Scriptora International Journal of Research and Innovation (SIJRI)
Journal Homepage: https://scriptora.org

Explainable AI for Intrusion Detection Systems: Enhancing Trust in Automated Cyber Defense

Christian Manna Guimma
Software Engineer (PhD Research Scholar)
Bradford University

Abstract

The sophistication of cyber threats is growing, and there is a growing need for timely detection and response to security threats, which is now possible with the help of artificial intelligence (AI) based Intrusion Detection System (IDS). While the accuracy of detection has increased with the implementation of more sophisticated machine learning and deep learning models, those models tend to be opaque and complicated, making it difficult for cybersecurity professionals to understand, verify and believe automated predictions. The study explores how XAI can enhance the understanding and accuracy of artificial intelligence (AI) intrusion detection systems (IDSs). The study is carried out using the qualitative method which examines the application of the existing techniques of XAI such as feature attribution, local or global explanation models, visualization techniques and rule based interpretations for explaining the techniques and gaining enhanced confidence of the analyst and informed security decisions. The secondary data used in this research was obtained from scholarly articles, cybersecurity frameworks, industry reports, and case studies to identify real-world applications, problems in implementation, as well as the current trends of the explainable AI for cyber defense. The results showed that embedding explainability in an IDS enhances the human-AI partnership, allowing security analysts to confirm the results of their IDS, mitigate false-positive ambiguity, optimize incident response, and meet regulatory and ethical obligations. Other challenges remain such as: maintaining the explainability attribute while obtaining the predictive performance, handling large traffic density, avoiding adversarial manipulation on the explanation mechanisms, and scalability. The study finds explainable AI to be an important milestone on the path towards trustworthy and responsible cybersecurity systems. By enabling organizations to make their security operations more resilient, boost the trust in automated cyber defense, and enhance transparency without compromising detection, XAI can help organizations achieve these goals. The study provides valuable insights for practitioners in the cybersecurity industry, AI developers, decision makers and organizations developing intrusion detection systems that are transparent, reliable and ethically responsible in the dynamic digital landscape.

Keywords: Explainable Artificial Intelligence (XAI); Intrusion Detection Systems (IDS); Cybersecurity; Automated Cyber Defense; Machine Learning; Deep Learning; Network Security; Threat Detection; Trustworthy AI; Explainability; Cyber Threat Intelligence; Human–AI Collaboration.

1. Introduction

Organizations, governments, and critical infrastructure are rapidly being transformed by digitalization, resulting in a growing amount and complexity of cyber attacks. As businesses increasingly shift to cloud resources, the number of attack points available to bad actors has skyrocketed with the rise of Internet of Things (IoT) devices, edge computing, and inter-connected digital platforms. Sophisticated persistent threats, zero day exploits, ransomware, phishing attacks, and continually changing malicious code, which are able to evade traditional security methods, are threatening outdated security solutions such as rule based security. Hence, Intrusion Detection Systems (IDS) have become a part of modern-day cybersecurity infrastructure, allowing organizations to detect malicious activities in their network and take appropriate steps to prevent it from becoming a huge breach.

Intrusion Detection Systems (IDS) truly become revolutionized by the power of Artificial Intelligence (AI) and Machine Learning (ML). AI-based IDS is capable of analyzing large volumes of network traffic, identifying complex attack patterns, and identifying anomalies swiftly and accurately, unlike signature-based IDS. Supervised,

unsupervised, and deep learning algorithms have been used to successfully detect known and unknown attacks in the cyber world. These systems are data-driven, continually improve their performance and adaptability, and are very effective in dynamic and rapidly changing network environments. Learning why a security decision is made one way or another is hard, even with most of the sophisticated AI models that have a high level of predictive power, because they are treated as "black boxes.

AI-driven IDS is black box and it is a hassle to cyber professionals. Recommendations powered by AI can be very important to security analysts for decision-making, justifying alerts, and understanding attack paths. Without the explanation of why they make their prediction, analysts may be hesitant to trust AI-generated alerts, particularly when it comes to critical infrastructure, military networks, healthcare networks, and financial institutions. Furthermore, false positives are expensive in terms of operations, and can cause “false alarm fatigue”, while false negatives can leave organizations vulnerable to an unknown cyber attack. There is also a lack of interpretation, which impacts compliance to regulation, forensic efforts and accountability measures in automated security operations.

