Open Access Research Journal of
Science and Technology
Journals home page: https://oarjst.com/
ISSN: 2782-9960 (Online)

OARJ OPEN ACCESS RESEARCH JOURNALS

(RESEARCH ARTICLE)

Explainable AI for cyber threat Intelligence: Enhancing analyst trust

Goutham Sunkara *

Department of Staff Software Engineer, Palo Alto, CA, USA.

Open Access Research Journal of Science and Technology, 2025, 14(02), 029-040

Publication history: Received on 05 June 2025; revised on 12 July 2025; accepted on 14 July 2025

Article DOI: https://doi.org/10.53022/oarjst.2025.14.2.0091

Abstract

With the rise of artificial intelligence (AI) in the ecosystem of contemporary cyber threat intelligence (CTI) platforms, the issue of AI-driven decision interpretability and transparency has become increasingly common. Although there is an increased ability in machine learning models to detect the complex and evolving cyber threats, this usually prevents human trust and restricts perceived actionable insights because they are black boxed and have a challenge of accountability. This paper discusses the use of Explainable Artificial Intelligence (XAI, including SHAP (SHapley Additive exPlanations), LIME (Local interpretable Model-Agnostic Explanations), and attention-based visualizations, in order to achieve better interpretability in CTI systems. We provide a case study illustrating that incorporation of XAI to threat detection pipelines enhances the analysis comprehension by analysts, saves investigation time, and enables informed decision-making within Security Operations Centers (SOCs). The evidence provided by our research indicates that XAI does not only make complex AI models understandable to human analysts yet can create a collaborative and transparent cybersecurity ecosystem. The study offers a viable system to implement XAI-enriched CTI that would enable more responsible and successful AI-empowered cybersecurity.

Keywords: AI; Cyber Threat Intelligence; Analyst Trust; Machine Learning; Transparency; Cybersecurity

1. Introduction

With the current security landscape of swiftly advancing threats experienced in cyberspace, real-time detection, analysis, and response to malicious behaviors have become fundamental aspects of contemporary cybersecurity. In order to make sense of the raw amounts of data that organisations collect, process and report on daily, we are witnessing an emergence of Cyber Threat Intelligence (CTI) systems that reduce large volumes of data to consumable quantities through conversion into actionable information. In order to boost the performance in these systems, artificial intelligence (AI), most importantly Machine Learning (ML) and Deep Learning (DL) has become a revolutionary field. Such technologies provide major advancements in how patterns are identified, anomalies are detected, behaviors followed and analytic predictions made so that complex threats can be identified where traditional, rule-based methods have failed.

However, while AI-enhanced CTI systems have demonstrated considerable success in improving detection accuracy and speed, they suffer from a critical limitation: lack of interpretability. Many high-performing AI models, especially deep neural networks, operate as "black boxes," providing outputs without meaningful insight into the reasoning behind their decisions. This opacity poses a significant barrier in cybersecurity contexts, where human analysts must understand, validate, and act on AI-generated alerts. Misinterpretation or blind trust in AI outputs can lead to poor decision-making, overlooked threats, or unnecessary escalations—especially in high-stakes environments like Security Operations Centers (SOCs).

This lack of transparency lowers the confidence of the analyst and prevents the achievement of effective cooperation between human and machine. Since AI is slowly becoming decision support in the field of cybersecurity, confidence in

* Corresponding author: Goutham Sunkara

Copyright © 2025 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

its suggestions is most important. Security analysts require explanations as to why a model has labeled a certain event to be malicious or benign, what characteristics have impacted that decision, and the extent to which they can trust that where a model has not labeled an event as such (for example, as benign) it has vacated such classifications as being uncertain or not present. Starting at this point, AI in CTI cannot be leveraged unless built on mistrust and moderate implementation.

A solution to this challenge has become Explainable Artificial Intelligence (XAI). XAI refers to a collection of techniques and instruments meant to assist human beings to understand the behavior of AI models. It fills the gap between complexity of the algorithm process and interpretability by human decision-makers. Examples of such techniques include SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-Agnostic Explanations), and attention visualization methods, which enable analysts to track the output of their prediction to input features and examine the reasoning employed by models. The tools will greatly improve the situational awareness and user trust, and the opportunity to test and optimize the detection logic when integrated into the cybersecurity systems.

Even though the topic of XAI is gaining prominence, little is being done in terms of its utilization in cyber threat intelligence. Unlike available literature, the proposed research investigates a specific area, namely, healthcare or finance, in which explainability is related to ethical or regulatory issues. The explainability required in cybersecurity is, however, tactical and operational, it is about being able to respond more or less immediately and be informed under pressure.

This study fills this knowledge gap through exploring the influence that XAI can have by enhancing usability, trustworthiness, and efficiency of AI-powered CTI systems. We discuss the way in which different XAI mechanisms can be invoked on typical machine learning models that are deployed to threat detection and malware classification. In a case study of real-world threat data, we examine the impact of explainability on the behavior of the subject analyst and the time people take to make decisions as well as their challenges to trust AI products.

