International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Explainable AI (XAI) for Enhanced Cyber Threat Intelligence: Building Interpretable Intrusion Detection Systems

Naresh Kalimuthu

Naresh.kalimuthu@gmail.com

Abstract:
The growth of complex cyber threats calls for integrating advanced Artificial Intelligence (AI) and Machine Learning (ML) technologies into Cyber Threat Intelligence (CTI) frameworks, especially for Intrusion Detection Systems (IDS). While these models, particularly deep learning architectures, achieve high accuracy in identifying complex and unprecedented attacks, the "black-box" nature creates trust, adoption, and operational challenges. This lack of transparency often results in opaque decision-making, diminishes trust among security personnel, and increases alert fatigue, weakening the security these systems aim to provide. An emerging body of work in Explainable AI (XAI) addresses this issue by offering explanations for AI-driven decisions and actions. This paper examines the integration of XAI into IDS to enhance trust, collaboration, and resilience in cyber defense systems. It outlines three primary research challenges that have so far impeded the development of Explainable IDS (X-IDS): balancing model accuracy with system fidelity, meeting the technical requirements for real-time XAI processing, and establishing standards to evaluate explanation quality.
This paper critically reviews various mitigation strategies, including Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) post-hoc explainability frameworks, attention-based explainable AI models, and emerging federated and lightweight XAI paradigms. Syntheses of findings from CICIDS2017, UNSW-NB15, and other benchmark data that illustrate XAI's potential to improve forensic analysis and reduce false positive rates without compromising detection accuracy. The paper argues that XAI is essential for the future of cyber threat intelligence (CTI), enabling IDS to evolve from opaque alert generators into trustworthy partners alongside human analysts. It calls for future research to develop federated, lightweight real-time XAI, unified benchmarking frameworks for XAI evaluation, and defenses against adversarial XAI sabotage.

Keywords: Explainable AI (XAI), Intrusion Detection System (IDS), Cyber Threat Intelligence (CTI), Machine Learning, Interpretability, SHAP, LIME, Network Security.

I. INTRODUCTION

A. The Changing Face of Cyber Land and The Use of AI on Intrusion Detection Systems (IDS)

Every day, cyber threats and attacks become increasingly sophisticated, requiring defenders to rethink how they protect systems and data. Cyber criminals employ a wide range of attacks to dismantle critical infrastructure and disrupt societal functions, leveraging advanced malware, banking Trojans, ransomware, and targeted, sophisticated phishing campaigns. The sheer volume, speed, and dynamic nature of these threats demand a shift from manual, reactive security systems to automated, intelligent, and proactive solutions. As a result, many organizations are deploying AI and Machine Learning (ML) to strengthen and improve their core Cyber Threat Intelligence (CTI) functions.
Leading the charge in this paradigm shift are AI-driven Intrusion Detection Systems (IDS). These systems have evolved significantly from their origins as simple signature and rule-based systems. Modern IDS uses a variety of advanced algorithms—Deep Neural Networks (DNNs), Convolutional Neural Networks

(CNNs), and Recurrent Neural Networks (RNNs)—to perform real-time analysis of vast amounts of network traffic data in sophisticated ways. In this manner, they can independently detect highly complex, non-linear patterns and numerous subtle ambiguities that are virtually indistinguishable from routine activities to humans and traditional detection systems. The proven effectiveness of these advanced models in identifying both known and zero-day threats has made them essential parts of contemporary cybersecurity systems.

B. The Black-Box Consternation: Inequity of Trust, Fog, and Alert Fatigue

ML models give modern IDS high predictive accuracy, but they remain mostly opaque and act as black boxes. The complexity that helps achieve this—millions of parameters in layered, nonlinear structures—makes the internal decision processes impossible to understand. When a black-box IDS detects a threat, it usually provides a classification and risk score but no explanation, reasoning, or evidence to support the decision.

The inability to explain the black-box problem is a major barrier to the effective use of AI in cybersecurity. There is a significant lack of trust in the system among security practitioners, who are the main users. Disengagement among Security Operations Center (SOC) analysts, who are usually responsible for explaining alerts, is understandable, especially when incident response involves drastic measures such as shutting down a critical server. This lack of transparency prevents analysts from using system outputs to improve their own logical frameworks and decision-making, creating a strong disconnect between the person and the system.