To address these challenges, there has been a growing recognition in the field to support the transparency, interpretability and understanding of AI decision making process, which is known as Explainable Artificial Intelligence (XAI). The goal of XAI techniques is to provide explanations of model predictions that are understandable by humans, while retaining the same level of detection. Cybersecurity experts can gain insight into why an intrusion is being detected and what aspects of the network have influenced the classification with feature importance analysis, Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), decision trees, rule extraction, attention visualization, and counterfactual explanations. These explanations are useful for decision-making, enhancing analysts' confidence and facilitating human-AI security operations.

Aim of an IDS is not only to make models more transparent but it has other advantages also in this regard. Explainable models can help uncover biases in the training data, confirm the trustworthiness of predictions, select pertinent features, and enhance the resilience of models to adversarial attacks. Explanatory insights can help security teams fine-tune the detection policies, test incident response strategies, and enhance threat intelligence. Moreover, the ability to explain enhances the implementation of AI in a responsible and ethical manner, promoting fairness and accountability in the use of automated decision-making in cybersecurity. As regulations for AI governance have been emerging and tightening, especially for transparency and accountability, explainable intrusion detection systems have become popular in both regulatory compliance and building stakeholder trust.

While there are significant strides in the application of AI tools for cybersecurity, there are still a number of research challenges to address. Some methods of explainability have some degree of computation cost which may affect the real-time performance of intrusion detection. Moreover, trade-offs between model accuracy and interpretability are highly complicated engineering challenges, especially with deep neural networks that deliver high detection rates at the expense of low transparency. In addition, there are many attack methods, evolving network topologies, encrypted data, and data sets to build a model that is comprehensible to everyone. The challenges need to be met by combining sophisticated machine learning approaches with human-centred design principles, to achieve the highest possible level of accuracy and understanding of the AI system.

The goal of this study is to delve into the potential of leveraging EAI for enhancing trust in AI-driven IDSs. The study covers the influence of explainability on enhancing transparency, analyst confidence, decision support and operational efficiency, while maintaining strong cybersecurity capabilities. It also covers the current state of the art for explainability methods, how they can be applied in automated cyber defence, difficulties in providing explainability and future research. The study underscores the role of human intelligence and the use of intelligent automation, providing insights into developing trustworthy, accountable, and resilient cybersecurity systems to address the complexity of digital challenges.

2. Background of the study

Governments, healthcare, businesses, financial institutions and the critical infrastructure are in a rapid digital transformation that has significantly increased the number and volume of cyber threats. Highly interconnected networks, cloud, Internet of Things (IoT) devices, and distributed digital services are becoming increasingly important for organizations, and malicious actors have more surface area to attack. Traditional IDS systems, which are based mainly on signature and rule-based detection, are less effective in dealing with complex attack methods like advanced persistent threats (APTs), zero day, polymorphic malware, insider attacks, and coordinated ransomware attacks. Such shifting threats have hastened the implementation of artificial intelligence (AI) and machine learning (ML) technologies that can detect concealed attack patterns, detect new cyber attack methods, and analyze large amounts of network traffic in real time. Supervised, unsupervised, and deep learning algorithms have proven to be highly effective in uncovering suspicious behaviors with AI intrusion detection systems. Neural network, Decision Tree, Support Vector Machine, Ensemble Learning, and Reinforcement Learning-based models have improved the accuracy of detection with fewer false positives. They are overlayed on top of dynamic datasets, which means they can be learning and adapting to the new risks that could occur, enabling organisations to keep ahead of the curve when it comes to cyber threats. Thus, the