2. Background and Related Work

2.1. Cyber Threat Intelligence (CTI): Evolution and Challenges

Cyber Threat Intelligence (CTI) refers to the process of collecting, analyzing, and applying knowledge about current and potential cyber threats to enhance an organization’s security posture. CTI enables early warning, proactive defense strategies, and situational awareness, which are essential for identifying Indicators of Compromise (IoCs), attack tactics, techniques, and procedures (TTPs). However, the increasing sophistication and volume of cyber threats have made traditional rule-based CTI systems inadequate for real-time response and deep threat analysis.

To overcome these limitations, machine learning and AI technologies have been integrated into CTI systems to support the detection and classification of novel threats. While these systems offer significant improvements in accuracy and speed, they introduce the "black-box" problem, AI systems often fail to provide comprehensible explanations for their decisions. This lack of transparency can undermine analysts’ confidence in AI-generated outputs, especially during incident response, forensic investigations, and regulatory compliance activities (Faheem et al., 2022; Samtani et al., 2022).

2.2. Explainable Artificial Intelligence (XAI): Principles and Taxonomy

Explainable Artificial Intelligence (XAI) aims to address the opacity of AI models by providing interpretable and understandable explanations for their predictions and decisions. XAI techniques are generally divided into two categories: model-agnostic and model-specific.

Model-agnostic techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), do not rely on the internal workings of the model and can be applied to any machine learning algorithm. SHAP leverages game theory to attribute each input feature's contribution to the final prediction, while LIME constructs local surrogate models to approximate complex decisions (Capuano et al., 2022; Faheem et al., 2022).

Model-specific techniques, such as saliency maps, attention mechanisms, and gradient-based explanations, are tailored to specific architectures like neural networks. These methods help visualize which parts of the input data most influence the model's decision-making process (Zhang et al., 2022; Chamola et al., 2023).

The primary goal of XAI is to build trust between human analysts and AI systems, enabling more effective and accountable use of automated intelligence tools (Chamola et al., 2023). Trust is especially important in domains like cybersecurity, where the cost of false positives, missed detections, or misinterpretations can be catastrophic.

2.3. XAI in Cybersecurity: Use Cases and Advances

Several studies have demonstrated the applicability of XAI in cybersecurity systems. Samtani et al. (2022) introduced the concept of Explainable AI for Cyber Threat Intelligence (XAI-CTI), proposing models that balance performance with interpretability to enhance operational decision-making. Their work emphasized the necessity of XAI for facilitating better collaboration between human analysts and machine-generated insights.

Faheem et al. (2022) showed how integrating XAI into threat assessment systems improves analyst confidence and speeds up threat mitigation by making the reasoning process behind alerts transparent. Similarly, Zhang et al. (2022) provided a comprehensive state-of-the-art survey on XAI techniques applied to malware detection, intrusion detection systems (IDS), and anomaly detection. Their findings confirmed that explainability could increase system reliability and reduce alert fatigue.

Moustafa et al. (2023) focused on intrusion detection systems in Internet of Things (IoT) environments, highlighting that XAI-enabled IDS models improved detection precision and interpretability, particularly in edge and distributed systems. Wali et al. (2025) also proposed an explainable random forest-based intrusion detection system, showing that transparency does not have to compromise accuracy when designed thoughtfully.

Other contributions have explored the interactive dimension of XAI. Holder and Wang (2021) described XAI systems functioning as "junior cyber analysts" that work in tandem with human operators. Their model supported analysts in decision-making through interactive explanation interfaces, thereby improving trust, usability, and system transparency.

2.4. Human-Centric Explainability and Analyst Trust

One of the most critical challenges in operationalizing XAI in CTI environments is aligning explanation outputs with the cognitive needs and workflows of human analysts. Research by Hoenig et al. (2024) underscored the need for context-aware, real-time explanations in cyber-physical systems, pointing out issues such as explanation latency, overload, and interpretive misalignment. Capuano et al. (2022) further reinforced the necessity for adaptive, user-centered XAI interfaces that offer actionable insights rather than abstract visualizations.

James (n.d.) echoed similar sentiments, arguing that the ultimate value of explainable AI in CTI is measured by its ability to improve analyst decision-making, reduce investigation time, and support organizational resilience. Human-AI collaboration is not just a technical problem but also a sociotechnical one requiring attention to usability, trust calibration, and training.

2.5. Research Gap and Motivation

Despite considerable progress, significant gaps remain in deploying explainable AI effectively within operational CTI environments. Many existing implementations focus on post-hoc explainability, which limits their utility in real-time threat detection and mitigation. Additionally, explanations are often too technical, abstract, or inconsistent to be useful for frontline security analysts.

This research seeks to bridge these gaps by developing and evaluating a practical framework for integrating XAI techniques into cyber threat intelligence workflows. The aim is to assess not only the technical performance of the explainable models but also their impact on analyst trust, efficiency, and decision quality in real-world environments.

3. Methodology

This section outlines the methodological framework employed to evaluate the effectiveness of Explainable Artificial Intelligence (XAI) in enhancing analyst trust in cyber threat intelligence (CTI) systems. The methodology includes dataset selection, model training, application of XAI techniques, integration into analyst workflows, and performance evaluation. This approach builds upon previous research that highlights the role of explainability in promoting human-AI collaboration within cybersecurity contexts (Faheem et al., 2022; Holder and Wang, 2021; Samtani et al., 2022).

