

Original Article

Explainable Artificial Intelligence for Cyber Security Threat Detection: Improving Transparency and Trust in AI-based Intrusion Detection Systems

Sunny Dhabhai

Cyber Resilience & Crisis Leadership, ESILV, France.

Received: 06-09-2026

Revised: 07-04-2026

Accepted: 07-10-2026

Published: 07-14-2026

ABSTRACT

The increasing adoption of artificial intelligence and deep learning in cybersecurity has significantly improved the ability of intrusion detection systems to identify complex and previously unseen network threats. However, many high-performing AI-based intrusion detection models operate as black boxes, making their decisions difficult for security analysts to interpret, validate, and trust. This study examines the role of Explainable Artificial Intelligence (XAI) in improving the transparency, interpretability, and trustworthiness of AI-driven intrusion detection systems. A structured comparative analysis of 30 foundational and domain-specific studies published between 2009 and 2022 was conducted, covering machine learning-based intrusion detection, deep learning approaches, cybersecurity datasets, interpretability methods, and XAI applications in intrusion detection. Particular attention was given to Local Interpretable Model-Agnostic Explanations, SHAP, decision trees, feature-importance techniques, surrogate models, and human-centered explanation approaches. The analysis indicates that XAI can support security analysts by clarifying the factors influencing threat classifications, improving model validation, exposing spurious correlations, and facilitating more informed incident-response decisions. Nevertheless, important challenges remain, including explanation instability, computational overhead, adversarial manipulation, limited standardization of explainability metrics, and differences in stakeholder interpretation needs. The study proposes a multidimensional evaluation perspective that considers predictive performance, fidelity, stability, comprehensibility, efficiency, robustness, and trust calibration. The findings emphasize that explainability should be integrated into intrusion detection design as a core cybersecurity requirement rather than treated as an optional post-processing feature.

KEYWORDS

Explainable Artificial Intelligence, Cybersecurity, Intrusion Detection Systems, Machine Learning, Deep Learning, Interpretability, Trust.

1. INTRODUCTION

1.1. Background of Cybersecurity Threat Detection

The rapid expansion of interconnected digital infrastructure has increased both the scale and complexity of cybersecurity threats. Modern networks now support cloud computing, mobile platforms, Internet of Things devices, industrial control systems, and distributed enterprise services, creating broader attack surfaces and more dynamic traffic patterns. As adversaries increasingly use automated, polymorphic, and evasive techniques, conventional security controls are often insufficient for detecting sophisticated intrusions. Intrusion detection systems therefore remain an essential defensive mechanism because they continuously monitor network or host activities and identify behavior associated with unauthorized access, misuse, or malicious exploitation.

Traditional IDS approaches are commonly divided into signature-based and anomaly-based methods. Signature-based systems compare observed activity with known attack patterns, providing strong performance against previously identified threats but limited capability against novel attacks. In contrast, anomaly-based IDSs attempt to detect deviations from normal behavior and can therefore identify previously unseen attacks, although they may generate higher false-positive rates (Garcia-Teodoro et al., 2009; Liao et al., 2013).

The growing complexity of network traffic has encouraged the adoption of machine learning and deep learning in intrusion detection. Machine learning techniques can identify nonlinear relationships and attack patterns that are difficult to encode manually, while deep learning can automatically learn hierarchical representations from high-dimensional data. Buczak and Guven (2015) showed the broad applicability of machine learning and data-mining methods to cybersecurity, while Yin et al. (2017), Shone et al. (2018), and Vinayakumar et al. (2019) demonstrated the potential of recurrent neural networks, deep architectures, and intelligent learning systems for intrusion detection. Ferrag et al. (2020) further highlighted the increasing importance of deep learning across cybersecurity datasets and threat-detection scenarios. These approaches can improve scalability, adaptability, and detection capability, particularly when handling large, complex, and changing traffic streams.

1.2. The Black-Box Problem

Despite their predictive advantages, advanced learning models create an important transparency problem. Deep neural networks may contain many hidden layers and millions of parameters, making their internal reasoning difficult for human users to interpret. High predictive accuracy therefore does not necessarily imply understandable decision-making. An IDS may correctly classify a connection as malicious while providing no explanation of why that decision was reached. Lipton (2018) emphasized that interpretability is often treated ambiguously, while Gilpin et al. (2018) distinguished predictive performance from the ability to explain model behavior. In operational cybersecurity, this distinction is critical because analysts must investigate alerts, prioritize incidents, and justify defensive actions.

A system that only reports “attack” or “normal” may offer insufficient evidence for confident intervention. Rudin (2019) argued that high-stakes applications should favor interpretable models when appropriate, while Arrieta et al. (2020) identified explainability as a central requirement for responsible and trustworthy artificial intelligence.

1.3. Explainable AI as a Solution

Explainable Artificial Intelligence offers methods for making complex model decisions more understandable. Interpretability generally refers to the degree to which a human can understand model behavior, whereas explainability focuses on producing understandable reasons for particular outputs. XAI methods may be local, explaining individual predictions, or global, describing the overall behavior of a model. They may also be model-specific or model-agnostic. Ribeiro et al. (2016) introduced LIME as a model-agnostic method for locally explaining classifier predictions, while Lundberg and Lee (2017) proposed SHAP for attributing predictions to individual features. Guidotti et al. (2018), Adadi and Berrada (2018), and Carvalho et al. (2019) provided broader taxonomies of black-box explanation and interpretability methods. Miller (2019) further emphasized that useful explanations should reflect how humans understand causes and decisions. Within intrusion detection, such explanations can help analysts determine which traffic features influenced an alert, validate whether model reasoning is plausible, and support more informed response decisions.