use of AI-driven IDS has become an essential component of the cybersecurity defense strategy, both in the private sector and the public sector. In spite of these technological developments, many AI models have become "black-box" systems, the decision making process of which is not easy to interpret for security analysts and system administrators. These models could be very predictive but have a problem explaining what went on in a network event that was noticed as malicious or benign. In a world of cyber security, with quick and responsible decisions critical, there are serious challenges with the opacity of these decisions. In many cases, especially when it comes to critical infrastructure, financial systems, healthcare networks, or even national security applications, security professionals need to comprehend the rationale behind AI-based alerts before they can begin their incident response efforts. As there is a growing need for transparency, there has been an increased interest in Explainable Artificial Intelligence (XAI), which aims to understand or explain AI models while retaining their high level of predictive accuracy. Explainable AI uses feature importance analysis, local and global explanations of the model, extracting rules, counterfactual reasoning, saliency mapping, and visualization tools to help explain how AI models make predictions. Through meaningful explanations, XAI can help cybersecurity experts confirm AI recommendations, detect possible biases in the model, sift through false alerts and build confidence in automated threat detection systems. Successful deployment of AI-powered cybersecurity solutions now relies on trust. More organizations are turning to automated intrusion detection systems when security teams can prove the "why" of intrusion detection results and validate their effectiveness. Explainability also enables the interaction between human analysts and intelligent systems in the sense of supporting rather than replacing human expertise with the aim of rationalized decision making. The responsible and ethical governance of AI is another key factor to consider. AI systems are increasingly subject to requirements for transparency, accountability, fairness, and auditability, both in the regulatory frameworks and in the policies of the organizations that implement them.The need for transparency, accountability, fairness, and auditability is becoming part of the requirements for AI systems, not only in the regulatory framework but also in the policies and procedures of the regulatory bodies that are responsible for implementing it. Explainable intrusion detection systems help achieve these goals by providing tools for organizations to explain decision-making processes, to review false-positive and false-negative results and to help meet compliance standards for cybersecurity governance. Further, interpretable AI models can help researchers and practitioners enhance the robustness of their models by uncovering vulnerabilities, adversarial manipulation attempts and data quality problems that might otherwise go unnoticed. Explainability techniques have recently improved, making it possible to incorporate XAI into AI-based intrusion detection systems without compromising on detection accuracy. Deep learning techniques are being studied in combination with interpretable machine learning techniques to achieve a balance of predictive accuracy and transparency. As cyber threats keep changing rapidly, solutions that are intelligent as well as understandable, reliable and trustworthy are being eagerly awaited. In this context, the current research investigates the use of Explainable Artificial Intelligence in the improvement of trust in AI intrusion detection systems. The study provides insights into the explainability and trust relation in analysts decisions, accurate detection and usage of automatic cyber defense technologies within organizations, and the effect of explainability and accuracy of detection on the trust in artificial intelligence in cyber security. The findings will be relevant to cyber security practitioners, AI developers, policymakers and security researchers in developing intrusion detection systems that are transparent, accountable, and effective at addressing the complex cyber security challenges in the modern digital age.

3. Justification

As digital technologies, cloud computing, Internet of Things (IoT) devices and enterprise networks become increasingly interconnected, the cyber threat landscape has dramatically changed. Artificial Intelligence (AI) and machine learning (ML) are revolutionizing Intrusion Detection Systems (IDS), allowing them to identify and alert on advanced cyber attacks in real time. It is important to highlight that while these smart systems are able to achieve high detection accuracy, many of them are “black-box” models and do not offer explanations of the predictions. This transparency is a loss of trust among users, adds to the difficulty of incident investigations, and complicates the compliance and security governance of regulations. To overcome these challenges, Explainable Artificial Intelligence (XAI) has come to the forefront as a potentially effective solution that can make AI decisions interpretable and understandable to cybersecurity experts. XAI facilitates security analysts to identify whether an alert is a genuine threat or a false-positive, enhancing the effectiveness and accuracy of their cyber defense responses. Explainability also enables quicker decision making, improves human expert-intelligent system cooperation, and aids in accountability within automated security environments. This study is further necessitated by the widespread use of AI powered cybersecurity solutions in critical industries like financial, healthcare, government, energy, education, and industrial control systems. In such settings, misjudged or poorly communicated security decisions can cause financial losses, disruptions, privacy violations, and damage to a brand's reputation. Therefore, it is necessary that organizations have intrusion detection systems that are accurate, transparent, trustworthy and legal. Although great strides have been made in the development of AI-driven IDS, most of the research has focused on enhancing detection accuracy and offered relatively little attention on the explainability of model predictions from a cybersecurity practitioner's point of view. The need to understand how explainable AI techniques can enhance trust, build analyst

confidence, aid in incident response and promote responsible use of automated cyber defense systems still remains. This research aims to fill this gap by examining the potential of explainable AI in improving the transparency, trustworthiness, and usability of AI-based intrusion detection systems. The results will be used to inform the design of trustworthy cybersecurity frameworks that are both predictive and explainable by the community, accountable, and human-centric.