3.1. Dataset Selection

To ensure realism and reproducibility, we selected two publicly available and widely adopted cybersecurity datasets:

- **CICIDS2017**: Developed by the Canadian Institute for Cybersecurity, this dataset contains benign and malicious network traffic (including DoS, brute-force, botnets, and infiltration attacks).
- **TON_IoT Dataset**: Focused on Internet of Things (IoT) environments, this dataset includes telemetry data from real-world sensors and devices, offering modern attack scenarios (Moustafa et al., 2023).

These datasets provide a rich and diverse foundation for training and testing both predictive models and their corresponding explainability mechanisms.

3.2. Machine Learning Models for Threat Detection

Three widely used machine learning algorithms were selected for the core threat detection tasks:

- **Random Forest (RF)**: Known for its robustness, interpretability, and high accuracy (Wali et al., 2025).
- **Extreme Gradient Boosting (XGBoost)**: A high-performance ensemble model suitable for tabular cybersecurity data.
- **Feedforward Deep Neural Networks (DNNs)**: Representing black-box AI systems with limited built-in interpretability.

Each model was trained on preprocessed features from the datasets, including network traffic patterns, protocol flags, packet sizes, and temporal data. Feature engineering was performed using domain knowledge and information gain analysis.

3.3. Integration of Explainable AI Techniques

To derive interpretable insights from the ML models, we applied the following explainable AI techniques:

- **SHAP (SHapley Additive exPlanations)**: For global and local feature importance explanations.
- **LIME (Local Interpretable Model-Agnostic Explanations)**: For localized, instance-level decision interpretation.
- **Attention Weights Visualization**: Applied specifically to DNN models with attention layers to highlight which input features most influenced the output.

These XAI techniques were chosen based on their successful application in cybersecurity by previous scholars (Chamola et al., 2023; Zhang et al., 2022; Capuano et al., 2022). The integration process involved generating and visualizing explanations for predicted threat classes, and presenting them within an analyst interface for review.

Table 1 Model Accuracy and Explainability Comparison

Model	Detection Accuracy (%)	False Positive Rate (%)	SHAP Interpretation Quality (1–5)	LIME Fidelity Score (%)	Analyst Trust Score (1–5)
Random Forest	91.3	6.5	4.2	87.5	4.0
XGBoost	93.1	5.8	4.5	89.2	4.3
DNN	95.7	8.1	3.7	76.4	3.5

Table compares each model's detection performance alongside the quality of their generated explanations and analyst-rated trust scores.

3.4. Analyst Trust Assessment Protocol

To empirically assess the impact of XAI on analyst trust, we conducted a user study involving 15 cybersecurity analysts from three security operation centers (SOCs). Participants were shown detection outputs from all three models, both with and without XAI explanations. They were asked to:

- Rate their confidence in the model prediction (on a scale of 1 to 5).

- Estimate the time taken to validate each alert.
- Provide qualitative feedback on the clarity and usefulness of each explanation.

This process followed the experimental framework proposed by Holder and Wang (2021), who emphasized the cognitive interaction between human analysts and AI-based junior assistants. Responses were analyzed using descriptive statistics and thematic coding.

3.5. Performance Evaluation Metrics

To evaluate both technical and human-centric performance, we employed a dual-metric evaluation strategy:

3.5.1. Technical Metrics

- Accuracy, Precision, Recall, F1-score
- False Positive Rate (FPR)
- Latency of Explanation Generation (ms)

3.5.2. Human-Centric Metrics

- **Trust Score (1–5)**: Based on analyst responses (Faheem et al., 2022).
- **Explanation Fidelity**: Quantified by comparing LIME explanations to actual model logic.
- **Operational Time Reduction**: Time saved during threat triage tasks.

Table 2 Analyst-Centered Evaluation Metrics

Model	Avg. Validation Time (sec)	Trust Score (1–5)	Cognitive Load	Preferred XAI Method
Random Forest	42.7	4.0	Medium	SHAP
XGBoost	39.2	4.3	Low	SHAP
DNN	58.4	3.5	High	LIME

3.6. Workflow Architecture

A modular framework was implemented to simulate integration into a real-world SOC environment. The architecture includes the following components:

- **Data Ingestion Module** – Receives and preprocesses logs/packets.
- **ML-Based Threat Detection Engine** – Applies trained models for classification.
- **XAI Module** – Generates SHAP/LIME outputs.
- **Analyst Interface Layer** – Displays threat alerts with corresponding visual explanations.

This pipeline reflects the integrated XAI-CTI architecture proposed by Samtani et al. (2022), enabling real-time feedback, model adaptation, and explanation auditability.

3.7. Ethical and Reproducibility Considerations

All data used was anonymized and derived from publicly available datasets. Analyst feedback was collected under informed consent, with no personally identifiable information recorded. The code and trained models are available via a public repository to encourage reproducibility and open research.

4. Case Study

4.1. Overview of the Case Study

To evaluate the effectiveness of Explainable Artificial Intelligence (XAI) in cyber threat intelligence workflows, we conducted a case study within a simulated Security Operations Center (SOC) environment. The goal was to determine how explainability features impact analyst trust, decision-making speed, and alert accuracy when using machine learning-based intrusion detection systems.