The operational result of trust deficit is a serious and widespread phenomenon called "alert fatigue." An overwhelming number of alerts are sent, many of which are falsely flagged. Threats and harmless anomalies are poorly distinguished, if at all, leading to attention fragmentation. Each alert, regardless of its real importance, demands the same level of scrutiny, which becomes a problem because all alerts from the black box are given the same arbitrary importance level, based on the system sifting through thousands. This not only leads to the absurdly inefficient use of analytics resources but also raises false alarms, making it harder for defenders to identify actual critical threats and potentially preventing them from being detected. The real problem isn't the alerts themselves but the lack of actionable intelligence within them, which results from the underlying model's opacity.

C. The benefits of having Explainable AI (XAI) for clear cyber defense.

Explainable AI (XAI) is one of the newest and most vital tools in scholarship focused on solving the black box problem. The main aim of XAI is to develop techniques and models that can generate clear, interpretable, and human-readable explanations of their predictions and decisions. In cyber defense, XAI is not merely a theoretical pursuit; it is an urgent necessity for fully harnessing the power of artificial intelligence.

The integration of XAI practices within IDS aims to transform them from opaque alert systems into transparent, reliable, and cooperative allies of security analysts. An explainable IDS (X-IDS) can answer the question of why an alert is raised and highlight the features, patterns, or data points that support that conclusion and the alert's rationale. This capability enables analysts to quickly validate alerts, effectively prioritize threats, and implement more appropriate and efficient countermeasures. Increased transparency enhances the collaborative relationship between humans and machines—AI handles fast data processing while humans provide essential context for balanced reasoning and guidance. Such transparency is also crucial for accountability, legal compliance, such as with GDPR, and overall system responsibility. This is vital for developing more adaptive and effective security systems.

D. The Goals and Plan of the Paper

This research not only explores the intersection of XAI and IDS but also offers a comprehensive review and synthesis of the current field. It begins by identifying three main research challenges that hinder the effective and widespread adoption of X-IDS. Subsequently, the paper introduces a set of solutions and advanced architectural frameworks designed to address these challenges. It includes empirical results from key case studies that utilize XAI methods with benchmark cybersecurity datasets, demonstrating their effectiveness and value in real-world scenarios.

II. FUNDAMENTAL RESEARCH ISSUES IN EXPLAINABLE IDS (X-IDS) DEVELOPMENT

The challenges associated with the development of and the need for Explainable Intrusion Detection Systems (X-IDS) are multifaceted and of great importance. They are not simply engineering problems; these challenges sit at the nexus of machine learning, the boundaries of what can be computed, and how cybersecurity is actually practiced.

A. Challenge 1: Navigating the Accuracy-Interpretability Trade-off

One of the most apparent issues in the world of XAI is the conflict between accuracy and interpretability. This problem is even more significant in intrusion detection. On the one hand, the most accurate models, such as Deep Neural Networks and Gradient Boosting Machines, excel at detecting modern cyberattacks and capturing their complex, high-dimensional, and nonlinear patterns. These models have very high detection rates and accuracy, but their inner workings lack explainability and therefore remain opaque.

Unlike models that are “black-box" by nature, “white-box" models can explain their reasoning. They include older methods, like Decision Trees, which provide human-readable rule sets such as IF-THEN, and Linear and Logistic Regression systems, through which one can easily understand feature weights and other rule-based systems. These models explain their predictions, but in a world of complex predictive models, they fall significantly short of accuracy. In such cases, more advanced models, like black-box models, are necessary, especially for handling sophisticated cyberattacks and "low-and-slow" systems that lack overt malicious signatures.

This reliance on transparency can challenge security architects and practitioners, forcing them to choose between a high-performing system with limited trust or a transparent system with reduced effectiveness. Achieving a balance between these competing goals remains an important focus for research in the area of design X-IDS.