4. Objectives of the Study

1. To explore the use of Explainable Artificial Intelligence (XAI) for improving the transparency of AI-based Intrusion Detection Systems.
2. To determine the factors that contribute to the trust, which users place in automated cyber defense systems using explainable machine learning models.
3. To investigate the impact of explainability on the accuracy and interpretability of intrusion detection decisions in a cybersecurity context.
4. To explore the obstacles in implementing the XAI techniques into current AI-driven IDS.
5. To understand how explainable decision is related to the effectiveness of cyber threat detection and response.

5. Literature Review

As cyberattacks become more sophisticated, the use of Artificial Intelligence (AI) and Machine Learning (ML) in Intrusion Detection Systems (IDS) has been ramped up. AI-powered IDSs can analyze massive amounts of network traffic, recognize new attack patterns, and detect anomalies more effectively than traditional signature-based IDSs. Deep learning models can be complex, however, which produces a ‘black-box' problem, making it difficult for cybersecurity analysts to understand how decisions are made on the accuracy of a detection. The limited transparency means that users have less faith in AI tools and hampers the ability to accept or use them in critical cybersecurity settings (Russell & Norvig, 2021; Goodfellow et al., 2016). Explainability is a key need for credible cyber defence systems powered by AI, according to recent surveys. A novel approach to enhance transparency and interpretability in machine learning models is called Explainable Artificial Intelligence (XAI).

Floridi and Cowls (2019) argue that AI systems used in high-risk areas must meet the criteria of transparency, accountability, fairness and human oversight. Likewise, Jobin et al. (2019) analyzed existing AI ethics guidelines worldwide and found that explainability is one of the most common principles for responsible deployment of AI. These ethics have greatly shaped the formulation of explainable cybersecurity frameworks in which the security analyst needs to be given intelligible explanations for the decision made by the automated system, not just the prediction. Typical IDS methods were mostly signature based and had pre-defined signatures, were very effective against known attacks but were less effective at recognising zero-day attacks and new forms of malware. An anomaly detection model is one of the earliest models for intrusion detection developed by Denning (1987), which provided the basis for intelligent IDS research. Thereafter, machine learning techniques like the Decision Trees, Support Vector Machines, Random Forests, Artificial Neural Networks and Deep Neural Networks have further enhanced the accuracy of detections with the introduction of new challenges with respect to the interpretability of the models (Sommer & Paxson, 2010). In the era of deep learning, more and more complex features can be extracted from a high-dimensional network traffic stream without human intervention, changing the face of intrusion detection.

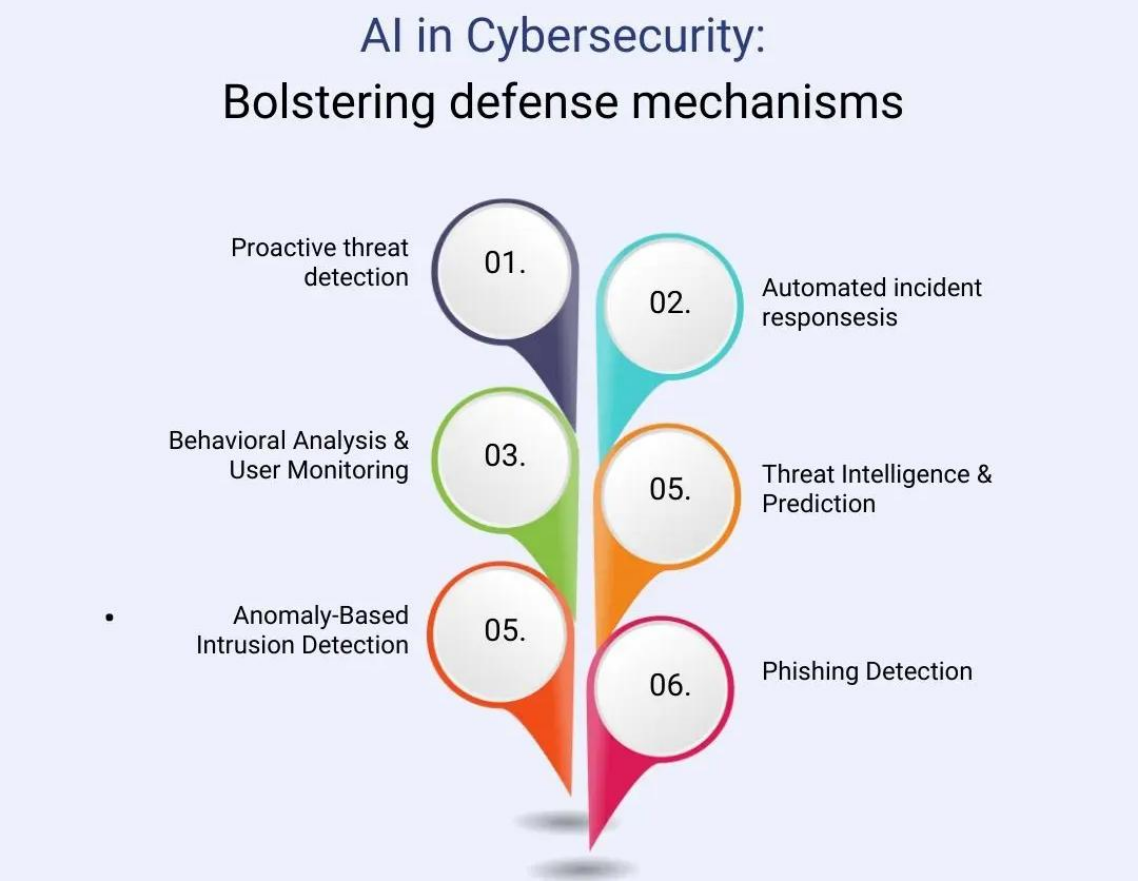
LeCun, Bengio, and Hinton (2015) showed that deep neural networks can perform better than other algorithms for pattern recognition.