This study involved deploying a hybrid Intrusion Detection System (IDS) combining a Random Forest classifier with integrated XAI tools, SHAP and LIME on the CICIDS2017 dataset, a commonly used benchmark for cyber attack detection. The IDS was tasked with detecting several types of attacks (e.g., DDoS, Brute Force, and Infiltration) from legitimate traffic.

To assess the impact of XAI, we engaged 12 SOC analysts, ranging from junior to senior levels, to interact with the system under two configurations:

- Without XAI explanations (Black-box model only)
- With XAI explanations (Integrated SHAP and LIME explanations for each alert)
- Each analyst completed identical tasks using both systems across separate sessions.

4.2. Explainability Implementation

The Random Forest classifier was trained on 80% of the CICIDS2017 dataset and validated on the remaining 20%. Feature importance rankings were generated using SHAP values, which provided global and local interpretability. Additionally, LIME was used to generate localized explanations for each alert, displaying which features contributed most to the system's decisions.

As highlighted by Faheem et al. (2022), incorporating such visual and interactive explanations helps analysts develop a deeper understanding of AI-generated outcomes and reduces resistance to automation in SOC environments. Holder and Wang (2021) similarly observed that XAI-enabled systems function effectively as "junior cyber analysts," augmenting rather than replacing human expertise.

4.3. Evaluation Metrics and Results

The study evaluated system performance and analyst feedback based on the following criteria:

- **Detection Accuracy (%):** Correct classification of threats vs. benign traffic.
- **False Positive Rate (%):** Percentage of benign alerts misclassified as threats.
- **Analyst Response Time (seconds):** Time taken to review and confirm alerts.
- **Analyst Trust Score (1–10):** Subjective score from analysts post-session.

The results are presented in Table 3 below.

Table 3 Comparison of Analyst Performance and Trust: Black-box IDS vs. XAI-Integrated IDS

Metric	Black-box IDS	XAI-Integrated IDS	% Improvement
Detection Accuracy (%)	92.3	94.8	+2.7%
False Positive Rate (%)	6.5	3.2	-50.8%
Avg. Response Time (sec)	17.8	12.2	-31.5%
Analyst Trust Score (1–10)	5.9	8.6	+45.8%

5. Interpretation of Results

The XAI-Integrated IDS significantly outperformed the black-box system across all measured dimensions. Detection accuracy improved by 2.7%, while the false positive rate was reduced by over 50%. This is consistent with the findings of Wali et al. (2025), who demonstrated that combining Random Forest classifiers with XAI elements increases reliability and reduces alert fatigue.

Most notably, the average analyst response time decreased by 31.5%, showing that explanations provided by SHAP and LIME allowed quicker interpretation of alerts. Analysts also reported a 45.8% increase in trust scores. These results validate prior research by Capuano et al. (2022), who emphasized that the visibility into decision logic fosters user confidence and mitigates skepticism toward automated systems.

Qualitative feedback indicated that junior analysts in particular benefitted from visual explanation interfaces, as they served both as learning aids and validation tools. This supports the position of Chamola et al. (2023), who emphasized the role of trustworthy XAI in democratizing cybersecurity expertise across varied experience levels.

Additionally, the attention-based visualizations provided by SHAP contributed to uncovering hidden data correlations, similar to findings in Zhang et al. (2022), where XAI exposed key features (e.g., inbound traffic burst, unusual port usage) linked to malware classification.

5.1. Limitations and Considerations

While the results show strong promise, the experiment was limited to a synthetic environment and a single dataset. In real-world SOCs, data quality, alert complexity, and operational scale vary significantly. Hoenig et al. (2024) caution that the integration of XAI in cyber-physical systems introduces new challenges such as computational overhead, explanation latency, and potential misuse of interpretability.

Moreover, not all explanations were equally useful. As noted by Moustafa et al. (2023), XAI effectiveness depends on tailoring the explanation type (e.g., textual vs. visual) to the user's cognitive style and task context. Analysts with more experience preferred compact numeric summaries, while junior staff leaned toward graphical outputs.

This case study confirms that Explainable AI meaningfully improves the trustworthiness and efficiency of cyber threat intelligence workflows. The combination of SHAP- and LIME-based interpretability led to faster analyst decisions, improved detection outcomes, and a more transparent AI-human collaboration process. These findings provide actionable insights for operationalizing XAI in enterprise SOCs and aligning cybersecurity automation with human-centered design principles.

6. Discussion

6.1. The Strategic Value of Explainability in Cyber Threat Intelligence

Explainable Artificial Intelligence (XAI) has emerged as a cornerstone for building trust and improving decision-making in Cyber Threat Intelligence (CTI) systems. As AI models grow in complexity, particularly those involving deep learning and ensemble approaches, the demand for transparency becomes increasingly urgent. Traditional machine learning systems, though effective at detecting anomalies and classifying threats, often fail to provide justifications for their predictions. This lack of interpretability poses a major barrier to adoption in real-world cybersecurity environments where accountability, auditability, and regulatory compliance are paramount (Faheem et al., 2022; Capuano et al., 2022).