1.4. Research Gap

Most intrusion-detection research has traditionally emphasized predictive measures such as accuracy, precision, recall, F1-score, and detection rate. These metrics remain essential, but they do not reveal why a model reaches a specific decision or whether that reasoning is understandable to security analysts. Recent studies have increasingly addressed explainability in cybersecurity, including explainable IDS frameworks and trust-oriented approaches (Wang et al., 2020; Mahbooba et al., 2021). However, important gaps remain concerning explanation stability, analyst comprehension, computational efficiency, real-time applicability, and the extent to which explanations genuinely improve justified trust. Reviews by Zhang et al. (2022), Neupane et al. (2022), and Charmet et al. (2022) indicate that standardized evaluation of explainable cybersecurity systems remains an unresolved challenge.

1.5. Research Aim and Research Questions

This study critically examines how XAI can improve transparency, interpretability, and trust in AI-based intrusion detection while identifying associated technical and human-centered challenges. It addresses five questions: why explainability is required in AI-based IDSs; which XAI techniques are most suitable for cybersecurity threat detection; how XAI can improve transparency and analyst trust; which technical, operational, and security challenges affect implementation; and how explainable IDSs should be evaluated beyond conventional detection accuracy.

2. BACKGROUND AND RELATED LITERATURE

2.1. Evolution of Intrusion Detection Systems

Intrusion detection systems (IDSs) have evolved from rule-based mechanisms to data-driven intelligent systems designed to identify increasingly complex cyber threats. Early IDSs relied mainly on signature-based detection, matching network events against predefined attack patterns. This approach is effective for known threats and relatively transparent because the triggering rule can usually be identified. Its major weakness is limited ability to recognize zero-day attacks or substantially modified attack patterns.

Anomaly-based IDSs address this limitation by learning or defining normal behavior and identifying significant deviations. Garcia-Teodoro et al. (2009) showed that anomaly detection can identify previously unseen attacks, although false positives remain a major challenge. Liao et al. (2013) similarly emphasized the need to balance detection capability with false-alarm reduction.

The growth of network traffic and attack diversity encouraged machine-learning methods such as support vector machines, random forests, decision trees, and logistic regression. These models learn discriminative traffic patterns rather than relying solely on manually constructed rules. Buczak and Guven (2015) demonstrated their broad applicability but also showed that performance depends heavily on feature quality and training data.

Deep learning further extended IDS capability by learning complex representations. Yin et al. (2017) applied recurrent neural networks to intrusion detection, exploiting their ability to capture sequential dependencies. Shone et al. (2018) used deep representation learning to improve network intrusion detection, while Mirsky et al. (2018) introduced Kitsune, an online IDS based on an ensemble of autoencoders. Vinayakumar et al. (2019) also demonstrated the effectiveness of deep architectures across intrusion-detection tasks. However, deep models are difficult to interpret, and Aldweesh et al. (2020) identified model complexity, computational requirements, and limited transparency as important challenges. These limitations have driven the transition toward explainable IDSs that combine predictive performance with understandable reasoning.

Table 1: Comparative Characteristics of Major Intrusion-Detection Approaches

IDS approach	Typical technique	Main strength	Main limitation	Explainability level
Signature-based IDS	Rules/signatures	Effective for known attacks	Poor detection of unseen attacks	High
Traditional anomaly IDS	Statistical methods	Detects behavioral deviations	High false-positive rate	Moderate-High
Machine-learning IDS	SVM, RF, DT, LR	Strong classification capability	Depends on model complexity	Moderate
Deep-learning IDS	DNN, RNN, AE	Learns complex attack patterns	Black-box behavior	Low

Explainable AI-based IDS	ML/DL + XAI	Combines detection with explanations	Added complexity and overhead	High
-----------------------------	-------------	---	----------------------------------	------

2.2. Intrusion-Detection Datasets

Dataset quality is fundamental to both IDS performance and the credibility of model explanations. KDD Cup 99 became a major benchmark for intrusion-detection research, but Tavallae et al. (2009) identified redundancy and structural weaknesses that can distort classifier evaluation. Models trained on repetitive or outdated traffic may achieve high reported performance without generalizing effectively to modern networks.

Moustafa and Slay (2015) introduced UNSW-NB15 to provide a more contemporary benchmark containing normal traffic and multiple modern attack behaviors. Ring et al. (2019) later showed that intrusion-detection datasets differ substantially in traffic realism, attack coverage, labeling, feature representation, and class balance. These differences also affect explainability. If a model learns dataset artifacts rather than genuine threat characteristics, its explanation may still mislead analysts. Representative datasets are therefore necessary for both prediction and explanation validity.

2.3. Foundations of Explainable Artificial Intelligence

Explainable Artificial Intelligence addresses decisions produced by complex machine-learning models. Lipton (2018) argued that interpretability is often used imprecisely and may refer to transparency, simplicity, or post-hoc explanation. Gilpin et al. (2018) similarly distinguished interpretability from the broader goal of producing meaningful explanations. Guidotti et al. (2018) and Adadi and Berrada (2018) classified XAI methods according to whether they explain individual predictions or global model behavior and whether explanations are intrinsic or generated after prediction.

Transparency concerns visibility into model operation, while interpretability concerns whether humans can understand relationships between inputs and outputs. Explainability provides understandable reasons for predictions. Fidelity describes how accurately an explanation represents the behavior of the original model, while understandability concerns whether the intended user can meaningfully interpret it. Miller (2019) emphasized that useful explanations should also reflect how people naturally reason about causes and decisions.

A key distinction is between intrinsic and post-hoc interpretability. Intrinsically interpretable models, including small decision trees and sparse linear models, expose their decision logic directly. Post-hoc approaches explain more complex models after training. Rudin (2019) argued that interpretable models should be preferred in high-stakes applications when suitable alternatives exist. Carvalho et al. (2019) stressed that interpretability itself requires evaluation, while Arrieta et al. (2020) positioned explainability as a central component of responsible and trustworthy AI.

