraction, robust anomaly detection, and predictive modelling of evolving threats. ML algorithms such as Support Vector Machines, Random Forests, Naïve Bayes, and k-NN have demonstrated strong performance in identifying suspicious traffic patterns with improved generalization [9]. Deep learning architectures—including CNNs, RNNs, Autoencoders, LSTMs, and hybrid DL models—further enhance IDS by capturing temporal dependencies, high-dimensional representations, and complex spatial relationships in traffic data. These models outperform traditional systems in both speed and accuracy, especially for sophisticated attacks like DDoS, botnets, and advanced persistent threats. However, their reliance on opaque, non-linear transformations creates a “black-box effect,” where security analysts cannot interpret why certain decisions were made [10]. This lack of transparency poses significant operational and regulatory challenges, making the integration of explainability an essential requirement in modern AI-driven IDS.

2.3 Need for Explainability in AI-Based IDS

Despite the exceptional performance of AI-driven IDS, their black-box nature limits their acceptance in mission-critical security operations. Security analysts often lack trust in model outputs when the reasoning behind alerts remains opaque. Explainability becomes essential for validating predictions, debugging model behaviour, reducing false positives, and enabling meaningful collaboration between human analysts and automated systems [11]. Transparent explanations also help reveal biased or misleading patterns learned by models, ensuring that IDS does not generate unreliable or inconsistent decisions. Regulatory frameworks such as GDPR, NIST AI RMF, and ISO 27001 increasingly emphasize accountability and justification for automated decisions, making XAI a compliance necessity. In cybersecurity contexts where rapid and accurate decision-making could prevent significant financial and reputational losses, explainability ensures clarity, confidence, and human oversight over AI-driven security processes [12].

Received: August 08, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

2.4 Explainable Artificial Intelligence (XAI): Concepts and Methods

Explainable AI refers to a collection of methodologies that make the inner workings of complex AI models interpretable to human users. Post-hoc explainability methods, such as LIME and SHAP, generate instance-level explanations by identifying the features that most strongly influence a specific prediction. Gradient-based visualization techniques such as Grad-CAM help interpret deep learning models by highlighting critical areas in input data responsible for detection decisions [13]. Model-agnostic surrogate models, rule-based approximations, and counterfactual explanations further provide intuitive interpretations of model reasoning. In contrast, inherently interpretable models—such as decision trees, linear models, rule lists, and Generalized Additive Models—offer transparency by design but may sacrifice predictive power compared to deep learning. Visualization-driven explainability approaches, including heatmaps, temporal behaviour plots, and feature attribution graphs, allow analysts to understand traffic patterns contributing to alerts [14]. Although these methods enhance transparency, their adaptation to real-time cybersecurity contexts remains limited, as many XAI algorithms are computationally expensive or not optimized for streaming data.

2.5 Review of Existing XAI-Driven IDS Models

Several recent studies have integrated XAI into intrusion detection to improve model interpretability and analyst trust. Research involving SHAP-based anomaly explanation has shown improvements in identifying attack root causes and reducing analyst validation time. Hybrid models combining CNN-LSTM architectures with attention mechanisms provide both high accuracy and intermediate-level interpretability through attention visualizations. Surrogate decision trees built around deep learning predictions enable analysts to understand complex classification boundaries. Other studies utilize LIME to generate localized explanations for individual traffic samples, facilitating forensic analysis of suspicious events. Despite these advancements, existing XAI-driven IDS approaches often focus on single-method explanations, lack scalable real-time deployment, or produce explanations too technical for non-expert analysts [15]. A unified architecture that integrates local and global explanations, visualization dashboards, and human-in-the-loop mechanisms remains underdeveloped in the current literature.

Table 1. Related research on Explainable AI Techniques Applied to Intrusion Detection Systems.