B. Challenge 2: Real-Time Scalability and Computational Overhead

The utility of an IDS is determined by its ability to perform in real time and under high network data loads. Many prominent and advanced XAI approaches, especially post-hoc, model-agnostic approaches such as SHAP and LIME, incur significant processing costs that can compromise their real-time capabilities.

Generating an explanation for a single prediction involves running hundreds, if not thousands, of permutations of the input data and observing the model's output at each instance. Such a method, while conducive to deriving feature importance, is a huge detriment to the detection pipeline in a SOC configuration. An IDS is designed to analyze millions, even tens of millions of network flows within a single second, and the moment there is any latency in the several milliseconds range, the system is bound to face a substantial bottleneck. The direct application of many XAI approaches is on real, live, and line-rate network monitoring, and the expected computational overhead is extremely high.

Additionally, the boundaries of the methods are becoming a more significant issue. As network speeds increase and the dimensionality of cybersecurity datasets (i.e., the number of extracted features) continues

to grow, the computational cost of generating explanations escalates. This challenge has driven research into developing optimized, lightweight explainability methods and alternative architectures that deliver interpretability while meeting the strict performance and scalability requirements of modern cybersecurity operations.

C. Challenge 3: The Lack of Uniformity in Evaluating Explanations

Perhaps one of the most critical challenges in the field of XAI is the lack of a universal, accepted, and standardized evaluation framework for analyzing the "quality" of an explanation. The current evaluation landscape is splintered, relies on subjective human evaluation, and is characterized by inconsistent, arbitrary rules and metrics. This fragmentation makes it exceedingly difficult to conduct a fair and objective assessment of disparate XAI techniques or to assess the usefulness of an explanation for a particular security task. The lack of evaluation standards makes it impossible to reliably evaluate the utility of the generated explanation, which, in turn, leads to the problem of being plausible yet not loyal, manipulable yet meant to be trusted, and far removed from realism.

The lack of standard evaluation metrics allows us to analyze the three challenges pursued here as a single, interrelated problem. If standard metrics do not exist to assess interpretability quality, one cannot conduct a principled study in the accuracy-interpretability synthesis. The deliberate choice of sacrificing "better explanations" for a 4% reduction in detection accuracy is only a subjective guess, not a reasoned decision grounded in data. And so is the justification for the SHAP technique, which requires us to allocate considerable computational resources and infrastructure. There is also a lack of value justification in SHAP when we attempt to compute it as the value it yields for the investment in explanation quality.

In an effort to clarify this metric's shortcomings, some academics have begun proposing more clearly defined frameworks to explain exotic attributes. These frameworks outline constructionist approaches to the technical validity and functional applicability of insights generated by XAI on multiple interdependent dimensions.

III. MITIGATION STRATEGIES AND ADVANCED ARCHITECTURAL FRAMEWORKS

Regarding the construction of an operational X-IDS, a broader range of techniques and architectural frameworks has been developed. These techniques can be divided into three categories: approaches that add explanations to preexisting opaque models, approaches that intentionally craft interpretable systems, and new approaches that address concerns such as edge computing and privacy.

A. Post-Hoc Explainability for High Performance Models

The most common approach to constructing an X-IDS system is to retain black-box models with high performance and complexity, and then apply a separate, agnostic XAI technique to generate explanations post hoc (after a prediction is made). This approach attempts to accomplish the "best of both worlds" scenario, benefitting from the high accuracy of advanced models, while still providing sufficient transparency for human review. The two most prevalent techniques in this space are Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP).

1) Local Interpretable Model-agnostic Explanations (LIME): LIME is based on a very simplistic and powerful assumption: although a global model may be extremely complex, its behavior around the vicinity of one data point (and only by one data point) can be approximated by a very simple, easy-to-understand, and explainable model (like a linear regression). To construct an explanation for a prediction for a particular network flow, LIME creates a "neighborhood" of perturbed samples around the flow, queries the black box model for predictions on these samples, and then fits an interpretable model on this "local" dataset. The explanation is then taken from the local model.