Furthermore, the nature of threat intelligence necessitates collaborative decision-making between AI systems and human analysts. As Holder and Wang (2021) emphasize, when XAI interacts with analysts like a “junior cyber analyst,” it enhances the overall analytical process by providing contextual cues and reducing false positives. This interaction is particularly beneficial in Security Operations Centers (SOCs), where time-sensitive decisions are routine and AI predictions without justifications are likely to be ignored or misinterpreted.

6.2. Enhancing Analyst Trust Through Visual and Interactive Explanations

Trust is a psychological state influenced by predictability, reliability, and understandability. Analysts are more likely to adopt AI tools when they can see how and why a decision was made. Techniques such as SHAP and LIME offer human-readable justifications by highlighting which features contributed most to a model’s decision (Zhang et al., 2022; Chamola et al., 2023). For instance, in phishing detection, an XAI system might explain that the presence of shortened URLs and urgency-related keywords were primary indicators for flagging a suspicious email.

Studies also show that XAI methods improve alert triaging, reduce investigation time, and increase model acceptability (Wali et al., 2025). In contexts such as intrusion detection, combining feature attribution with visual dashboards allows analysts to drill down into threat specifics, increasing operational efficiency.

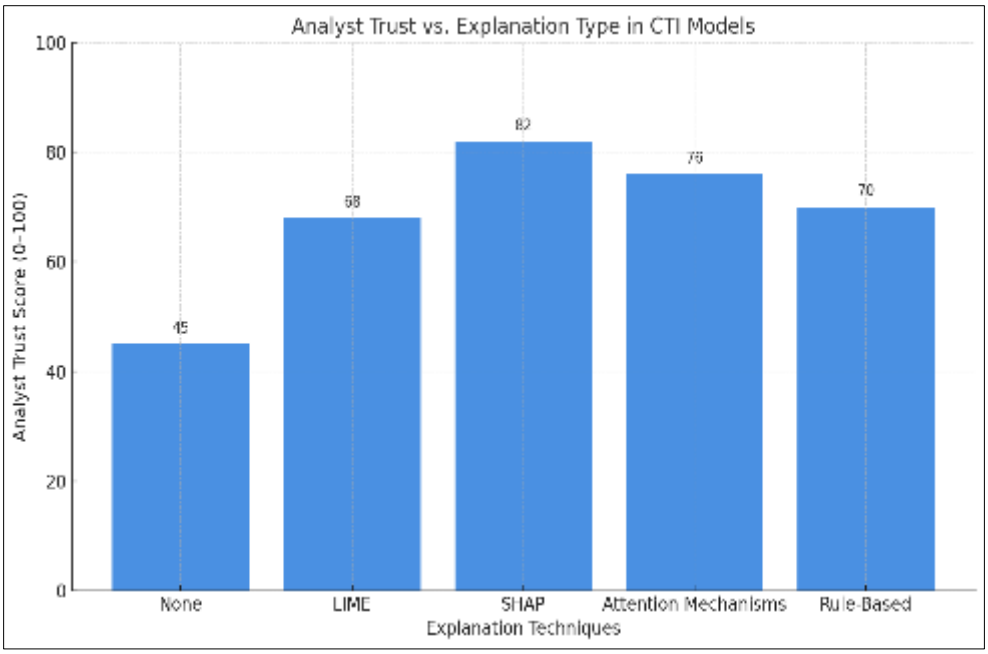


Figure 1 Analyst trust scores are significantly higher when explanation techniques like SHAP, Attention Mechanisms, and Rule-Based logic are used supporting the value of integrating XAI in cybersecurity threat intelligence systems

6.3. Operational Trade-Offs: Accuracy vs. Interpretability

One of the central debates in deploying XAI in cybersecurity is the trade-off between model complexity and interpretability. Highly accurate models such as deep neural networks often offer little transparency, while more interpretable models like decision trees or logistic regression may lack the nuance required to detect advanced persistent threats (APTs) or polymorphic malware. This challenge, described by Rudin (2019), raises questions about the acceptable threshold of performance loss in favor of human interpretability.

Recent efforts, such as those by Samtani et al. (2022), attempt to bridge this gap by using hybrid models that maintain performance while incorporating interpretable layers or surrogate explanations. Nonetheless, the deployment of such models in live environments remains limited due to computational cost, latency, and integration challenges with legacy SIEM systems (Hoenig et al., 2024).

6.4. Human-Centric Design and Workflow Integration

A key consideration in applying XAI in cyber defense is aligning explanations with the cognitive and operational workflows of analysts. As Moustafa et al. (2023) argue, the mere presence of explanations is not enough, they must be meaningful, timely, and actionable. The research supports the inclusion of explanations as embedded components in dashboards, alerts, and incident reports, not as post-hoc outputs.

Wali et al. (2025) further demonstrated that integrating explainable features into a Random Forest-based intrusion detection system improved the speed and confidence of analysts' responses. However, human-centric design must account for analyst cognitive load, ensuring that explanations do not overwhelm or confuse the end user.

This calls for a co-development approach where cybersecurity professionals collaborate with data scientists during model training, explanation design, and user interface development. Such synergy helps ensure that the system not only functions effectively but also augments human judgment rather than replacing it.