Ref	Dataset(s)	AI / IDS Method	XAI Technique Used	Major Contribution	Key Limitations
[1]	NSL-KDD, CICIDS2017	Hybrid ML IDS	SHAP, LIME	Comprehensive survey on XAI for IDS and taxonomy of methods	Theoretical; no implementation
[2]	Multiple cyber datasets	DL-based IDS	SHAP, Grad-CAM	Evaluates performance–explainability trade-off in cyber systems	Limited real-time applicability

Received: August 08, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[3]	UNSW-NB15	ML IDS	LIME	Identifies challenges in integrating XAI into IDS	No DL model analysis
[4]	—	AI-driven cybersecurity systems	Surrogate Models	Reviews transparency & interpretability challenges	No dataset-based experiments
[5]	CICIDS2017	Deep Learning IDS	SHAP, LIME	Demonstrates XAI applications in threat & malware detection	No insider threat focus
[6]	UNSW-NB15	Ensemble IDS	Rule-based XAI	Describes dual nature of XAI (challenges & innovation)	No deployment guidelines
[7]	CAN logs (Automotive)	DL IDS	SHAP	XAI IDS for connected vehicles	Limited to automotive domain
[8]	CICIDS2017	Adversarial ML IDS	Explanation Stability	Studies adversarial effect on explanations	No defensive countermeasures
[9]	Multiple datasets	DL IDS	Global SHAP	Taxonomy of explainable deep IDS	No hybrid XAI implementation
[10]	CICDDoS2019	CNN-based IDS	LIME, SHAP	Explainable DL for DDoS & cyber defense	No encrypted traffic evaluation
[11]	CICIDS2017	XGBoost IDS	SHAP	High-performance interpretable IDS	High computational overhead
[12]	IoT-23	Federated IDS	Interpretable Trees	Federated, privacy-preserving XAI-IDS	Limited IoT scope
[13]	Cloud traffic logs	DL anomaly detection	SHAP, LIME	Explainable cloud anomaly detection pipeline	Not tested in large SOC

Received: August 08, 2025

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[14]	NSL-KDD	Attention-based DL IDS	Attention Visualization	High-accuracy IDS with built-in interpretability	Attention may not equal true explanation
[15]	Encrypted traffic	GNN-based IDS	SHAP Analysis	XAI applied to encrypted anomaly detection	Requires high computational resources

2.6 Research Gaps

The literature review highlights several critical gaps in current AI-based IDS research. First, although AI significantly improves detection accuracy, the opacity of these models limits their operational usability, highlighting the absence of comprehensive explainability frameworks. Second, existing XAI-driven IDS models often provide explanations that lack domain relevance, interpretability for SOC analysts, or real-time responsiveness. Third, many studies do not integrate both local and global explanations, resulting in partial interpretability. Fourth, most models have not been evaluated in real-world SOC environments where network behaviour is complex, multi-modal, and dynamic. Finally, there is limited research on XAI approaches tailored for zero-day attacks, encrypted traffic, or cloud-native architectures. These gaps motivate the need for a robust, transparent, and analyst-friendly XAI-enabled IDS framework that enhances both detection performance and interpretability.

3. XAI for Intrusion Detection Systems: Core Foundations

Explainable Artificial Intelligence (XAI) has emerged as a fundamental requirement in cybersecurity, particularly for AI-driven Intrusion Detection Systems (IDS) that operate in complex and high-stakes environments. As cyber threats evolve rapidly, automated detection mechanisms must not only demonstrate high accuracy but also offer clear interpretability to support informed decision-making by analysts. This section presents the conceptual foundations of XAI in IDS by explaining AI-driven IDS components, transparency requirements, the role of human analysts, integration challenges, and the ethical and legal implications of deploying explainable security technologies.

3.1 AI-Driven IDS Components and Workflow

AI-enhanced IDS employ machine learning and deep learning models to monitor and classify network behaviour as benign or malicious. The typical workflow begins with data acquisition from network traffic logs, system events, sensors, or cloud environments. Raw data undergo preprocessing steps such as normalization, encoding, feature extraction, or sequence modelling to prepare high-quality inputs for AI models. Machine learning (ML) classifiers like SVM, Random Forests, and Logistic Regression offer fast detection but rely heavily on manual feature engineering. Deep learning (DL) architectures, including CNNs, Autoencoders, LSTMs, GRUs, and hybrid models, automatically learn hierarchical representations from high-dimensional traffic patterns. These models identify anomalies by detecting deviations in temporal, spatial, or behavioural features. However, their complex internal structure—multiple hidden layers, millions of parameters, and nonlinear activation functions—creates a "black-box" phenomenon where the rationale behind predictions remains concealed. XAI bridges this interpretability gap by integrating explanation mechanisms alongside the existing AI modules.

