International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 13 Issue 5 ǁ May 2025 ǁ PP. 174-180

Enhancing Threat Detection and Response using Explainable AI (XAI): A Literature Review

Harish M S[1], Ankur Khare[2] and [3] Shivamurthaiah M
[1]*Research Scholar, Department of Computer Science and Engineering, Rabindranath Tagore University, Raisen, India, harishbecs@gmail.com*
[2]*Assistant Professor, Department of Computer Science and Information technology, Rabindranath Tagore University, Raisen, India, khareankur94@gmail.com*
[3]*Assistant Professor and HOD, Department of Computer Science & Engineering, APS College of Engineering – VTU, Bengaluru, India,*
Corresponding Author: khareankur94@gmail.com

Abstract
With the quick increase in cyber threats, the request for vigorous and interpretable threat detection systems has convert a critical attention in cybersecurity. Explainable Artificial Intelligence (XAI) delivers a promising pathway to increase cybersecurity by transporting interpretable insights into model choices, offering transparency in threat uncovering and response systems. Unlike outdated machine learning models, which activate as "black boxes," XAI integrates interpretability addicted to the decision-making process, permitting cybersecurity professionals to healthier understand model behaviour, verify forecasts, and identify potential vulnerabilities. This literature review inspects recent developments in XAI techniques applied to cybersecurity, importance how interpretability helps in enhancing detection accuracy, sanitising anomaly detection, and supporting swift response. Additionally, this review converses the strengths and limitations of several XAI methods and classifies emerging challenges and research openings. By consolidating current findings, this learning aims to provide a widespread understanding of XAI presentations in cybersecurity, shedding light on prospect research directions to reinforce model transparency, improve security carriage, and build trust in AI-driven security resolutions.

Date of Submission: 12-05-2025 Date of acceptance: 26-05-2025

I. Introduction

As cyber fears become increasingly sophisticated, administrations face challenges in meritoriously identifying and responding to safety incidents. Traditional cybersecurity solutions rely on configuration recognition and anomaly recognition, yet the complexity and inconsistency of attacks often extract these methods insufficient. In rejoinder, artificial intelligence (AI) and machine learning (ML) have remained incorporated into cybersecurity frameworks to expand detection aptitudes and speed up response periods. However, the "black box" nature of several AI models raises concerns nearby trust, accountability, and the capability to accurately interpret predictions, particularly in high-stakes security environments anywhere transparency is paramount.

Explainable Artificial Intelligence (XAI) has performed as a transformative approach to address these alarms. XAI methods endeavour to make AI verdicts understandable to humans, offering visions into model reasoning, feature standing, and decision pathways. In cybersecurity, XAI can elucidate how models identify potential threats, differentiate caring from malicious activity, and familiarise to evolving attack patterns. This clearness allows security teams to confirm model decisions, identify favouritisms, and establish trust in AI-powered tools.

This literature review affords a comprehensive indication of recent research on the request of XAI in cybersecurity. We discover a diversity of interpretability techniques, such as rule-based methods, object attribution, and visualizations, which contribute to a more vigorous threat detection and retort system. By examining XAI's character in enhancing model interpretability inside cybersecurity, this review aims to classify current advancements, appraise existing challenges, and outline possible pathways for future research. With XAI dignified to become an important part of cybersecurity approaches, understanding its impact on risk detection and response is indispensable to advancing the security landscape.

II. Literature Review

Here is a comprehensive summary for each of the listed papers:

H. C. and P. J. M. P. (2024)

This paper suggests a deep learning-based network imposition detection model planned to optimize detection efficiency while retaining interpretability. The model influences feature selection and pruning to reorganise the architecture, reducing computational cargo without sacrificing accuracy. The explainability of model is heightened by incorporating interpretable layer productivities, which allow cybersecurity analysts to comprehend model predictions and recover threat mitigation strategies. Key discoveries show that this optimized model recovers detection rates and decreases false positives, making it appropriate for real-time applications.