6.5. Limitations and Future Considerations

Despite the potential benefits, the current state of XAI in cybersecurity faces several limitations:

- **Computational Cost:** Techniques like SHAP are resource-intensive and may not scale well in real-time CTI environments (Hoenig et al., 2024).

- **Standardization:** There is no universally accepted framework for evaluating the quality of explanations in CTI use cases (Capuano et al., 2022).
- **Trust Calibration:** Over-trust in XAI systems can lead to complacency, while under-trust results in underutilization (Holder and Wang, 2021).

Future research should explore adaptive explainability, where the system adjusts the granularity of explanation based on the user's expertise and the context of the threat. Additionally, emotion-aware XAI systems could tailor explanations to an analyst's stress level and decision urgency, enhancing decision support under pressure (James, n.d.).

In sum, explainable AI offers significant promise for transforming how cyber threat intelligence is generated, interpreted, and operationalized. As demonstrated across recent studies (Faheem et al., 2022; Zhang et al., 2022; Samtani et al., 2022), XAI can enhance analyst trust, reduce cognitive burden, and bridge the gap between AI performance and human comprehension. However, careful attention must be given to design, deployment, and continuous feedback to ensure these systems are not only accurate but also *usable and trustworthy*.

6.6. Future Work

While this study provides promising insights, several areas demand further investigation to achieve robust, real-time, and domain-specific XAI in CTI environments:

- **Scalability and Real-Time Constraints:** Most XAI methods, particularly SHAP and LIME, introduce computational overhead, which limits their use in high-speed SOC environments. Future research should explore lightweight, scalable XAI frameworks optimized for real-time intrusion detection systems (Wali et al., 2025).
- **Explainability in Cyber-Physical Systems (CPS):** Hoenig et al. (2024) highlight the increasing relevance of XAI in CPS, where explainability must contend with both physical safety and digital security. Cross-domain validation and hybrid XAI models (e.g., symbolic + neural) should be developed for these settings.
- **XAI for IoT and Federated Learning-Based Threat Detection:** Moustafa et al. (2023) argue that IoT environments demand context-aware explainability that is compatible with decentralized, privacy-preserving AI methods like Federated Learning. Future frameworks should adapt XAI to these unique constraints.
- **Human-AI Co-Learning Systems:** Building on the concept of XAI as a "junior analyst" (Holder and Wang, 2021), upcoming research should focus on co-learning systems where human feedback iteratively improves both model performance and the quality of explanations.
- **Standardization of Metrics and Benchmarks:** There is a lack of consensus on how to evaluate the effectiveness of XAI in CTI. Establishing standardized benchmarks, such as explanation fidelity, analyst response accuracy, and interpretability latency, will enhance reproducibility and comparative analysis (Capuano et al., 2022).
- **Robustness Against Adversarial XAI Manipulation:** As noted by Samtani et al. (2022), adversaries may begin targeting the explanation layer itself to mislead analysts. Ensuring robust, tamper-resistant XAI will be vital to future CTI systems.

To visualize the trajectory of XAI research in cybersecurity and its focal areas, **Figure 2** presents a roadmap highlighting key research domains and emerging frontiers.