Received: August 08, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.2 Interpretability, Transparency, and Trust Requirements

In cybersecurity, trust is not optional—it is essential. Security analysts must be confident that AI-based IDS are making correct, reliable, and unbiased decisions, especially in environments where misclassification can lead to operational disruption or financial loss. Interpretability refers to the ability of humans to understand the factors influencing a model's output, while transparency denotes the visibility into internal model processes and decision pathways. Together, these attributes enhance trust by helping analysts verify predictions, validate risk assessments, and justify mitigation actions. Transparent explanations also illuminate whether a model has learned genuine attack patterns or has overfitted to noise or irrelevant correlations. Trustworthy IDS improve collaboration between AI systems and human experts, enabling analysts to prioritize high-risk alerts effectively, reduce false positives, and enhance situational awareness. Without explainability, even highly accurate models may be rejected in operational settings due to doubts about their reliability.

3.3 Human-in-the-Loop Explainable Cybersecurity

Human-in-the-loop (HITL) integration plays a critical role in operationalizing XAI within cyber defense workflows. In real-world SOCs, analysts interact with alert dashboards, threat intelligence systems, and incident response platforms. XAI augments these workflows by presenting interpretable insights such as feature attributions, visual heatmaps, temporal evolution of anomalies, or rule-based explanations. By understanding *why* an alert was triggered, analysts can validate if the decision aligns with the observed network context. HITL explainability also enables iterative improvement of IDS models by allowing analysts to label ambiguous samples, correct model assumptions, and provide behavioural feedback through reinforcement learning mechanisms. This collaboration reduces alert fatigue by suppressing routine false positives and enhances threat investigation by supporting incident triage, root cause analysis, and remediation strategies. Ultimately, HITL-driven XAI transforms AI from an opaque detection tool into a cooperative decision-support partner.

3.4 Challenges in Integrating XAI With Cyber Defense Systems

Despite its benefits, integrating XAI into IDS presents multiple operational and technical challenges. First, most XAI algorithms—such as SHAP or LIME—are computationally expensive and may not scale effectively to real-time, high-throughput network environments where millisecond-level response is essential. Second, explanations may be technically correct but not easily interpretable by security analysts who require domain-specific insights rather than mathematical approximations. Third, adversarial attacks pose new risks, as malicious actors may exploit explanation patterns to reverse-engineer model behaviour or craft evasive traffic. Fourth, many XAI techniques have been developed for static tabular datasets, whereas IDS must handle sequential, streaming, or multi-modal network data. Fifth, balancing accuracy and interpretability remains challenging, as interpretable models may underperform in highly complex attack scenarios. Finally, integrating XAI into existing SOC workflows, visualization dashboards, and legacy security systems requires significant engineering and domain adaptation efforts. These challenges highlight the need for optimized, cybersecurity-specific XAI techniques.

3.5 Ethical, Legal, and Accountability Considerations

As AI systems increasingly influence cybersecurity decisions, ethical and legal considerations become central to responsible deployment. Explainability supports fairness by exposing potential biases in training data, such as skewed representations of certain user groups, services,

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

or protocols. It enhances accountability by providing documented reasoning for each detection, enabling organizations to justify automated decisions during audits or regulatory inspections. Legal frameworks such as the GDPR "Right to Explanation," NIST AI Risk Management Framework, EU AI Act, and ISO 42001 mandate transparency, accuracy, and human oversight in automated decision-making. In cybersecurity, opaque IDS decisions can lead to wrongful access denial, unauthorized user blocking, or delayed response to genuine threats, potentially causing financial or operational harm. XAI mitigates these concerns by fostering ethical AI usage, improving decision reliability, and ensuring that defensive actions conform to compliance standards. As cybersecurity becomes increasingly AI-driven, XAI will be essential in bridging the gap between technical efficiency and ethical governance.

4. Proposed XAI-Enabled IDS Framework

The increasing reliance on AI-based intrusion detection systems highlights the need for transparency, interpretability, and human-aligned decision-making in cybersecurity operations. To address these requirements, this section presents a comprehensive Explainable Artificial Intelligence–enabled Intrusion Detection System (XAI-IDS) framework that integrates machine learning, deep learning, and multi-level explainability mechanisms. The framework introduces both *local* and *global* interpretability modules, supports multi-modal network telemetry, and provides analyst-friendly visualizations for Security Operations Centers (SOCs). The design emphasizes three core principles: transparency, operational usability, and integration with existing security ecosystems. The following subsections describe the layered design, data handling pipeline, learning components, and the explainability mechanisms embedded in the framework.