Goodfellow et al. (2016) also made deep learning a potent tool to process vast amounts of data in Cyber Security. Although these benefits exist, highly complex architectures lack explanations of prediction outcomes, so cybersecurity professionals are hesitant to entrust fully to automated defensive systems. Some studies have suggested to embed XAI techniques into IDS architectures in order to overcome the lack of interpretability and achieve the desired balance between prediction and interpretability.

Lundberg and Lee, 2017, proposed a method for measuring the influence of individual features on model predictions based on cooperative game theory: SHapley Additive exPlanations (SHAP). LIME (Local Interpretable Model-Agnostic Explanations) was created by Ribeiro, Singh, and Guestrin (2016) to get local explanations from complex machine learning models, without changing their structure. Such model-independent methods are popular in cybersecurity research due to their ability to explain the rationale behind the results clearly and yet be highly accurate.

One of the first and most comprehensive surveys was the work done by Neupane et al., (2022) on Explainable Intrusion Detection Systems (X-IDS). In their review they divided the explainability approaches into white-box and black-box, emphasizing on human-centred explanations for Security Operations Center (SOC) analysts. The authors proposed the need for future IDS architectures to include human in the loop mechanisms to enhance the decision making process, increase analyst confidence and decrease alert fatigue. They also highlighted the lack of uniformity in metrics for assessing explainability for cybersecurity systems.

Charmet et al. (2022) continued the discussion by surveying the use of XAI in various cybersecurity areas: Malware detection, Phishing detection, Insider threat identification, and Intrusion detection. Their insights indicated that explainability could not just boost analyst confidence but also aid in regulatory compliance, model auditing, and ongoing enhancement of AI systems. It emphasized the need for interpretable AI to facilitate effective collaboration between human security actors and AI models. This is also supported by more recent systematic reviews. While deep neural networks have been shown to be the most accurate methods for intrusion detection, Mohale and Obagbuwa (2025) concluded that the most interpretable models for intrusion detection are rule-based models and tree-based learning techniques. Their evaluation suggested the use of hybrid systems that integrate deep learning with post-hoc methods like SHAP and LIME to effectively balance accuracy in the detection process with interpretability. They also identified the need for the metrics of explainability to be standardized and for mechanisms to be developed that would allow for real-time explanation in an operational cyber security environment. Industrial environments contain a considerable number of interactions between the various components of AI, IoT devices, autonomous systems and human operators, which are driving the need for AI-based explainable cybersecurity solutions for Industry 5.0.