2.4. XAI in Cybersecurity

The convergence of XAI and cybersecurity has produced explainable intrusion detection systems (X-IDS). Marino et al. (2018) and Amarasinghe and Manic (2018) investigated explanations for deep IDS decisions and user trust. Wang et al. (2020) developed an explainable machine-learning framework for intrusion detection, while Mahbooba et al. (2021) demonstrated the value of interpretable decision trees for trust management. More recent reviews by Zhang et al. (2022), Neupane et al. (2022), and Charmet et al. (2022) identify continuing challenges involving black-box models, explanation quality, stakeholder needs, computational cost, and security risks. X-IDS therefore represents a shift from merely detecting attacks toward producing decisions that security analysts can understand, validate, and appropriately trust.

3. RESEARCH METHODOLOGY

3.1. Research Design

This study adopted a structured comparative literature-analysis design to investigate how Explainable Artificial Intelligence (XAI) contributes to transparency, interpretability, and trust in artificial intelligence-based intrusion detection systems (IDSs). The design was selected because the study does not rely on a newly generated experimental dataset or original model-training results. Instead, it synthesizes established evidence from prior research and compares the major technological, methodological, and human-centered dimensions reported across the literature. The approach combines structured literature analysis, comparative assessment, thematic synthesis, and descriptive bibliometric analysis.

The structured literature component was used to organize evidence relating to intrusion detection, machine learning, deep learning, explainability, cybersecurity datasets, and trust in AI-assisted security systems. Comparative analysis was then applied to identify similarities and differences among IDS approaches and XAI techniques, particularly with respect to model transparency, explanation type, usability, and operational limitations. Thematic synthesis was used to integrate recurring concepts such as interpretability, feature attribution, analyst trust, explanation fidelity, computational efficiency, and adversarial robustness. In addition, descriptive bibliometric analysis was used to examine the temporal distribution of the selected studies and determine how research activity in explainable cybersecurity evolved over the review period.

3.2. Literature Corpus

The literature corpus consisted of all 30 references selected for this study, covering publications from 2009 to 2022. The corpus includes journal articles, conference papers, and one preprint relevant to XAI, intrusion detection, cybersecurity datasets, machine learning, deep learning, and explainable cybersecurity. The studies were retained because, collectively, they provide both the theoretical foundations of explainability and the domain-specific evidence required to assess XAI-enabled intrusion detection.

For analytical purposes, the literature was organized into five broad categories. The first category covered foundational XAI and interpretability research, including studies on model transparency, post-hoc explanation, interpretable machine learning, LIME, SHAP, and explanation evaluation. The second category included IDS and anomaly-detection research addressing conventional and machine-learning-based approaches. The third category focused on deep-learning IDS studies, including recurrent neural networks, autoencoders, and other neural architectures. The fourth category comprised research on cybersecurity datasets, particularly studies discussing KDD Cup 99, UNSW-NB15, and the broader limitations of network intrusion datasets. The fifth category consisted of explainable IDS and XAI cybersecurity studies directly examining transparency, trust, feature-level explanation, and human interpretation in security applications.

3.3. Inclusion and Analytical Criteria

Each publication was examined using a common analytical framework to ensure consistency across the review. The extracted variables included publication year, research domain, IDS approach, XAI method, explanation scope, contribution to transparency, contribution to trust, reported limitations, and evaluation methodology. Explanation scope was classified as local, global, or both, depending on whether the method explained individual predictions, overall model behavior, or both levels. Transparency contribution referred to the extent to which a study enabled inspection or understanding of model decisions, while trust contribution captured claims or evidence concerning analyst confidence, user acceptance, or decision support.

The analysis also recorded whether a study emphasized predictive metrics, explanation metrics, human evaluation, computational performance, or qualitative assessment. This made it possible to distinguish studies that treated explainability as a technical feature from those that evaluated its practical value for cybersecurity users.

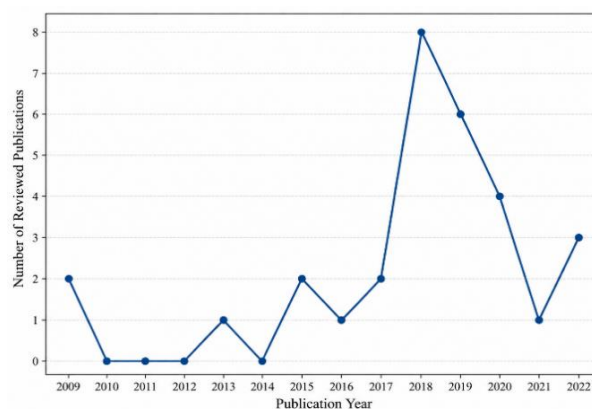


Fig 1 - Publication Trend of the Reviewed Literature, 2009–2022

The graph shows the annual distribution of the 30 studies included in the review. Research activity remained relatively limited before 2018, followed by a sharp increase between 2018 and 2020.

This pattern reflects the growing academic interest in deep learning, explainable artificial intelligence, and trustworthy intrusion detection systems during this period.

3.4. Analytical Procedure

The analysis was conducted in four stages. First, relevant information was extracted from each study and entered into a structured coding matrix. Second, recurring concepts were coded into thematic groups, including model interpretability, feature attribution, explanation quality, transparency, analyst trust, computational cost, dataset limitations, and adversarial concerns. Third, XAI methods were compared according to explanation scope, model dependency, strengths, limitations, and suitability for IDS environments. Particular attention was given to LIME, SHAP, interpretable decision trees, feature-importance techniques, surrogate models, and human-centered explanation approaches.