4.1 System Overview

The proposed XAI-IDS follows a modular, layered architecture designed to detect, interpret, and contextualize cyber threats. The system ingests real-time network traffic, application logs, and endpoint telemetry, which are fed into preprocessing and feature-engineering layers. The processed data then passes through ML/DL detection engines capable of identifying malware patterns, anomalous behavior, insider threats, and emerging zero-day signatures. Unlike traditional IDS, which only generates binary decisions, the proposed system records internal model reasoning and activates explainability components to generate human-interpretable insight for each alert. Finally, a visualization dashboard communicates explanations to SOC analysts through color-coded attributions, ranking charts, traffic heatmaps, and textual reasoning, enabling faster incident triage and informed decision-making.

Received: August 08, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

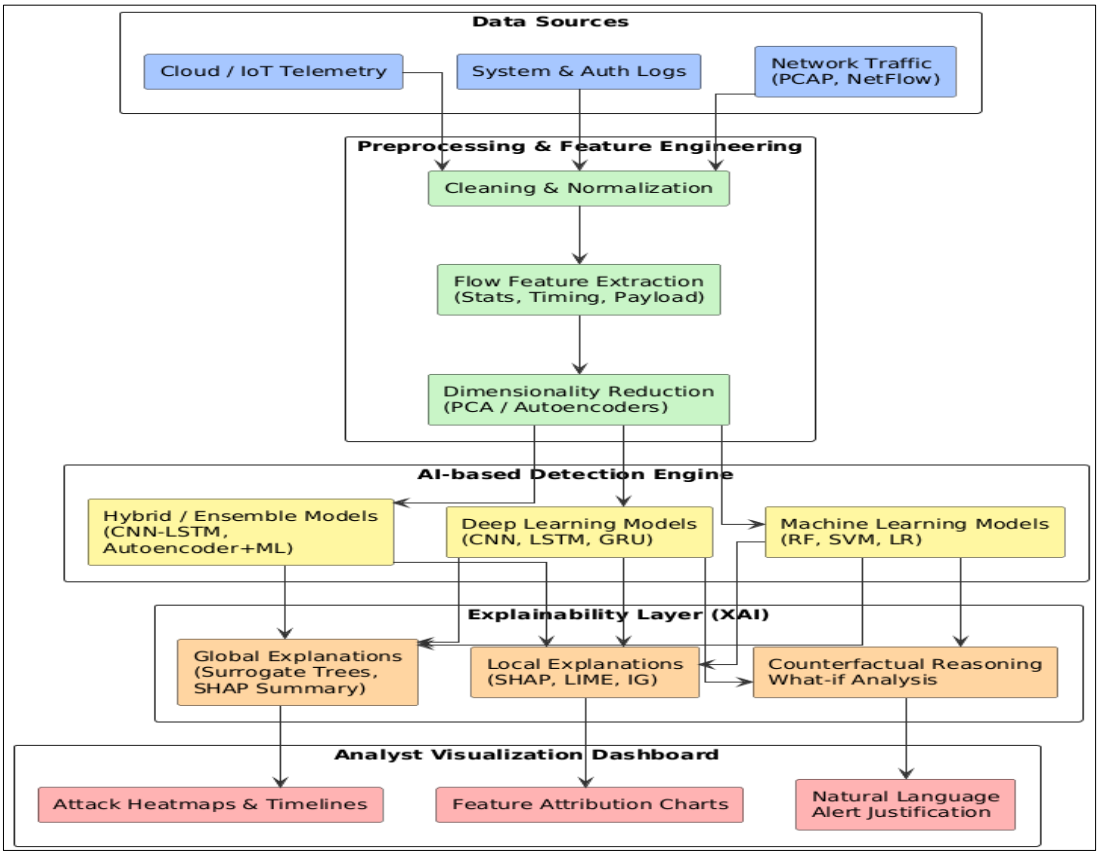

Figure 1. XAI-Enabled Intrusion Detection System

4.2 Dataset Considerations

The framework is designed to accommodate commonly used IDS datasets such as **NSL-KDD**, **CICIDS2017**, **UNSW-NB15**, and **BoT-IoT**, along with custom enterprise logs. These datasets represent diverse traffic categories including DoS, DDoS, brute force attempts, port scanning, infiltration, and heartbleed attacks. The system supports structured tabular data, raw packet captures (PCAPs), time-series flows, and payload metadata. Since dataset imbalance is common in real-world network environments, the framework incorporates balancing strategies such as SMOTE, under-sampling, and adaptive weighting. Additionally, the architecture supports incremental data ingestion from streaming telemetry pipelines such as Kafka, Zeek logs, and SIEM platforms, ensuring adaptability to real-time enterprise environments.