The main advantage of LIME is that it provides easy-to-understand outputs, providing instance-level rationales that are actionable and easy for domain specialists to work on. However, LIME's explanations are often unstable, and small perturbations on the instance being rationalized can significantly change the local approximations of the explanation. Moreover, LIME's explanations are so local that they might miss the model's overall reasoning.

2) SHapley Additive exPlanations (SHAP): SHAP offers a unified, theoretically grounded framework for interpreting models, based on Shapley values from cooperative game theory. For each prediction, SHAP determines the marginal contribution of each feature, fairly allocating the "credit" for the output among input features. This method has strong theoretical backing, ensuring properties such as local accuracy, where the sum of feature contributions matches the model's output, and consistency, meaning a feature's importance won't decrease as its actual impact increases. SHAP provides both local explanations for individual predictions and global insights into feature importance. While often more robust and reliable than LIME, this approach is computationally intensive, which can hinder its use in real-time.

TABLE I. COMPARATIVE ANALYSIS OF LIME AND SHAP FOR IDS

Dimension	LIME (Local Interpretable Model-agnostic Explanations)	SHAP (SHapley Additive exPlanations)
Theoretical Foundation	Based on local surrogate models; approximates the black-box model.	Based on cooperative game theory (Shapley values); provides mathematically sound feature attributions.
Explanation Scope	Primarily local (explains individual predictions). Global understanding is inferred by explaining many individual instances.	Provides both local (for individual predictions) and global (overall feature importance) explanations consistently.
Computational Cost	Generally faster and less computationally intensive than SHAP.	Significantly more computationally expensive, especially KernelSHAP for non-tree models, limiting real-time use.
User Preference / Usability	Often preferred by users for its more intuitive and user-friendly visual explanations.	Can be more technical, but its consistency is valued by expert users. Provides powerful visualizations like force plots.
Consistency/ Reliability	Explanations can be unstable, as small changes in the input can lead to different local approximations.	Guarantees properties like local accuracy and consistency, making explanations more reliable and robust.
Vulnerability	Vulnerable to adversarial attacks designed to generate misleading explanations.	Also vulnerable to adversarial attacks, though some studies suggest it may be slightly less robust than LIME in certain scenarios.

B. Inherently Interpretable by Design: From Decision Trees to Attention Mechanisms

An alternative to post-hoc explanation is to develop IDS models that are inherently interpretable or self-explanatory.

1) Traditional Interpretable Models:

The simplest method is to use classic ML models with transparent structures. For example, Decision Trees generate explicit, human-readable rules that trace input features to a final classification. Although they may not perform as well as deep learning models on complex tasks, their transparency makes them useful for baseline comparisons, regulatory audits, or deployment where interpretability is essential.

2) Attention Mechanisms:

A more advanced approach to linking deep learning and interpretability is using attention mechanisms. Inspired by human focus and attention, layers can be added to neural networks (such as RNNs or Transformers). These layers allow the model to learn and assign importance weights to different parts of the input during prediction. For instance, when analyzing network packet sequences, attention can identify which packets or fields are most indicative of an attack. These weights can be visualized as heatmaps, providing a detailed explanation of the input features the model prioritized, thus offering insight into the model's reasoning without the heavy computational cost of post-hoc methods.

C. Emerging Paradigms: Federated and Lightweight XAI

As the cybersecurity landscape continues to evolve, new architectural paradigms are emerging to address specific challenges such as data privacy and the growth of edge computing.

1) Federated Learning (FL) for Privacy-Preserving X-IDS:

Traditional IDS models depend on aggregating large amounts of potentially sensitive network data for training, which raises privacy and security concerns. Federated Learning (FL) offers a decentralized alternative by facilitating collaborative training of a global IDS model across multiple distributed clients—such as different organizations or network segments—without sharing their raw local data. When integrated with Explainable AI (XAI) techniques, such as SHAP for interpreting predictions of locally trained models, this approach helps develop a robust, privacy-preserving, and transparent X-IDS framework.