I. H. Sarker et al. (2024)

This work discovers the application of understandable AI (XAI) in digital doubles for cybersecurity. It introduces an arrangement of XAI methods aimed at powering threat detection, attractive trust, and fostering intelligent rejoinders within digital twin systems. The authors discourse the challenges accompanying with model transparency and suggest a framework for integrating XAI into digital identical architectures. This methodology promises improvements in risk detection accuracy and response time though addressing model opacity, which is a communal obstacle in cybersecurity submissions

S. B. A. Maricar et al. (2024)

This paper contributions an enhanced intrusion detection system (IDS) that participates XAI methods to increase transparency in the risk detection process. The model usages an ensemble of machine learning algorithms with editorial acknowledgment methods to provide an interpretable decision-making procedure. The authors attention on improving model robustness alongside evasion techniques commonly used by assailants and demonstrate that their method yields higher detection correctness and interpretability, supporting informed conclusions for incident response teams.

B. T. Familoni (2024)

This paper observes theoretical frameworks and applied solutions to address the cybersecurity tasks presented by AI. The learning discusses the twin role of AI as together a tool for enhancing cybersecurity and for example a target of adversarial outbreaks. The paper highpoints key challenges, including model vulnerabilities and information privacy issues, and recommends strategies like adversarial exercise and secure multi-party computation to mitigate these threats. It accomplishes by stressing the need for robust, understandable AI to progress resilience against emerging cyber threats.

N. Katiyar et al. (2024)

This review refuges the integration of AI in risk detection and comeback systems, specifically focusing on machine learning's starring role in attractive cybersecurity. The authors assess several AI-driven techniques for anomaly discovery and threat identification, noting the reputation of scalability and productivity in real-world applications. While the paper accepts machine learning's probable, it also addresses challenges connected to model interpretability and the necessity for transparent AI to safeguard trust among cybersecurity authorities.

O. Arreche et al. (2024)

This paper recommends the XAI-IDS framework, expected at improving intrusion discovery by incorporating interpretability into the continuous. It utilizes feature standing ranking and visualization tools, permitting security analysts to understand just how the model arrives at this one conclusions. Results indicate that XAI-IDS improves detection accuracy and affords a foundation for translucent, justifiable cybersecurity decisions, with a precise emphasis on the rank of interpretability in complex threat sceneries.

P. P. Kundu et al. (2024)

The authors present a deep learning-based model for identifying and categorising botnet traffic. By integrating

explainability methods like SHAP (SHapley Additive exPlanations), they safeguard that the model's forecasts are transparent, allowing security groups to understand which features donate to identifying botnet behaviour. The learning demonstrates improved accuracy in botnet recognition and highlights the efficiency of using XAI methods for model transparency and operative trustworthiness.

M. Binhammad et al. (2024)

This paper converses AI's role in protection digital identities, focusing on machine learning techniques for verification, incongruity detection, and fraud prevention. The author's current a comprehensive evaluation of AI-driven methods for individuality protection and the importance of explainability in these requests. They contend that while AI improves digital security, model transparency is decisive for building user trust, specifically in sensitive areas like individuality verification.

I. H. Sarker et al. (2024)

The paper propositions a multi-aspect rule-based methodology for AI in dangerous infrastructure protection, categorizing rule-based procedures by functionality and transparency. The authors afford a taxonomy of rule-based AI methods, emphasizing their probable for automated and interpretable risk detection in complex atmospheres. Challenges, such as scalability and adaptableness, are discussed, and instructions for future work on accomplishing high interpretability without compromising performance are charted.

L. Almuqren et al. (2023)

This learning presents an XAI-enabled IDS explicitly designed for cyber-physical systems. The model services feature picturing to enhance interpretability and safeguard that model predictions align with safety policies within cyber-physical atmospheres. Results indicate that the system progresses threat detection accuracy and provisions a transparent approach to perfect decision-making, which is important in settings where human inaccuracy is critical.