4.3 Data Preprocessing and Feature Engineering Layer

The preprocessing layer ensures that the input data is transformed into high-quality, noise-free features suitable for learning models. This layer includes missing value handling, normalization, outlier removal, and categorical encoding. For raw network flows, statistical features such as packet count, flow duration, average payload size, and inter-arrival times are extracted. For sequential data, the model uses sliding-window segmentation to capture temporal dependencies. Dimensionality reduction approaches such as PCA or autoencoders can be applied to reduce feature redundancy and enhance model efficiency. Additionally, the preprocessing layer retains metadata essential for explainability—such as protocol type, source/destination attributes, and temporal markers—so that explanations remain contextually meaningful for SOC analysts.

Received: August 08, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4.4 Machine Learning and Deep Learning Components

The detection engine forms the analytical core of the framework, comprising a combination of classical machine learning algorithms and advanced deep learning models. Traditional models such as Random Forest, SVM, and Logistic Regression are incorporated due to their fast interpretability and suitability for tabular data. Deep learning models—including CNNs for spatial pattern extraction, LSTMs for temporal flow modeling, and Autoencoders for anomaly reconstruction—enable advanced threat detection across heterogeneous data sources. Hybrid models such as CNN-LSTM or GRU-Attention architectures combine temporal and spatial insights for improved detection accuracy. Multi-model ensembling strengthens robustness by leveraging complementary strengths of different classifiers. Each detection output is routed into the explainability layer, ensuring that every prediction is accompanied by a clear rationale.

4.5 Explainability Layer

The central innovation of the proposed framework lies in its multi-level explainability module that integrates both post-hoc and intrinsic interpretability techniques.

4.5.1 Local Explanations (Instance-Level Interpretability)

Local explanations justify why a particular sample (e.g., a suspicious network flow) was classified as malicious or benign. Techniques such as SHAP, LIME, Integrated Gradients, and Grad-CAM identify which features contributed most to the specific prediction. For tabular data, SHAP values rank features influencing the alert, while for sequential packet data, attention maps highlight anomalous temporal behavior. These explanations support immediate analyst validation of alerts.

4.5.2 Model-Level Interpretability

Global explanations describe the overall learning patterns of the model, offering insights into what features the model considers most important during training. Surrogate models (e.g., decision trees approximating deep models), global SHAP summary plots, feature interaction maps, and rule-based abstractions help analysts understand model behavior at a high level. These explanations are essential for verifying fairness, reducing bias, and auditing IDS performance over time.

4.5.3 Analyst-Oriented Visualization Dashboard

To enhance usability, explanations are communicated through an interactive dashboard that displays:

- Feature attribution bar charts
- Global feature importance heatmaps
- Time-series anomaly maps
- Counterfactual “what-if” scenarios
- Explanation summaries generated in natural language

This visualization layer converts mathematical explanations into practical, intuitive intelligence for SOC teams.

Received: August 08, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4.6 Framework Architecture