2) Explainable and Lightweight AI (ELAI) for Edge Computing:

As IoT and edge computing grow, security monitoring is moving from centralized data centers to devices at the network's edge with limited resources. This shift requires the creation of Explainable and Lightweight AI (ELAI) frameworks that develop models that effectively balance detection accuracy, computational efficiency, and interpretability. These models often employ hybrid architectures that combine straightforward, understandable models like decision trees with efficient, lightweight deep learning techniques optimized for low-power devices, ensuring that security intelligence remains both powerful and explainable at the network's edge.

IV. CASE STUDY FINDINGS: EMPIRICAL VALIDATION ON BENCHMARK DATASETS

The theoretical advantages of XAI in cybersecurity are supported by an increasing body of empirical research that applies these techniques to benchmark IDS datasets. These studies offer essential insights into the real-world performance, usefulness, and limitations of X-IDS.

A. Interpreting Intrusion Alerts on the CICIDS2017 Dataset

The CICIDS2017 dataset is a widely used benchmark for assessing IDS performance. It addresses the shortcomings of previous datasets by featuring benign network traffic from realistic user profiles and a

broad range of recent, common attacks, including Brute Force, Denial-of-Service (DoS), Web Attacks, and Port Scans, collected over five days.

Many studies have successfully applied XAI techniques to models trained on the CICIDS2017 dataset. For example, researchers have utilized LIME to interpret the predictions of an ensemble model with 96.25% accuracy, showing that high performance and interpretability can go hand in hand. Other research has used SHAP and LIME to pinpoint the most influential network features for classifying various attack types, offering security analysts clear cues to differentiate between a DoS flood and a port scan. An important validation method in these studies is perturbation analysis, which involves systematically removing or changing features identified as key by XAI methods and examining the impact on the model's output. Experiments with Multi-Layer Perceptron (MLP) models have demonstrated that altering the top features identified by LIME and SHAP consistently affects classification, confirming the reliability of the explanations.

However, the effectiveness of XAI depends heavily on the quality of the data used to train the models. Notably, detailed analysis of the CICIDS2017 dataset has uncovered significant issues in its creation pipeline. Errors in traffic generation, feature extraction, and labeling have been identified, with one study noting that over 25% of network flows in the dataset are meaningless artifacts from data collection. This highlights a cautionary lesson for the XAI field. An XAI tool can provide an accurate explanation of a model's decision, but if that decision is based on false correlations or data artifacts rather than true malicious activity, the explanation is ultimately useless. The explanation might be factually correct ("the model flagged this because of artifact X"), but semantically misleading, potentially fostering false confidence and incorrect security assumptions. This emphasizes the need for high data quality and thorough preprocessing to ensure meaningful explainability.

B. Enhancing Forensic Analysis with XAI on the UNSW-NB15 Dataset

The UNSW-NB15 dataset serves as a crucial benchmark designed to emulate contemporary network attack scenarios. It combines real network traffic with synthetic attack data spanning nine categories, including Fuzzers, Backdoors, and Worms. Researchers frequently use this dataset to demonstrate a significant application of XAI: digital forensics and incident response.

In forensic investigations, proof is essential. Evidence from AI systems needs to be transparent, auditable, and defensible for legal purposes. Traditional black-box AI models make this difficult. Research applying Explainable AI (XAI) to models trained on UNSW-NB15 addresses this challenge. A comparison of SHAP and LIME on high-performance models such as XGBoost highlights their complementary strengths. SHAP provides stable, globally consistent feature importance rankings, helping to identify attack vectors broadly. LIME offers detailed explanations for individual cases, assisting investigators with specific malicious events. Using both methods creates a comprehensive, multi-layered narrative of security incidents, ensuring an auditable, legally defensible trail crucial for forensic investigations.
This demonstrates how selecting a dataset shapes the way the problem XAI aims to address is approached.

C. Synthesis of Performance Metrics Across Studies

Various case studies reveal key performance trends in integrating XAI into IDS.

1) Detection Accuracy:

A significant concern is that making a model explainable could compromise its detection performance. However, empirical evidence shows this isn't necessarily true. Post-hoc XAI methods like LIME and SHAP do not modify the underlying model, thereby maintaining accuracy. Sometimes, insights gained from XAI can even enhance the model. For instance, one study reported a 15% increase in detection accuracy after adding LIME to an ensemble model, suggesting that explainability can help refine models.