A. Oseni et al. (2023)

This paper announces an explainable deep learning framework personalised to IoT-based transference networks, where explainability is important for rapid response to sanctuary incidents. The framework integrates SHAP and LIME techniques to increase model interpretability, enabling sanctuary teams to identify threats and measure their sources. It establishes improved resilience and robustness touching IoT-specific threats, emphasizing the reputation of XAI in maintaining protected, resilient transportation networks.

N. N. tai et al. (2023)

This work attentions on enhancing IDS responsibility by designing interpretable models impervious to evasion attacks. The author's employment adversarial training and feature acknowledgement methods to progress model resilience and interpretability. Their outcomes show that interpretability not individual aids in detecting evasive behaviours however also builds belief in IDS systems, essential for operative cybersecurity defence.

F. Ullah et al. (2023)

This paper progresses a malware detection system that usages transformer-based assignment learning and multi-model visualization methods to provide explainability. The model accomplishes high accuracy in malware detection whereas enabling security teams to imagine the threat's appearances, contributing to a better considerate of the detection process and supplementary in more knowledgeable decision-making.

T. T. H. Le et al. (2023)

This paper exploits ensemble tree models lengthwise with SHAP for feature prominence to enhance the interpretability of IDS productivities. The methodology allows security analysts to measure how different features impact interruption detection and classification decisions, manufacture the system more transparent and cultivating trust in automated sanctuary solutions.

T. Zebin et al. (2022)

This learning addresses DNS over HTTPS (DoH) outbreak detection through a reasonable AI-based approach. The model employments SHAP values to explicate feature contributions, ensuring interpretability. Results highpoint improved accuracy in discovering DoH attacks and demonstrate the probable of XAI to increase trust and usability in compound threat scenarios.

X. Yuan (2019)

This introductory paper examines adversarial outbreaks on deep learning models and offerings various defence mechanisms. The author converses methods to improve model robustness against adversarial management, emphasizing the importance of interpretability in categorizing and mitigating argumentative threats, which is crucial for structure resilient AI in cybersecurity.

Table 1: Comprehensive Review

Paper	Proposed Methodology	Performance Parameters	Advantages	Limitations
H. C. and P. J. M. P. (2024)	**Heightened deep learning-based network intrusion discovery using feature selection and clipping**	**Detection correctness, false positive rate**	**Improved accurateness and interpretability, lower computational cargo**	**May scrap with new, unknown attack categories**
I. H. Sarker et al. (2024)	**XAI methods for cybersecurity in digital doubles, with a focus on mechanization, intelligence, and transparency**	**Trustworthiness, reply time**	**Enhanced limpidity in digital twins, better-quality automation in detection**	**Complex addition with digital twin schemes**
S. B. A. Maricar et al. (2024)	**Enhanced XAI-enabled IDS using ensemble knowledge and feature attribution**	**Detection correctness, model robustness**	**High recognition accuracy, better interpretability for safety teams**	**Possible computational cost due to ensemble systems**
Paper	**Proposed Methodology**	**Performance Parameters**	**Advantages**	**Limitations**
B. T. Familoni (2024)	Hypothetical frameworks addressing cybersecurity challenges in AI structures, with adversarial application and secure multi-party computation	Model robustness, flexibility	Improved resilience and robustness, theoretical deepness	Practical application can be challenging
N. Katiyar et al. (2024)	Machine learning-based risk detection and reaction, with emphasis on scalability	Scalability, effectiveness, interpretability	Enhanced scalability for large systems, efficient discovery	Limited attention on interpretability methods
O. Arreche et al. (2024)	XAI-IDS framework integrating feature ranking and conception	Detection accurateness, interpretability	Transparent risk detection, supports justifiable decisions	May deficiency scalability in large networks
P. P. Kundu et al. (2024)	Deep learning model for botnet discovery with SHAP for explainability	Classification accurateness, transparency	High correctness in botnet detection, improved interpretability	Difficulty in model explanation
M. Binhammad et al. (2024)	AI-driven digital distinctiveness security with attention on explainability in identity protection	Fraud discovery rate, user trust	Builds user belief, effective in individuality protection	Limited examination of AI vulnerabilities
I. H. Sarker et al. (2024)	Multi-aspect rule-based AI designed for transparent risk detection in critical infrastructure	Transparency, flexibility	High transparency, flexibility to complex environments	Incomplete scalability in rapidly evolving cyber sceneries
L. Almuqren et al. (2023)	XAI-enabled IDS aimed at cyber-physical systems with feature conception	Detection accurateness, interpretability	Improved correctness in cyber-physical systems, supports clearness	Necessitates human oversight for critical replies
A. Oseni et al. (2023)	Explicable deep learning for resilient IDS in IoT- transportation networks with SHAP and LIME	Resilience, correctness	Enhanced resilience in IoT networks, limpidity in model output	High source demands for complex IoT structures
N. N. Tai et al. (2023)	Adversarial exercise and feature attribution aimed at evasion-resistant IDS	Model trustworthiness, circumvention resistance	Trustworthy IDS, flexibility against evasive attacks	Incomplete generalizability to diverse attack courses
F. Ullah et al. (2023)	Malware discovery with transformers-based transfer knowledge and multi-model visualization	Detection accurateness, interpretability	High detection correctness, visual insights for risk analysis	Complication in model implementation
T. T. H. Le et al. (2023)	Collective trees with SHAP for feature position in IDS	Classification correctness, feature	Translucent feature impact analysis, effective	Collaborative methods can be resource-intensive