Finally, the coded evidence was synthesized to identify recurring patterns, methodological gaps, and unresolved research challenges. Findings were compared across the five literature categories to determine how general XAI principles translate into cybersecurity practice. Because the study is literature-based, no artificial performance values were generated. All quantitative statements, including the publication trend, were derived solely from the selected references, thereby preserving transparency and reproducibility in the analytical process.

4. XAI TECHNIQUES FOR CYBERSECURITY THREAT DETECTION

Explainable Artificial Intelligence (XAI) is increasingly important in cybersecurity because modern intrusion detection systems (IDSs) often rely on complex machine-learning and deep-learning models whose decision processes are difficult to interpret. Although such models can detect sophisticated attack patterns, their black-box nature creates operational concerns when analysts must determine whether an alert represents a genuine threat. XAI addresses this problem by providing local or global explanations that clarify why particular network events are classified as malicious.

4.1. LIME

Local Interpretable Model-Agnostic Explanations (LIME) is a widely used post-hoc XAI technique. Ribeiro et al. (2016) proposed LIME as a model-agnostic method for explaining individual predictions by approximating the behavior of a complex model within a small neighborhood around the observation being explained. Rather than reproducing the complete decision boundary of the original classifier, LIME creates a simple local surrogate model, often a sparse linear model, that behaves similarly to the black-box model near one selected instance. The procedure begins by perturbing the features of the original observation and submitting the generated samples to the trained classifier. The resulting predictions are weighted according to their proximity to the original instance. LIME then fits an interpretable surrogate model to the weighted samples, and the surrogate coefficients indicate which features contributed most strongly to the prediction.

In cybersecurity, this mechanism is useful for analyzing individual intrusion alerts. If a network flow is classified as malicious, LIME may show that unusual connection duration, abnormal packet rate, repeated connection failures, or a particular service characteristic strongly influenced the prediction. This can help an analyst distinguish a plausible attack from a false positive and support incident triage. However, LIME can be unstable because small changes in sampled neighborhoods may produce different explanations. Its output also depends on the perturbation strategy and definition of local similarity. In addition, it primarily explains individual predictions rather than complete global model behavior.

4.2. SHAP

SHAP, introduced by Lundberg and Lee (2017), is grounded in Shapley values from cooperative game theory. Each feature is treated as a participant in a coalition, while the prediction is interpreted as the result of their combined contributions. A SHAP value estimates how much each feature increases or decreases the prediction relative to a baseline expectation.

This feature-attribution mechanism is particularly useful in intrusion detection because it captures both positive and negative feature effects. A high packet count may push a prediction toward the attack class, whereas a normal traffic characteristic may shift it toward the benign class. Local SHAP explanations clarify why one specific network flow was classified as malicious, while aggregated SHAP values provide a global view of which features are most influential across the IDS.

SHAP can therefore support malicious-flow explanation, attack-feature ranking, and model auditing. It may also reveal when a detector relies excessively on irrelevant or dataset-specific variables. Its main limitation is computational complexity. Exact Shapley-value calculation becomes expensive as feature dimensionality increases, and even approximation techniques may create undesirable latency in real-time environments processing large volumes of network traffic.

4.3. Intrinsically Interpretable Models

Not all explainability methods require post-hoc interpretation. Intrinsically interpretable models are designed so that their decision logic can be examined directly. Examples include decision trees, rule-based classifiers, and sparse linear models. Rudin (2019) argues that interpretable models should be preferred to black-box systems in high-stakes applications when they can provide comparable predictive performance.

Decision trees are attractive for cybersecurity because analysts can trace a prediction through a sequence of feature-based decisions. Mahbooba et al. (2021) demonstrated how decision-tree models can enhance trust management in intrusion detection by making the reasoning behind predictions directly visible. Rule-based classifiers express decisions through explicit if-then conditions, while sparse linear models expose feature direction and magnitude through a limited set of coefficients.

Their main advantage is that the explanation is part of the predictive model rather than an approximation generated afterward, reducing concerns about explanation fidelity. However, interpretability may decline as trees or rule sets become large, and very simple structures may struggle to represent highly nonlinear attack patterns. Cybersecurity systems must therefore balance transparency with predictive capacity.

4.4. Feature Importance and Surrogate Models

Feature-importance techniques provide a broader view of the variables that influence model decisions. Global importance measures identify features that dominate predictions across many observations, whereas local surrogate models explain individual decisions. Global surrogate models instead approximate the overall behavior of a black-box detector using a simpler interpretable model.

These techniques are particularly valuable for model auditing. If an IDS performs well but feature analysis shows that it depends heavily on an irrelevant identifier, timestamp artifact, or dataset-specific variable, the model may fail outside the benchmark environment. XAI can therefore expose spurious correlations, feature leakage, and unexpected decision patterns before deployment. Nevertheless, surrogate models introduce a fidelity problem: the explanation may be understandable while still failing to reproduce the original model perfectly.

4.5. Human-Centered Explanations

The value of an explanation ultimately depends on the person interpreting it. Miller (2019) emphasizes that explanation is fundamentally human-centered and that technically correct information is not automatically useful. In cybersecurity, a data scientist may require detailed diagnostics, while a security analyst may need a concise explanation showing influential features, confidence, and similarity to known attack behavior. Managers or auditors may instead require higher-level evidence about reliability, accountability, and risk.

Effective XAI should therefore adapt explanation depth and format to stakeholder needs. Explanations should be understandable, relevant, and actionable rather than unnecessarily technical. The objective is not merely to expose model internals, but to communicate reasoning in a form that supports secure, efficient, and defensible human decision-making.