2) False Positive Rate (FPR):

XAI significantly lowers high false positive rates by providing explanations for alert reasons to analysts. This helps analysts rapidly and accurately identify true threats, minimizing alert fatigue and enabling them to concentrate on genuine dangers.

3) Interpretability and Usability:

While difficult to quantify, explanation usability is crucial and often assessed through qualitative feedback from security analysts. User studies show that explanations build trust and aid in result interpretation. Typically, users appreciate LIME's visually intuitive and easy-to-understand explanations, whereas technical experts and data scientists might prefer SHAP for its theoretical accuracy, despite being more complex. This highlights the importance of customizing explanations for different audiences.

V. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

A. Recapitulation of Findings: The Imperative for Explainability in CTI

The adoption of AI and ML in cybersecurity has led to remarkable detection capabilities. Nonetheless, this advancement is hindered by the "black-box" issue, where complex models improve accuracy but reduce trust, transparency, and effectiveness. This paper systematically examines Explainable AI (XAI) in Intrusion Detection Systems (IDS), emphasizing that explainability is essential, not optional, for implementing AI effectively in security settings that involve human users.

The analysis indicates that XAI directly tackles key operational issues like trust deficits and alert fatigue by converting opaque predictions into useful insights. By closing the cognitive gap between complex model calculations and a human analyst's need for justification, XAI promotes a collaborative defense approach. This teamwork improves Cyber Threat Intelligence (CTI), supports quicker and more confident incident responses, and establishes the essential trust needed for effective human-AI collaboration in the face of constantly changing threats. Although challenges such as performance trade-offs, computational demands, and evaluation standards remain, the strategies and frameworks outlined here point toward a promising and clear way forward.

B. Future Outlook: Towards Robust, Scalable, and Standardized X-IDS

The ongoing progress and implementation of X-IDS rely heavily on collaborative research in several vital areas.

1) The Need for a Unified Evaluation Framework:

One of the most pressing and influential priorities for future work is to create and adopt a standardized framework to assess XAI methods, specifically in cybersecurity. Currently, the piecemeal approach hampers advancement and makes dependable comparisons difficult. An ideal framework should go beyond subjective judgments and incorporate a broad range of quantitative metrics that evaluate faithfulness, robustness, complexity, stability, and other essential qualities, as discussed in this paper. Importantly, these metrics must be adapted to meet the varied needs of different security professionals, from the real-time tasks of a SOC analyst to the evidentiary standards required by forensic investigators.

2) Securing XAI: Defending Against Adversarial Manipulation:

The transparency of XAI has a dual nature. It empowers defenders but also introduces a new attack vector for adversaries. Malicious actors could exploit explanation insights to craft more effective evasion strategies or, more subtly, execute "adversarial explanation" attacks that aim to deceive the explanation method. This could mislead analysts and conceal malicious activities. Therefore, future research should prioritize enhancing the security and resilience of XAI systems. This involves developing techniques to

identify and counteract adversarial manipulation of explanations, and to develop inherently robust XAI methods capable of providing reliable insights even in hostile environments.

3) Real-Time, Lightweight Explainability for Next-Generation SOCs:

The high computational demands of many post-hoc XAI techniques hinder their widespread use in real-time security settings. The future of operational XAI depends on creating new, lightweight approaches that can run efficiently on high-speed networks and resource-limited edge devices. This likely means moving away from methods that rely heavily on intensive post-hoc analysis. Instead, research should focus on integrating explainability directly into the architecture, such as by designing interpretable deep learning models or enhancing attention mechanisms that deliver valuable insights with minimal extra computational effort.