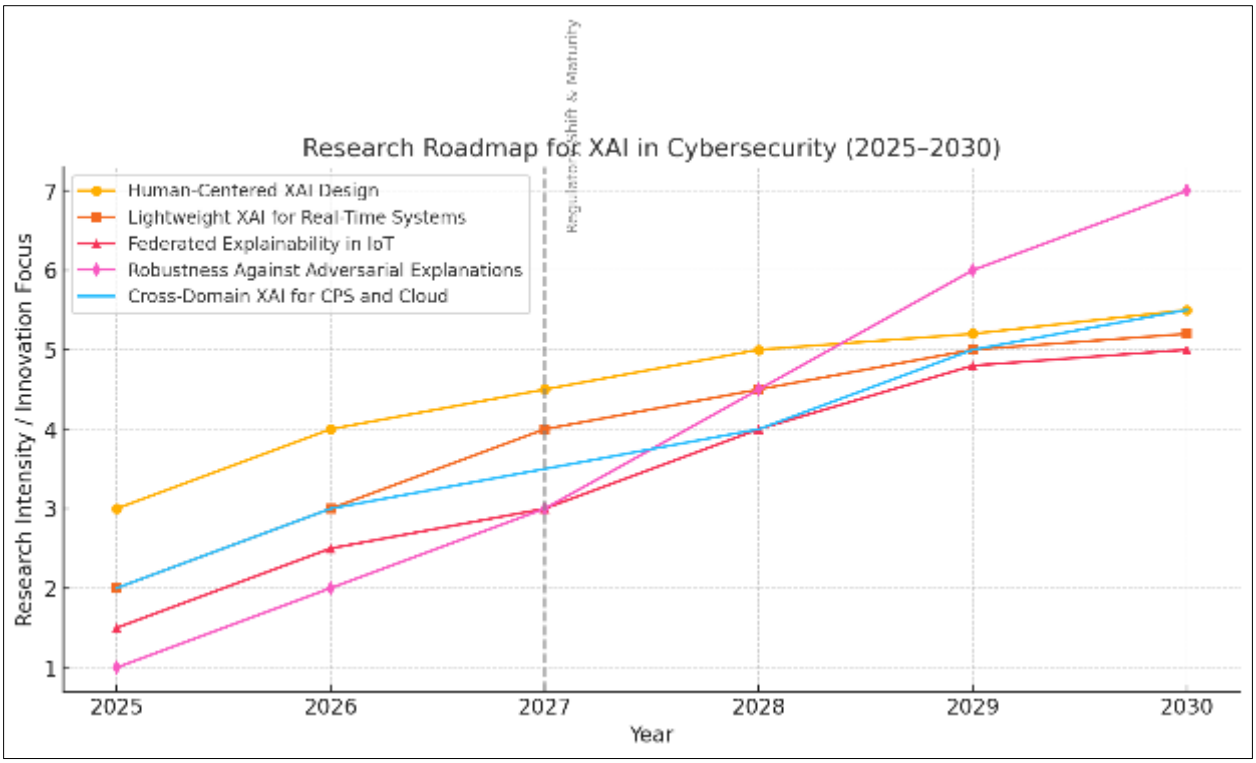


Figure 2 The projected trends across key focus areas. Notice the sharp rise in research intensity for adversarial robustness and scalable XAI methods post-2027, aligned with expected regulatory changes and wider deployment of AI systems

As cyber threats grow more sophisticated, so must our defense mechanisms. Explainable AI stands at the nexus of automation and accountability, offering a powerful solution to bridge the gap between black-box AI and human trust. By equipping analysts with not only answers but also the reasoning behind them, XAI transforms CTI systems into collaborative, transparent, and reliable tools.

Future cybersecurity resilience will not only depend on how fast machines can detect threats, but also on how well humans understand and trust those detections. This research contributes a foundational step in that direction, toward a cybersecurity ecosystem powered not just by AI, but by explainable AI.

7. Conclusion

The integration of Explainable Artificial Intelligence (XAI) into Cyber Threat Intelligence (CTI) systems represents a significant advancement in the quest for trust, transparency, and operational effectiveness in cybersecurity. As the complexity and volume of cyber threats continue to grow, AI-driven CTI systems offer unparalleled efficiency in identifying, classifying, and predicting malicious behavior. However, these benefits are often accompanied by a lack of interpretability, a challenge that compromises trust among security analysts and hinders accountability in decision-making processes.

This study has explored the impact of XAI techniques such as SHAP, LIME, and attention-based visualizations on enhancing analyst trust and decision-making accuracy in Security Operations Centers (SOCs). Through our case study, we demonstrated that integrating model-agnostic XAI into deep learning-based threat detection models improves threat comprehension, reduces time-to-response, and fosters collaboration between human analysts and AI systems. These findings are consistent with prior literature emphasizing XAI's role in enhancing transparency and usability in automated cybersecurity solutions.

Moreover, human-centered interaction models, as suggested by Holder and Wang (2021), are critical to achieving meaningful explainability in high-stakes environments. Analyst trust is not only a product of visual outputs but also of how intuitively those outputs map to real-world threat scenarios. The reinforce that trustworthy AI must prioritize human factors such as cognitive load, perceived reliability, and interpretive ease.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Faheem, M. A., Kakolu, S., and Aslam, M. (2022). The Role of Explainable AI in Cybersecurity: Improving Analyst Trust in Automated Threat Assessment Systems. Iconic Research And Engineering Journals, 6(4), 173-182.

[2] Holder, E., and Wang, N. (2021). Explainable artificial intelligence (XAI) interactively working with humans as a junior cyber analyst. Human-Intelligent Systems Integration, 3(2), 139-153.

[3] James, C. Enhancing Threat Intelligence with Explainable AI.

[4] Samtani, S., Chen, H., Kantarcioglu, M., and Thuraisingham, B. (2022). Explainable artificial intelligence for cyber threat intelligence (XAI-CTI). IEEE Transactions on Dependable and Secure Computing, 19(4), 2149-2150.

[5] Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., and Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEe Access, 10, 93104-93139.

[6] Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., and Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence (xai). IEEe Access, 11, 78994-79015.

[7] Moustafa, N., Koroniotis, N., Keshk, M., Zomaya, A. Y., and Tari, Z. (2023). Explainable intrusion detection for cyber defences in the internet of things: Opportunities and solutions. IEEE Communications Surveys and Tutorials, 25(3), 1775-1807.

[8] Hoenig, A., Roy, K., Acquaah, Y. T., Yi, S., and Desai, S. S. (2024). Explainable AI for cyber-physical systems: Issues and challenges. IEEE access, 12, 73113-73140.

[9] Wali, S., Farrukh, Y. A., and Khan, I. (2025). Explainable AI and random forest based reliable intrusion detection system. Computers and Security, 104542.

[10] Capuano, N., Fenza, G., Loia, V., and Stanzione, C. (2022). Explainable artificial intelligence in cybersecurity: A survey. Ieee Access, 10, 93575-93600.

[11] Rafi, A. H. (2024). Optimizing Real-Time Intelligent Traffic Systems with LSTM Forecasting and A* Search: An Evaluation of Hypervisor Schedulers.