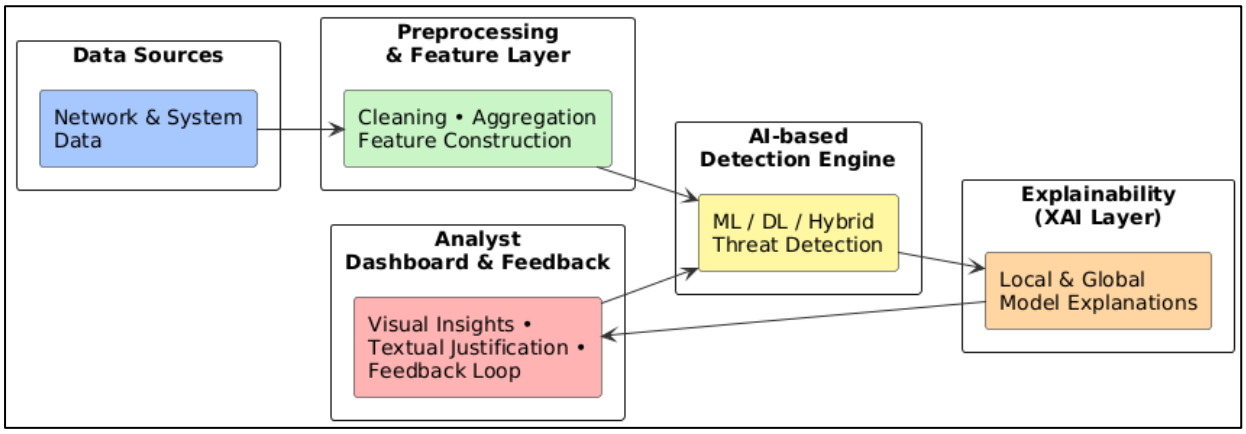

Figure 2. proposed XAI-IDS architecture

The proposed XAI-IDS architecture follows a structured data-to-explanation workflow involving ingestion, preprocessing, detection, interpretability, and visualization. Each component is designed to operate independently while maintaining seamless integration across the pipeline. Analysts can drill down into explanations at any stage—from raw feature contributions to high-level decision summaries. This modular workflow ensures flexibility across enterprise networks, cloud infrastructures, IoT ecosystems, and edge-computing environments.

5. Applications of Explainable AI in Cybersecurity

Explainable Artificial Intelligence (XAI) has emerged as an indispensable component of modern cybersecurity frameworks due to its ability to reveal the internal reasoning of AI-based intrusion detection models. Traditional IDS often fail to convey *why* a particular alert is generated, leading to analyst distrust, alert fatigue, and inefficient decision-making. XAI addresses these gaps by providing transparent, interpretable, and human-aligned explanations that enhance operational efficiency and strengthen security posture. This section discusses key real-world applications of XAI in cybersecurity, highlighting its growing relevance in threat detection, incident response, forensic analysis, compliance, and decision support systems.

5.1 Enhanced Threat Detection and Transparency

AI-driven IDS frequently detect anomalies using deep learning and ensemble models, but their decision-making processes are typically opaque. XAI enables analysts to understand the reasoning behind each detection by revealing influential features, traffic attributes, or patterns that contributed to a model's decision. Techniques such as SHAP and LIME quantify the contribution of each feature to a prediction, allowing analysts to differentiate between true malicious behaviour and benign anomalies. This transparency strengthens trust in automated detection, reduces false positives, and accelerates triage by providing contextual cues about suspicious activities such as unusual protocol usage, abnormal source-destination flows, or sudden spikes in traffic.

5.2 Explainable Root Cause Analysis (RCA)

One of the most significant challenges in incident response is uncovering the root cause of an intrusion. XAI enhances RCA by allowing investigators to reconstruct attack paths and identify specific behavioural deviations that triggered alerts. By attributing predictions to specific

Received: August 08, 2025

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

packet sequences, execution patterns, or misconfigurations, XAI helps analysts pinpoint vulnerabilities exploited by attackers. For example, in a brute-force login attack, local explanations may reveal features such as repeated authentication failures, rapid login attempts, or abnormal IP address distribution. This accelerates post-incident investigations and helps organizations implement targeted mitigation strategies.

5.3 Explainable Malware and Ransomware Analysis

Malware detection systems powered by deep learning models often classify binaries or traffic patterns without exposing underlying reasoning. XAI assists malware analysts by highlighting code segments, API calls, or behavioural indicators responsible for malicious classification. Visual explanations such as attention maps or feature contribution graphs can pinpoint suspicious file modifications, registry edits, encryption routines, or anomalous system calls associated with ransomware. This level of interpretability not only enhances reverse engineering efforts but also supports the development of more resilient signatures and behaviour-based detection models for future threats.

5.4 Insider Threat Detection Using Behavioral Explainability

Insider threats are notoriously difficult to detect because malicious activities often resemble legitimate user behaviour. AI-based behavioural analytics systems generate risk scores, but without explainability, these scores lack accountability. XAI enables the interpretation of user-specific patterns—such as unusual access times, unauthorized data transfers, privilege escalation attempts, or atypical query behaviour—by providing explanations of why a user was flagged as suspicious. This allows security teams to validate whether detected anomalies are intentional misconduct or legitimate operational variations, thereby reducing false accusations and improving internal threat monitoring.