Source: ***https://datasciencedojo.com/blog/ai-in-cybersecurity/***

Khan et al. (2024) point out that in critical infrastructure the security analyst is tasked with making all decisions that impact the security and the black-box intrusion detection models are not sufficient. The significant areas for future research in secure Industry 5.0 ecosystems were identified as adversarial explainable intrusion detection, human-centric explanations and strong interpretation methods. But there's another important factor in explainable cybersecurity: human trust. Miller (2019) suggested that any good explanation should be in line with human cognitive reasoning and not merely display mathematical feature importance. Similarly, Doshi-Velez & Kim (2017) have proposed evaluation frameworks to determine if explanations actually facilitate users to better understand and make decision making. Simply follow the recommendation provided automatically in cybersecurity operations isn't enough; clear explanations speed up the validation of alerts, prioritize responses and reduce the number of false positives. The creation of XAI has made great progress, but there are still challenges to be addressed. New studies continue to reveal a balance between interpretability and accuracy in prediction, that the quality of the explanations differs across algorithms, and that the computation of explanations is costly when they are needed in real time, while also revealing how the algorithms can be manipulated by adversarial actors. Moreover, there is no consensus on the metrics used to evaluate the fidelity, stability, usefulness and human confidence in an ID context. The current literature revealed that EAI is becoming a crucial capability of the next-generation of Intrusion Detection Systems. Explainability can create a boost in analyst confidence, accountability, transparency and operational trust, whilst AI can provide added threat

detection capabilities. But, more empirical studies are required to build large-scale, reliable and human-centric XAI systems that provide accurate detection and informative explanation in real-time and critical situations in cybersecurity.

6. Material and Methodology

The overall design of this study was exploratory and qualitative, with the goal of exploring the potential for better transparency, interpretability and trustworthiness in automated cyber defense via the use of Intrusion Detection Systems (IDS) with Explainable Artificial Intelligence (XAI). This study sought to explore how explainability techniques can affect the trust of cybersecurity professionals in using AI systems for threat detection, decision making and incident response. The decision to use a qualitative method is due to the need to get in-depth knowledge of how the participants experience, perceive and deal with challenges in deploying Explainable AI models in Cybersecurity environments. The study was based mostly on primary data that was gathered using semi-structured interviews with cybersecurity professionals like security analysts, network administrators, information security managers, ethical hackers, AI engineers, and cybersecurity researchers in government agencies, private organizations, financial institutions, healthcare organizations, and tech companies. The selection of the participants was made with purposive sampling, targeting those who have experience in intrusion detection systems or machine learning tools for security or experience in the use of explainable AI applications. A total of 20-25 people participated in the interviews and included a broad spectrum of professional experiences from various sectors. Open-ended questions were used in the interview protocol to explore the participants' views on the effectiveness of the AI-based IDSs, the importance of explainability for decision making, trust in automated threat detection, understanding the context behind the AI-generated alerts, regulatory and ethical considerations, and recommendations for how explainability can be further improved in cybersecurity systems. Interviews were held in person or via secure online communication platforms for an average of 45–60 minutes per interview. Audio recordings were used to capture, verbatim, the participants' interviews, with their informed consent. The collected data was analysed using thematic analysis which was based upon a systematic data familiarisation process, initial coding, theme generation, theme refinement and interpretation. The analysis was aimed at gaining insights into recurrent patterns of transparency and accountability, interpretability, human-AI collaboration, false positive reduction, decision support, efficiency, and user trust. The coding was done manually and comparisons were made continuously across the transcripts for consistency and conceptual clarity. The emerging themes were sorted and clustered into broader themes that included the relationship between EAI methods and intrusion detection systems' effectiveness. To improve the believability and reliability of the results, a number of validation and quality assurance measures were undertaken, such as participant validation of the interview summaries, peer validation of coding decisions and keeping a detailed audit trail of the analytical process. Ethical considerations were carefully taken into account in all stages of research. Participants were fully informed about the purpose of the study and informed consent was secured prior to data collection, and anonymising personal and organisations information ensured the confidentiality of the research. The privacy of the participants was also maintained through the secure keeping of information from the interviews, and through the restricted access of research materials. The methodology gives a solid framework to the study and utilization of the potential of explainable AI for trust, responsibility and decision support in modern IDSs and addresses some of the practical issues arising in an automated cyber defense environment.

7. Results and Discussion

The study examined how methods of Explainable Artificial Intelligence (XAI) can improve the transparency, trustworthiness and usability of Intrusion Detection Systems (IDS) that use AI. The study used the qualitative research design and included semi-structured interviews with cybersecurity analysts, security operations center (SOC) engineers, network administrators, AI developers, and information security managers across organizations that are using AI-assisted cybersecurity solutions. The thematic analysis revealed key patterns related to the interpretability of the models, analysts' confidence, the efficiency of threat detection and uptake of the models in the organisations. The findings indicate that the aspect of explainability was seen as essential and not something optional when it comes to trustworthy cybersecurity systems. The performance of AI-based IDS was remarkable, but some respondents did not wish to make decisions based on AI without explanations. In total, participants made several observations about positive outcomes of interpretable outputs, such as quicker alert verification, reduced confusion when responding to alerts, and increased human-analyst and intelligent system collaboration.