Table 2: Comparison of Major XAI Methods for Intrusion Detection

XAI method	Explanation type	Model dependency	Main advantage	Main limitation	Cybersecurity suitability
LIME	Local	Model-agnostic	Provides intuitive explanations for individual intrusion alerts	Explanations may be unstable and perturbation-sensitive	High

SHAP	Local and global	Model-agnostic and model-specific variants	Provides theoretically grounded and consistent feature attribution	Computationally demanding for complex or high-dimensional models	Very high
Decision tree	Local and global	Intrinsic	Decision paths can be inspected directly by analysts	Interpretability decreases as tree complexity increases	High
Feature importance	Global	Model-dependent or model-agnostic	Identifies variables that strongly influence overall model behavior	Limited ability to explain individual alerts	Moderate-High
Surrogate model	Local and global	Model-agnostic	Provides an interpretable approximation of black-box behavior	Surrogate fidelity may be imperfect	High
Rule-based explanation	Local and global	Usually intrinsic	Provides clear if-then reasoning that analysts can inspect	Simpler rules may sacrifice predictive flexibility	High

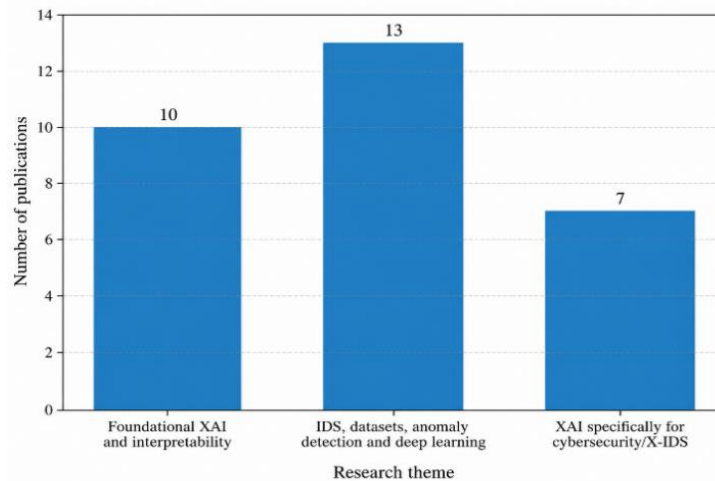


Fig 2 - Distribution of the Reviewed Literature by Research Theme

The graph shows that IDS, dataset, and deep-learning studies constitute the largest share of the reviewed literature, followed by foundational XAI research and cybersecurity-specific XAI studies. This highlights the growing integration of explainability with intrusion detection research.

5. TRANSPARENCY, TRUST, AND EVALUATION OF EXPLAINABLE IDS

5.1. Explainability and Transparency

Transparency is a critical requirement for AI-based intrusion detection because a prediction alone rarely provides enough information for operational response. A black-box IDS may report that a

connection is malicious, but the analyst still needs to know why the alert was generated, which traffic characteristics influenced the decision, and whether the model relied on security-relevant evidence. Explainable artificial intelligence addresses this limitation by exposing the factors that drive individual predictions and, in some cases, the broader decision logic of the detector. An explainable IDS can therefore show that attributes such as connection duration, failed-login frequency, source-byte volume, protocol behavior, or abnormal service usage contributed to the predicted attack class.

This transparency has two levels. Local explanations clarify why one observation was classified as malicious, whereas global explanations reveal the features and relationships that generally govern the model. Wang et al. (2020) used SHAP to provide both local and global interpretations of intrusion-detection predictions, enabling security personnel to inspect feature contributions and relationships between feature values and attack categories. This allows analysts to judge whether the model's reasoning is plausible. If a detector identifies denial-of-service traffic using security-relevant network characteristics, confidence in its reasoning is strengthened. Conversely, if predictions are dominated by an arbitrary identifier, collection artifact, or noncausal shortcut, the explanation can expose a weakness that accuracy alone may conceal.

5.2. Explainability and Human Trust

Explainability should not be used simply to make analysts trust an IDS more. The objective is calibrated trust, in which human confidence reflects the demonstrated reliability and intelligibility of the system. Amarasinghe and Manic (2018) argued that high detection accuracy is insufficient in critical security settings because black-box deep neural networks prevent users from understanding the basis of decisions. Their approach reported influential input features both offline and for individual online detections, making the IDS more communicative.

This human dimension matters because explanations are consumed by people rather than algorithms. Miller (2019) showed that useful AI explanations should reflect insights from the social sciences, including people's preference for selective, contrastive, and context-sensitive explanations rather than exhaustive technical descriptions. For a security analyst, a useful explanation should therefore answer an operational question such as "Why was this flow classified as an attack rather than normal traffic?" rather than display an uninterpretable collection of model parameters.

Mahbooba et al. (2021) similarly linked interpretability with trust management in IDS and demonstrated the value of decision-tree rules that expose the reasoning behind attack predictions. Such transparency can help analysts determine both when a model appears reliable and when its reasoning is questionable. An interpretable system that reveals uncertainty, contradictory evidence, or dependence on weak features can appropriately reduce analyst confidence and trigger further investigation.

5.3. Explainability for Model Validation

XAI is also valuable before deployment because it can serve as a diagnostic instrument for validating an IDS. High predictive scores may conceal spurious correlations, feature leakage, dataset bias, overfitting, or decision rules that will not transfer to operational networks. Feature-attribution and global explanation methods make these problems easier to identify by showing what information the model has actually learned to use.

Dataset quality is especially important. Tavallaee et al. (2009) demonstrated that KDD Cup 99 contains extensive redundant records capable of biasing learning algorithms toward frequently repeated patterns. A model evaluated on such data could produce strong metrics while learning relationships that poorly represent real network behavior. Moustafa and Slay (2015) introduced UNSW-NB15 to provide more contemporary normal and attack traffic, while Ring et al. (2019) showed that intrusion-detection datasets differ in recording environment, traffic composition, labeling, volume, and other properties affecting their suitability for evaluation.