5.5 XAI for Financial and Enterprise Fraud Detection

Fraud detection systems in financial networks increasingly use AI to analyze high-volume transactional data. However, regulatory frameworks require transparent and explainable decision-making. XAI-driven IDS models provide interpretable fraud alerts by revealing unusual transaction patterns, anomalous account behaviour, or deviations from historical financial baselines. By supplying evidence behind each flagged transaction—such as geolocation anomalies, spending pattern inconsistencies, or suspicious merchant interactions—XAI ensures compliance with auditing requirements and improves customer trust in automated fraud prevention systems.

5.6 SOC Decision Support and Alert Prioritization

Security Operations Centers (SOCs) rely on rapid, accurate analysis of thousands of daily alerts. XAI significantly enhances SOC workflows by enabling automated alert prioritization based on explanation strength and risk context. For example, alerts where SHAP values show strong malicious indicators can be marked as high-priority, while those showing weak patterns can be deprioritized. Natural language explanation modules further assist SOC analysts by summarizing model decisions into human-readable justifications. This reduces cognitive load, accelerates triage, and enables better allocation of analyst attention to critical threats.

5.7 Compliance, Auditing, and Governance

With increasing regulatory scrutiny over automated decision-making, XAI helps organizations satisfy legal requirements for transparency and accountability. Frameworks such as GDPR,

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ISO 27001, NIST 800-53, and the EU AI Act mandate explainable and auditable security decisions, especially when automated blocking or access denial is involved. XAI-enabled IDS can generate decision logs detailing why a specific action was taken, which features contributed to the decision, and how the model assessed risk. These logs serve as evidence during audits and help organizations demonstrate responsible AI practices.

5.8 Zero-Day Threat Identification Through Explainability

Zero-day attacks often involve novel, previously unknown patterns that may confuse traditional IDS. XAI helps identify these threats by revealing unusual feature interactions or outlier behaviours that differ significantly from known attack families. For instance, local explanations may highlight unexpected combinations of packet timing anomalies and payload characteristics indicative of an emerging exploit. This enables security analysts to discover evolving attack patterns early and update detection signatures accordingly.

5.9 Visualization-Enhanced Cyber Threat Intelligence (CTI)

Visualization plays a vital role in transforming raw explainability data into actionable threat intelligence. XAI supports CTI by generating network heatmaps, behaviour graphs, anomaly timelines, and protocol interaction maps that highlight key indicators of compromise (IoCs). These insights help analysts understand large-scale attack campaigns, detect coordinated botnet activities, and document behavioural patterns for intelligence sharing. Integrating XAI with CTI platforms amplifies situational awareness and enhances strategic decision-making.

6. Conclusion and Future Directions

Explainable Artificial Intelligence (XAI) has emerged as a critical enabler of trustworthy, transparent, and analyst-centric cybersecurity solutions, particularly within Intrusion Detection Systems (IDS) where decision reliability and operational interpretability are essential. While AI-driven IDS offer exceptional detection accuracy and adaptability to evolving threat landscapes, their opaque internal reasoning presents challenges related to trust, accountability, and regulatory compliance. This study addressed these shortcomings by exploring the conceptual foundations of XAI, reviewing existing approaches, and presenting a structured framework that integrates explainability across data preprocessing, model training, detection, and analyst visualization layers. The proposed XAI-enabled IDS architecture allows analysts to understand the rationale behind alerts, identify influential features, and interpret complex machine behaviours in a meaningful way. Through local and global explanations—supported by tools such as SHAP, LIME, surrogate models, counterfactual reasoning, and interactive dashboards—the framework enhances situational awareness, improves root cause analysis, and bridges the gap between automated decision-making and human cyber expertise. Despite the progress achieved through XAI integration, several challenges remain unresolved. Current explainability methods often struggle to deliver real-time insights, especially when applied to deep learning models operating on high-volume, high-velocity network traffic. Additionally, explanations may not always align with the mental models of security analysts, requiring further refinement to improve cognitive readability and domain relevance. Concerns about adversarial manipulation also highlight the need for robust, tamper-resistant explainability mechanisms that cannot be exploited to infer model weaknesses. Furthermore, the lack of standardized evaluation metrics for XAI in cybersecurity limits the consistency and comparability of existing systems, underscoring the need for unified validation frameworks. Addressing these challenges will require coordinated efforts spanning machine learning

Received: August 08, 2025

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

research, SOC practice, threat intelligence, regulatory compliance, and human–computer interaction.