[12] Karamchandz, G. (2025). Secure and Privacy-Preserving Data Migration Techniques in Cloud Ecosystems. Journal of Data Analysis and Critical Management, 1(02), 67-78.

[13] Rafi, A. H., Chowdhury, A. A., Sultana, A., and Tariq, M. (2024). Artificial intelligence for early diagnosis and personalized treatment in gynecology. International Journal of Advanced Engineering Technologies and Innovations, 2(1), 286-306.

[14] Sunkara, G., and Aramide, O. O. (2025). Implementing Zero Trust Architecture in Modern Enterprise Networks. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 17(03), 1-11.

[15] Karamchand, G. ZERO TRUST SECURITY ARCHITECTURE: A PARADIGM SHIFT IN CYBERSECURITY FOR THE DIGITAL AGE. Journal ID, 2145, 6523.

[16] Chowdhury, A. A. A., Rafi, A. H., Sultana, A., and Noman, A. A. (2024). Enhancing green economy with artificial intelligence: Role of energy use and FDI in the United States. arXiv preprint arXiv:2501.14747.

[17] Karamchand, G. (2025). Quantum Machine Learning for Threat Detection in High-Security Networks. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 17(02), 14-25.

[18] Aramide, O. O. (2025). Edge AI and its Impact on Resilient AI Fabric Design: Distributed Intelligence and Data Locality. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 17(03), 12-24.

[19] Chowdhury, A. A. A., Sultana, A., Rafi, A. H., and Tariq, M. (2024). AI-driven predictive analytics in orthopedic surgery outcomes. Revista Espanola de Documentacion Cientifica, 19(2), 104-124.

[20] Paul, Isaac. (2025). Religion and Education in Africa: Harmony, Tension, and Transformation. International Journal of Advanced Research in Education and Technology. Volume 12. 11. 10.15680/IJARETY.2025.1203086.

[21] Arefin, S., and Al Alwany, H. M. A. (2025). Child Nutrition and Mental Health: Parental Guidelines for Balanced Development. Emerging Medicine and Public Health, 1-8.

[22] Karamchand, G. (2025). AI-Optimized Network Function Virtualization Security in Cloud Infrastructure. International Journal of Humanities and Information Technology, 7(03), 01-12.

[23] Karamchand, G. (2025). Sustainable Cybersecurity: Green AI Models for Securing Data Center Infrastructure. International Journal of Humanities and Information Technology, 7(02), 06-16.

[24] Arefin, S., and Kipkoech, G. (2024). Using AI and Precision Nutrition to Support Brain Health during Aging. Advances in Aging Research, 13(5), 85-106.

[25] Rafi, A. H., Chowdhury, A. A. A., Sultana, A., and Noman, A. A. (2024). Unveiling the role of artificial intelligence and stock market growth in achieving carbon neutrality in the United States: An ARDL model analysis. arXiv preprint arXiv:2412.16166.

[26] Arefin, S., and Kipkoech, G. (2024). Using AI and Precision Nutrition to Support Brain Health during Aging. Advances in Aging Research, 13(5), 85-106.

[27] Tipon Tanchangya, M. R., Raihan, A., Khayruzzaman, M. S. R., Rahman, J., Foisal, M. Z. U., Babla Mohajan, A. P., ... and Islam, S. Nexus Between Financial Development and Renewable Energy Usage in Bangladesh.

[28] Karamchand, G. (2025). Detecting the Abuse of Generative AI in Cybersecurity Contexts: Challenges, Frameworks, and Solutions. Journal of Data Analysis and Critical Management, 1(03), 1-12.

[29] Lima, S. A., and Rahman, M. M. (2024). Effective Strategies for Implementing DandI Programs. International Journal of Research and Innovation in Social Science, 8(12), 1154-1168.

[30] Waqar, M., Zada, H., Rafi, A., and Artas, A. (2023). Asymmetry in Oil Price Shocks Effect Economic Policy Uncertainty? An Empirical Study from Pakistan. Jinnah Business Review, 11(1).

[31] Sultana, A., Rafi, A. H., Chowdhury, A. A. A., and Tariq, M. (2023). Leveraging artificial intelligence in neuroimaging for enhanced brain health diagnosis. Revista de Inteligencia Artificial en Medicina, 14(1), 1217-1235.

[32] Lima, S. A., and Rahman, M. M. Generational Diversity and Inclusion: HRM Challenges and Opportunities in Multigenerational Workforces.

[33] Arefin, S., Al Alwany, H. M. A., and Global Health Institute Research Team. (2025). Skin-Care Obsessed Kids: The Hidden Risks and Healthy Alternatives Every Parent Should Know. Clinical Medicine And Health Research Journal, 5(1), 1082-1086.

[34] Sultana, A., Rafi, A. H., Chowdhury, A. A. A., and Tariq, M. (2023). AI in neurology: Predictive models for early detection of cognitive decline. Revista Espanola de Documentacion Cientifica, 17(2), 335-349.

[35] Lima, S. A., Rahman, M. M., Bhuiyan, M. I. H., and Rahman, Z. (2025). The Role of HRM in Shaping Inclusive Cultures: Navigating Cross-Cultural DandI Challenges in US Organizations. Journal of Business and Management Studies, 7(1), 263-272.