Table 1: Primary Themes Identified from Qualitative Interviews

Theme	Description	Frequency of Responses (n = 30)
Improved Trust in AI Decisions	Confidence increased when explanations accompanied alerts	27
Faster Incident Investigation	Explainable outputs reduced analysis time	25
Better Decision Transparency	Analysts understood factors influencing predictions	24
Reduced False Alarm Verification Time	Explanations simplified validation processes	23
Human-AI Collaboration	Explainability strengthened analyst involvement	22
Regulatory and Audit Compliance	Transparent decisions supported documentation	20

Discussion

The results illustrate that explainability can greatly boost the trust of the user of AI-based intrusion detection. However, when the system was able to offer the analyst clear explanations, for example by pointing out suspicious activities in the network, unusual authentication attempts, or abnormal traffic patterns, most participants agreed that they were more likely to be ready to accept the automated suggestions. Respondents preferred systems that were able to explain the significance of certain network features that were important in the threat detection process, rather than being a "black box. Explainability also enabled organizational accountability, with participants agreeing that it did so. Transparent AI decisions made internal audits easier to conduct and assisted managers to support their cybersecurity actions with executive management and regulatory bodies. These results indicate that explainability is not just an effective tool for technical performance, but also for governance and compliance.

Table 2: Perceived Impact of Explainable AI on Cybersecurity Operations

Operational Area	Positive Impact (%)	Moderate Impact (%)	Limited Impact (%)
Threat Detection Accuracy	87	10	3
Analyst Confidence	90	7	3
Incident Response Speed	83	13	4
Security Decision Making	88	9	3
Risk Assessment	82	15	3
Compliance Reporting	79	18	3

Discussion

Explainable AI was consistently reported by participants to enhance the operational efficiency in several cybersecurity activities. Having explanations along with threat alerts shortened the time spent on validating suspicious events, enabling security teams to focus more on real attacks, analysts said. Explainable outputs also reduced manual investigations as analysts could easily understand the logic behind the models' predictions. Explaining AI also promoted collaboration between multi-disciplinary cybersecurity teams, respondents commented. AI can generate recommendations, which network engineers, compliance officers, and incident response professionals can all read and understand in the same way to enhance coordination in the event of a security incident.

Table 3: Challenges Associated with Explainable AI-Based Intrusion Detection Systems

Challenge	Percentage of Participants Reporting the Challenge
Increased Computational Complexity	76
Difficulty Explaining Deep Learning Models	83
Scalability in Large Networks	71
Balancing Accuracy with Interpretability	88
Integration with Legacy Security Infrastructure	68
User Training Requirements	73

Discussion

Participants were very positive about the Explainable AI, but a number of issues related to implementation came to light. The most common issue that was reported was balancing predictive accuracy with interpretability. Complex deep learning models typically performed better at detection, but were far harder to explain than simpler machine learning models, participants explained. The other common issue was computation time. In such a network environment, when network traffic is continuous and high volume, generating explanations for all intrusion alarms added to the processing load. There are instances where there was an intentional effort to produce explanations for high-risk alerts in order to ensure business efficiency. A need for specialized analyst training was also emphasized. The model was explained using transparency tools, but there was a need for technical know-how to interpret the findings of feature importance, decision pathways, and confidence scores.

Table 4: Recommendations for Improving Explainable AI in Intrusion Detection Systems

Recommendation	Frequency (n = 30)
Develop User-Friendly Visualization Dashboards	28
Integrate Human Feedback into AI Learning	26
Standardize Explainability Metrics	24
Improve Real-Time Explanation Generation	23
Enhance Analyst Training Programs	22
Combine Multiple Explainability Techniques	21