Explanations therefore provide an additional validation layer. If an IDS achieves high test accuracy but assigns high importance to dataset-specific artifacts, leaked labels, or irrelevant variables, the model should not be considered trustworthy. Likewise, substantial changes in feature importance across retraining runs or datasets may indicate instability. XAI can thus support model debugging, feature auditing, cross-dataset comparison, and detection of overfitting before deployment.

5.4. Human-in-the-Loop Cybersecurity

For high-impact cybersecurity decisions, a human-in-the-loop architecture can be safer than unrestricted black-box automation. Neupane et al. (2022) identify human involvement as an important component of explainable IDS design because security operations analysts must interpret model predictions and translate them into response actions. In this arrangement, the AI system performs rapid pattern recognition, the XAI component provides evidence for the prediction, and the analyst evaluates that evidence alongside contextual knowledge unavailable to the model.

The operational sequence can be summarized as AI prediction + explanation + human judgment. A detector may classify a flow as malicious and highlight the responsible features, but an analyst can determine whether the activity represents a genuine attack, legitimate administrative behavior, a misconfiguration, or an authorized anomaly. Human review is particularly important for high-consequence actions such as account suspension, host isolation, service interruption, or incident escalation.

Evaluation of X-IDS should therefore extend beyond accuracy, precision, recall, and F1-score. A trustworthy system must generate explanations that faithfully represent model behavior, remain stable for similar inputs, can be understood by intended users, arrive quickly enough for operational use, and support effective decisions. It should also withstand adversarial manipulation and generalize

beyond a single benchmark dataset. These requirements are summarized in Table 3. The thematic coding in Graph 3 reinforces this multidimensional view: transparency, feature importance, and interpretability are the most consistently emphasized factors in the cybersecurity-focused XAI studies reviewed, whereas stability, fidelity, computational efficiency, adversarial robustness, and human-centered evaluation receive less uniform attention. Future X-IDS research should therefore move beyond showing that explanations can be produced and demonstrate that they are reliable, usable, secure, and genuinely beneficial to analysts.

Table 3: Proposed Multidimensional Evaluation Framework for XAI-Based Intrusion Detection Systems

Dimension	Evaluation question	Possible metric
Detection performance	Does the model accurately distinguish attacks from legitimate traffic?	Accuracy, precision, recall, F1-score, false-positive rate
Fidelity	Does the explanation accurately represent the behavior of the underlying model?	Explanation-model agreement; faithfulness score
Stability	Do similar inputs and predictions produce reasonably consistent explanations?	Stability or consistency score; feature-rank correlation
Comprehensibility	Can the intended security analyst understand the explanation?	Human evaluation; explanation complexity; task comprehension
Actionability	Does the explanation help an analyst make a better security decision?	Response accuracy; investigation time; decision quality
Efficiency	Can an explanation be produced rapidly enough for operational IDS use?	Explanation-generation latency; computational overhead
Robustness	Does the explanation remain reliable under noise or adversarial manipulation?	Perturbation tests; adversarial robustness score
Generalizability	Do model explanations remain meaningful across datasets and network environments?	Cross-dataset consistency; external validation
Trust calibration	Does analyst confidence appropriately correspond to actual model reliability?	Trust-performance relationship; confidence calibration

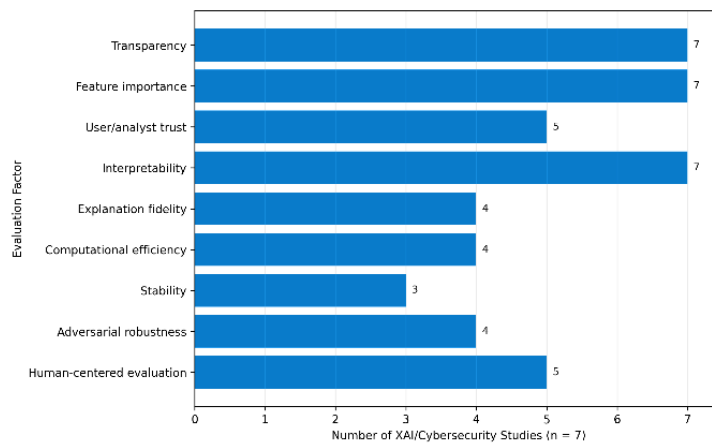


Fig 3 - Frequency of Explainability and Trust-Related Factors Reported in X-IDS Literature

The graph shows that transparency, feature importance, and interpretability are the most frequently emphasized factors in X-IDS research, while stability receives comparatively less attention.

6. CHALLENGES AND FUTURE RESEARCH DIRECTIONS

The integration of Explainable Artificial Intelligence into intrusion detection systems offers an important path toward greater transparency, accountability, and operational trust. However, several unresolved challenges limit the practical deployment of explainable intrusion detection systems in real-world cybersecurity environments. These challenges extend beyond model accuracy and include the quality, stability, efficiency, security, and human usability of explanations. Addressing them is essential if XAI is to move from experimental demonstrations to dependable security operations.

6.1. Lack of Standardized XAI Metrics

One of the most significant limitations in current XAI research is the absence of universally accepted criteria for defining a “good” explanation. Carvalho et al. (2019) showed that interpretability is assessed using diverse and often incompatible measures, while Arrieta et al. (2020) emphasized that explainability involves multiple dimensions, including transparency, comprehensibility, fidelity, and user relevance. This creates a serious problem for intrusion detection research because two XAI methods may produce explanations that appear useful but cannot be compared objectively.

Neupane et al. (2022) similarly identified the lack of standardized evaluation procedures as a major challenge for explainable intrusion detection systems. Future studies should therefore move beyond visual demonstrations of feature importance and adopt benchmark protocols that assess explanation fidelity, stability, consistency, computational cost, and human usefulness. Standardized cybersecurity XAI benchmarks would allow methods to be compared across models and datasets under consistent conditions.