Looking forward, several promising research directions can significantly advance XAI-driven cybersecurity. One important avenue is the development of **real-time explainable IDS**, capable of generating low-latency explanations without compromising detection accuracy. Another emerging direction is **federated explainable learning**, enabling distributed IDS nodes across cloud, edge, and IoT environments to collaborate securely without exchanging sensitive raw data. **Multi-agent explainable cybersecurity** systems hold potential for coordinated decision-making, where AI agents not only detect threats but also share interpretable reasoning among themselves and with human operators. Future work should also focus on **explainability for encrypted, obfuscated, and zero-day traffic patterns**, where traditional feature attribution methods may fall short. Additionally, **natural language explanation systems** could make IDS outputs more comprehensible for non-technical executives, auditors, and compliance regulators. Finally, advancing **visual analytics for XAI**—including dynamic attack path visualization, behaviour graphs, and temporal heatmaps—will further enhance analyst workflows and support more proactive cyber defense strategies.

References

[1] S. Neupane, J. Ables, W. Anderson, S. Mittal, S. Rahimi and I. Banicescu, "Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities," *arXiv preprint arXiv:2207.06236*, 2022.

[2] A. Sharma, "A comprehensive review of explainable AI in cybersecurity," *Journal of Information Security and Applications*, vol. 82, pp. 1–18, 2025.

[3] V. Z. Mohale and I. C. Obagbuwa, "A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhance transparency and interpretability in cybersecurity," *Frontiers in Artificial Intelligence*, vol. 8, art. 1526221, 2025.

[4] S. Reynaud and A. Roxin, "Review of Explainable Artificial Intelligence for cybersecurity systems," *Discover Artificial Intelligence*, vol. 5, no. 78, pp. 1–20, May 2025.

[5] Z. Zhang, H. Al Hamadi, E. Damiani, C. Y. Yeun and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," *IEEE Access*, vol. 10, pp. 93765–93791, 2022.

[6] M. Pawlicki, A. Pawlicka, R. Kozik and M. Choraś, "The dual nature of XAI challenges in intrusion detection and their potential for AI innovation," *Artificial Intelligence Review*, vol. 57, pp. 1–47, 2024.

[7] C. I. Nwakanma, "Explainable Artificial Intelligence (XAI) for intrusion detection systems in intelligent connected vehicles," *Applied Sciences*, vol. 13, no. 3, pp. 1–17, 2023.

[8] R. Bafna, B. Wressnegger and K. Rieck, "On the role of explanations in adversarial machine learning: A comprehensive survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8871–8891, Nov. 2023.

[9] R. Vinayakumar, K. Alazab and S. Srinivasan, "Explainable deep learning-based intrusion detection systems: A taxonomy and survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–45, 2024.

Received: August 08, 2025

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[10] A. B. Wahab, H. A. Yasin, M. Alazab, S. Khan and T. R. Gadekallu, "Explainable deep learning for robust cybersecurity: Trends, challenges and opportunities," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 3, pp. 820–840, June 2023.

[11] W. M. Alghamdi, M. Alenezi and M. A. Rassam, "Improving network intrusion detection using interpretable machine learning: SHAP-enhanced XGBoost-based framework," *IEEE Access*, vol. 11, pp. 45602–45617, 2023.

[12] N. Abubakar, H. Alqahtani and A. Alrawais, "Explainable machine learning for IoT intrusion detection systems: A federated and interpretable approach," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 4219–4231, Mar. 2023.

[13] D. Mahdavian and M. Malekinejad, "XAI-enhanced anomaly detection using LIME and SHAP for secure cloud environments," *IEEE Transactions on Cloud Computing*, vol. 12, no. 1, pp. 85–99, Jan. 2024.

[14] S. Alshammari, F. Alshehri and N. Alghamdi, "A hybrid attention-based deep learning architecture with explainability for next-generation intrusion detection," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 1312–1326, 2024.

[15] A. Mulyanto, P. H. Nguyen and J. S. Choi, "Explainable anomaly detection in encrypted traffic using graph neural networks and SHAP analysis," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 2, pp. 633–646, Feb. 2024.

Received: August 08, 2025