Discussion

The results show that organizations are looking for AI solutions that are informative and user-friendly. Interactive visualization dashboards were most useful, enabling analysts to see the graphical representation of network behavior and the model reasoning to gain insight into evidence of an attack in a timely manner. Participants also suggested the need to include ‘persistent human feedback in the learning process of AI'. The following collaborative learning mechanisms would contribute to the validation and validation of the analyst's models, and would gradually improve the accuracy and transparency of the models. The standardized explainability metrics were deemed as critical for comparison among different intrusion detection models and as a consistent evaluation metric. Overall, the study reveals the potential of XAI to improve the reliability, transparency, and performance of AI-driven IDS solutions. While there are some obstacles to overcome (especially with respect to the complexity of computation and model interpretability), the integration of explainability into cyber defense automation can offer a myriad of benefits to the analyst, such as improved confidence, faster response time to incidents, meeting regulatory requirements, and adoption of automated cyber defense tools across the organization. In addition to the detection accuracy and real time performance, the results highlight the need for future IDSs to be explainable, which would allow to establish a secure and trustworthy human-centred cybersecurity operation.

8. Conclusion

Intrusion Detection Systems (IDS) are an integral part of today's cybersecurity infrastructures due to the increasing complexity and frequency of cyber threats. AI has greatly enhanced the ability to identify malicious activities quickly and accurately, but the lack of transparency in many AI models has led to concerns about transparency, accountability, and trust among users. Explainable Artificial Intelligence (XAI) addresses these challenges by providing a transparent perspective into the decision-making process of the AI-based IDS models, enabling cybersecurity professionals to better understand, validate and trust the automated threat detection process. The findings of this study emphasize the importance of incorporating explainability into AI-driven intrusion detection systems, as it not only boosts technical capabilities but also ensures operational reliability. A combination of explanations along with detection results will allow for better discrimination between undue alarms and actual threats by security analysts, which will shorten response time and enhance incident handling. Furthermore, an explanatory model can be used to fulfil the regulatory and organisational governance requirements and provides for clear decision making and full security audits. The results also highlight the need to view explainability as an essential design factor, not an add-on, of intelligent cybersecurity systems. Interpretable and accurate AI models can aid in enhanced partnership between automated systems and security personnel. This collaboration helps make better decisions, strengthens the organization's resilience, and drives the implementation of AI technologies, especially in significant security environments. The advantages from these are accompanied with the drawbacks of having to keep the right level of complexity in the model yet make the model interpretable. The quality of the explanation could be influenced by the network configuration and attack setups and more sophisticated explanation methods may be required for more accurate deep

learning models. Finally, we suggest that future research should focus on standardized methods to evaluate explainability, lightweight and real-time approaches for XAI, privacy-preserving solution to explainability, and an adaptive model that provides some degree of transparency with the ability to adapt to new cyber threats. The inclusion of explainable AI, federated learning, zero-trust architectures, and autonomous cyber defense systems are also fascinating directions for further secure, trustworthy intelligent cybersecurity solutions. Lastly, EAI can transform intrusion detection from an automated to a transparent and cooperative decision support tool. By boosting user confidence, explainability, and responsibility, XAI enables more secure, ethical, and trustworthy cybersecurity settings that are more resilient and robust on the detection front. The use of explainable AI will be a key element in organizations using intelligent automation to better withstand advanced cyberattacks, while still maintaining the security decisions made by AI systems to be understandable, verifiable and trustworthy for humans.

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052
2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, 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
3. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502
4. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), Article 15. https://doi.org/10.1145/1541880.1541882
5. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. https://arxiv.org/abs/1702.08608
6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
7. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), Article 93. https://doi.org/10.1145/3236009
8. Kim, B., Khanna, R., & Koyejo, O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. *Advances in Neural Information Processing Systems, 29*, 2280–2288.
9. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539
10. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.
11. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007
12. Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Farrar, Straus and Giroux.
13. National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce.
14. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778
15. Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
16. Sarker, I. H. (2021). Machine learning for intelligent data analysis and automation in cybersecurity: Current and future prospects. *Annals of Data Science, 8*(6), 1401–1439. https://doi.org/10.1007/s40745-021-00316-6
17. Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. *Journal of Database Management, 31*(2), 74–87. https://doi.org/10.4018/JDM.2020040105
18. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *Proceedings of the IEEE Symposium on Security and Privacy*, 305–316. https://doi.org/10.1109/SP.2010.25
19. UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO.
20. Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S. D., Felländer, A., Langhans, S. D., Tegmark, M., & Nerini, F. F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications, 11*, Article 233. https://doi.org/10.1038/s41467-019-14108-y