6.2. Explanation Stability and Reliability

An explanation must not only be understandable but also reliable. Similar network events that receive similar predictions should ideally generate similar explanations. If small and operationally insignificant input changes cause major changes in the explanation, security analysts may question whether the explanation reflects genuine model reasoning.

This issue is particularly important for local post-hoc methods, which may be sensitive to sampling strategies, perturbations, or feature dependence. Future X-IDS research should therefore evaluate explanation stability systematically by introducing controlled perturbations into benign and malicious traffic samples and measuring how strongly explanations change. Reliability testing should become a standard component of XAI evaluation because unstable explanations may undermine the trust that explainability is intended to support.

6.3. Computational Overhead and Real-Time Detection

Cybersecurity systems frequently operate under strict timing constraints. Enterprise networks can generate extremely large volumes of traffic, and an IDS may be required to analyze events continuously with minimal delay. Although LIME and SHAP can provide informative explanations, they may require repeated model evaluations or complex feature-attribution computations. If explanation generation significantly increases latency, the practical value of the IDS can be reduced.

The challenge becomes even more important for online detection architectures. Mirsky et al. (2018), for example, demonstrated an ensemble of autoencoders for online network intrusion detection. Systems operating in such environments cannot afford explanation mechanisms that delay threat response. Future research should explore lightweight approximations, selective explanation of high-risk alerts, parallel processing, explanation caching, and hardware acceleration. Explanation latency should also be reported alongside conventional detection latency.

6.4. Adversarial Risks of Explainability

Explainability creates a security paradox in cybersecurity applications. Greater transparency helps legitimate analysts understand model decisions, but the same information may reveal sensitive details about the detector to an attacker. Marino et al. (2018) highlighted the adversarial dimension of explainable AI in intrusion detection, showing that explanation mechanisms must themselves be treated as part of the security surface.

An attacker who learns which features strongly influence classification may modify traffic to reduce those signals and improve the chance of evasion. Consequently, explainability should not imply unrestricted disclosure. Practical X-IDS architectures should incorporate controlled explanation access, role-based interfaces, secure logging, and different levels of explanation detail for different users. Sensitive model information should be protected, and explanation systems should be tested against adversarial probing and manipulation.

6.5. Dataset Limitations and Generalizability

The quality of an explanation depends partly on the representativeness of the data used to train the detector. Many intrusion detection studies rely on benchmark datasets that do not fully reproduce contemporary enterprise traffic, cloud environments, encrypted communication, or rapidly evolving attack techniques. Some datasets also suffer from class imbalance, duplicated observations, synthetic traffic, or limited attack diversity.

These limitations create a risk that an explanation may be faithful to a model while describing patterns that do not generalize to real networks. Future research should validate both detection performance and explanation behavior across multiple datasets. Cross-dataset testing is essential for determining whether influential features remain meaningful under different network conditions,

attack families, and traffic distributions. Realistic and continuously updated cybersecurity datasets will therefore be critical for trustworthy XAI evaluation.

6.6. Human-Centered XAI

Technical explainability does not automatically translate into operational usefulness. Future research should include real security operations center analysts rather than relying only on researcher judgments of whether explanations appear understandable. Human-centered studies should measure whether XAI improves the speed and quality of incident investigation, reduces cognitive burden, and supports better decisions.

Relevant measures should include alert investigation time, decision accuracy, analyst confidence, explanation comprehension, and cognitive load. Studies should also determine whether explanations encourage appropriate trust rather than blind reliance. Different stakeholders may require different explanation formats, so adaptive interfaces should be tested with analysts, engineers, managers, and auditors.

6.7. Future Research Agenda

Future X-IDS research should pursue a coordinated agenda combining technical innovation with operational validation. Priorities include real-time XAI methods, explainable deep-learning architectures, causal explanations, and hybrid systems that combine powerful black-box detectors with transparent interpretable components. Adversarially robust explanation mechanisms should also be developed so explanation systems cannot easily be manipulated or exploited.

Further work is needed on stakeholder-specific explanations, cross-dataset explanation testing, and quantitative trust-calibration metrics that determine whether user confidence corresponds to actual system reliability. Human-in-the-loop security operations should be studied under realistic workloads, while explainability should expand beyond conventional enterprise networks to IoT, cloud, edge, and industrial control environments. Ultimately, the next generation of explainable intrusion detection systems should be evaluated not only by whether they detect attacks accurately, but also by whether their explanations are stable, timely, secure, generalizable, understandable, and genuinely useful to cybersecurity professionals.

7. CONCLUSION

Artificial intelligence has become increasingly important to modern cybersecurity because the scale, speed, and complexity of network activity now exceed the capabilities of purely manual or rule-based threat detection. Machine learning and, in particular, deep learning have significantly improved the ability of intrusion detection systems to recognize complex, nonlinear, and previously unseen attack patterns across large volumes of network traffic. These capabilities make AI-based intrusion detection highly valuable for contemporary security operations. However, stronger predictive performance has also introduced a critical limitation: many advanced models operate as black boxes,

producing alerts or classifications without providing sufficient insight into the reasoning behind their decisions.

This lack of transparency can weaken trust in automated security systems, especially when analysts must decide whether to block traffic, isolate devices, escalate incidents, or dismiss alerts as false positives. Explainable Artificial Intelligence offers a practical response to this challenge by making model behavior more understandable, traceable, and assessable. Techniques such as LIME can provide local explanations for individual intrusion alerts, while SHAP can quantify the contribution of specific features to both individual and global model behavior. Intrinsically interpretable approaches, including decision trees and rule-based models, provide more direct transparency, while feature-importance methods can help reveal which network characteristics most strongly influence model predictions. Together, these approaches improve the ability of security professionals to inspect, validate, and question automated decisions.