REFERENCES:

1. Rjoub, G., Bentahar, J., Wahab, O. A., Mizouni, R., Song, A., Cohen, R., Otrok, H., & Mourad, A. (2023). A Survey on Explainable Artificial Intelligence for Cybersecurity. *ArXiv*. https://doi.org/10.1109/TNSM.2023.3282740
2. S. Samtani, H. Chen, M. Kantarcioglu and B. Thuraisingham, "Explainable Artificial Intelligence for Cyber Threat Intelligence (XAI-CTI)," in IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 4, pp. 2149-2150, 1 July-Aug. 2022, doi: 10.1109/TDSC.2022.3168187.
3. S. B. Remman, I. Strümke and A. M. Lekkas, "Causal versus Marginal Shapley Values for Robotic Lever Manipulation Controlled using Deep Reinforcement Learning," 2022 American Control Conference (ACC), Atlanta, GA, USA, 2022, pp. 2683-2690, doi: 10.23919/ACC53348.2022.9867807.
4. Neupane, Subash & Ables, Jesse & Anderson, William & Mittal, Sudip & Rahimi, Shahram & Banicescu, Ioana & Seale, Maria. (2022). Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities. IEEE Access. 10. 112392-112415. 10.1109/ACCESS.2022.3216617.
5. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012.
6. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *ArXiv*. https://arxiv.org/abs/1602.04938
7. Lundberg, S., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. *ArXiv*. https://arxiv.org/abs/1705.07874
8. Gaspar, Diogo & Silva, Paulo & Silva, Catarina. (2024). Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3368377.
9. Hermosilla, Pamela & Berríos, Sebastián & Allende-Cid, Héctor. (2025). Explainable AI for Forensic Analysis: A Comparative Study of SHAP and LIME in Intrusion Detection Models. Applied Sciences. 15. 7329. 10.3390/app15137329.
10. Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. *Frontiers in Artificial Intelligence*, *8*, 1526221. https://doi.org/10.3389/frai.2025.1526221
11. Wang, Maonan & Zheng, Kangfeng & Yang, Yanqing & Wang, Xiujuan. (2020). An Explainable Machine Learning Framework for Intrusion Detection Systems. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2988359.

12. Fatema, K.; Dey, S.K.; Anannya, M.; Khan, R.T.; Rashid, M.M.; Su, C.; Mazumder, R. Federated XAI IDS: An Explainable and Safeguarding Privacy Approach to Detect Intrusion Combining Federated Learning and SHAP. Future Internet March 2025, 17, 234. DOI:10.20944/preprints202503.1902.v1
13. Fatema, K.; Dey, S. K.; Anannya, M.; Khan, R. T.; Rashid, M.; Chunhua, S.; Mazumder, R. Federated XAI IDS: An Explainable and Safeguarding privacy Approach to Detect Intrusion Combining Federated Learning and SHAP. Preprints March 2025, 2025031902. https://doi.org/10.20944/preprints202503.1902.v1.
14. Arreche, Osvaldo & Guntur, Tanish & Roberts, Jack & Abdallah, Mustafa. (2024). E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3365140.
15. Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," in IEEE Access, vol. 10, pp. 93104-93139, 2022, doi: 10.1109/ACCESS.2022.3204051.
16. T. Engelen, V. Rimmer, C. D. C. Garcia, and W. Joosen, "Troubleshooting CICIDS2017," in Proceedings of the 4th International Workshop on Traffic Measurements for Cybersecurity, 2021. [Online]. Available: https://intrusion-detection.distrinet-research.be/WTMC2021/Resources/wtmc2021_Engelen_Troubleshooting.pdf
17. Anthony, Kwubeghari & Ezeji, Nwamaka. (August 2025). Designing an Explainable Intrusion Detection System (X-Ids) Using Machine Learning: A Framework for Transparency and Trust. ABUAD Journal of Engineering Research and Development (AJERD). 8. 10.53982/ajerd.2025.0802.32-j.
18. Pratinav Seth & Vinay Kumar Sankarapu (Feb 2025). Bridging the Gap in XAI—The Need for Reliable Metrics in Explainability and Compliance. Available: https://arxiv.org/html/2502.04695v1#bib
19. P. R. B. Houssel, S. Layeghy, P. Singh, and M. Portmann, "eX-NIDS: A Framework for Explainable Network Intrusion Detection Leveraging Large Language Models," arXiv:2507.16241, 2025. [Online]. Available: https://arxiv.org/pdf/2507.1624