The analysis further shows that effective explainability must be designed around the needs of cybersecurity analysts rather than around technical sophistication alone. An explanation is useful only when it helps the intended user understand why an event was flagged, which features were most influential, how the event differs from normal behavior, and whether the model's reasoning is consistent with known attack patterns. For this reason, the evaluation of explainable intrusion detection systems should extend well beyond conventional metrics such as accuracy, precision, recall, and F1-score. High detection performance remains essential, but it does not guarantee that a system is trustworthy, understandable, stable, or operationally useful.

This study therefore emphasizes a multidimensional evaluation perspective for X-IDS. In addition to predictive effectiveness, explainable intrusion detection systems should be assessed in terms of explanation fidelity, stability, comprehensibility, actionability, computational efficiency, robustness, security, generalizability, and trust calibration. Such an evaluation framework is necessary because an explanation may be technically correct yet too complex to use, easy to manipulate, computationally expensive, or inconsistent across similar events.

Several challenges remain unresolved. These include the absence of standardized explainability metrics, instability in some post-hoc explanation methods, computational overhead in real-time environments, variation in stakeholder explanation needs, dataset limitations, and the risk that detailed explanations may expose defensive logic to adversaries. Addressing these issues will require closer integration of machine learning, cybersecurity engineering, human-computer interaction, and security operations research.

Ultimately, explainability should not be treated as a cosmetic layer added after an intrusion detection model has been developed. It should be considered a core design requirement from the earliest stages of system development, alongside detection accuracy, latency, robustness, and security.

The future of trustworthy AI-based cybersecurity will depend not only on whether an intrusion detection system can identify threats accurately, but also on whether its reasoning can be understood, validated, challenged, and safely acted upon by the people responsible for defending digital environments.

REFERENCES

- [1] Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36-43.
- [2] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1-38.
- [3] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
- [4] Potla, R. (2023). Designing a BTP-centric integration mesh for shop-floor IoT, MES and ERP in discrete manufacturing. *Journal of Artificial Intelligence, Machine Learning and Data Science*, 1(2), 1-8.
- [5] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- [6] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5), 206-215.
- [7] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115.
- [8] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6, 52138-52160.
- [9] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 1-42.
- [10] Potla, R. B. (2024). A SOX/ITAR-aligned global ERP template for multi-plant manufacturers: Governance patterns and controls. *J Artif Intell Mach Learn & Data Sci*, 2(2), 3222-3232.
- [11] Jadala, S. K. (2026). Data-Centric Security for Generative AI Systems: Protecting Sensitive Data against Leakage, Inference Attacks, and Unauthorized Model Access. *International Journal of Data Engineering and Intelligent Computing*, 9(3), 01-32.
- [12] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009, July). A detailed analysis of the KDD CUP 99 data set. In *2009 IEEE symposium on computational intelligence for security and defense applications* (pp. 1-6). Ieee.
- [13] Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys & tutorials*, 18(2), 1153-1176.
- [14] Moustafa, N., & Slay, J. (2015, November). UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 military communications and information systems conference (MilCIS)* (pp. 1-6). Ieee.
- [15] Liao, H. J., Lin, C. H. R., Lin, Y. C., & Tung, K. Y. (2013). Intrusion detection system: A comprehensive review. *Journal of network and computer applications*, 36(1), 16-24.
- [16] Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security*, 28(1-2), 18-28.
- [17] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80-89). IEEE.
- [18] Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. *IEEE access*, 5, 21954-21961.
- [19] Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A deep learning approach to network intrusion detection. *IEEE transactions on emerging topics in computational intelligence*, 2(1), 41-50.

- [20] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *IEEE access*, 7, 41525-41550.
- [21] Ferrag, M. A., Maglaras, L., Moschogiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*, 50, 102419.
- [22] Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). Kitsune: An ensemble of autoencoders for online network intrusion detection. *arXiv preprint arXiv:1802.09089*.
- [23] Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*, 2(1), 20.
- [24] Aldweesh, A., Derhab, A., & Emam, A. Z. (2020). Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. *Knowledge-Based Systems*, 189, 105124.
- [25] Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. *Computers & security*, 86, 147-167.
- [26] Wang, M., Zheng, K., Yang, Y., & Wang, X. (2020). An explainable machine learning framework for intrusion detection systems. *IEEE access*, 8, 73127-73141.
- [27] Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, 2021(1), 6634811.
- [28] Marino, D. L., Wickramasinghe, C. S., & Manic, M. (2018, October). An adversarial approach for explainable ai in intrusion detection systems. In *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society* (pp. 3237-3243). IEEE.
- [29] Amarasinghe, K., & Manic, M. (2018, October). Improving user trust on deep neural networks based intrusion detection systems. In *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society* (pp. 3262-3268). IEEE.
- [30] Potla, R. B. (2024). Optimizing extended warehouse management for make-to-order plants: Slotting, wave picking, and yard orchestration at scale. *Journal of Computer Science and Technology Studies*, 6(3), 181-192.
- [31] Jadala, S. K. (2025). Ransomware Resilience in Cloud and Enterprise Systems: Detection, Response, and Recovery Frameworks. *Journal of Cyber-Physical Security and Robotics*, 1(02), 63-86.
- [32] Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEE Access*, 10, 93104-93139.
- [33] Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. *IEEE Access*, 10, 112392-112415.
- [34] Jadala, S. K. (2024). Zero Trust Architecture and Identity Threat Detection for Securing Cloud, IoT, and Hybrid Enterprise Systems. *International Journal of Humanities and Information Technology*, 6(04), 141-185.
- [35] Charmet, F., Tanuwidjaja, H. C., Ayoubi, S., Gimenez, P. F., Han, Y., Jmila, H., ... & Zhang, Z. (2022). Explainable artificial intelligence for cybersecurity: a literature survey. *Annals of Telecommunications*, 77(11), 789-812.
- [36] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 832.