		importance	discovery	
T. Zebin et al. (2022)	SHAP-based XAI for DoH attack discovery	Detection correctness, interpretability	High correctness in DoH detection,	Imperfect adaptability to non- DoH attacks

Paper	Proposed Methodology	Performance Parameters	Advantages	Limitations
			transparent risk analysis	
X. Yuan (2019)	Adversarial exercise methods and defences in deep learning aimed at cybersecurity	Model robustness, protection effectiveness	Enhanced robustness against adversarial examples	Defence effectiveness may destroy against novel attacks

Here are about research gaps and their solutions recognised across the specified papers related to Explainable AI (XAI) in cybersecurity:

1. **Scalability of XAI Models**: Several existing models focus on precise contexts (e.g., intrusion detection in particular atmospheres) without addressing how to measure XAI approaches to handle greater, more complex networks or different attack vectors effectively.

Solution: Progress modular XAI frameworks that can be definitely adapted and scaled to accommodate huge networks. Appliance federated learning approaches to empower the model to learn from dispersed data sources without compromising information privacy.

2. **Generalization Across Attack Types**: Nearby is limited research on emergent XAI frameworks that can oversimplify well across different types of cyber fears (e.g., phishing, ransomware, and insider threats), manufacture it challenging to implement a unified solution.

Solution: Produce ensemble models that combine numerous algorithms trained on various categories of cyber threats. This methodology can help enhance the oversimplification capabilities of XAI models crossways different attack vectors.

3. **Real-Time Interpretability**: Though many studies highlight the standing of interpretability, there is a gap in study focusing on accomplishing real-time interpretability and enlighten ability in dynamic environments, which is critical for effective event response.

Solution: Service lightweight models or approximation procedures that allow for real-time interpretability deprived of sacrificing performance. Techniques identical model distillation can create simplified types of complex models that preserve interpretability while offering speedy responses

4. **Integration with Existing Systems**: Around is a need for frameworks that simplify the seamless combination of XAI models with current cybersecurity systems and workflows to develop usability and efficiency.

Solution: Enterprise XAI frameworks with APIs and plug-in structural design that agree seamless integration with existing cybersecurity tools and structures. Providing clear documents and user-friendly interfaces can facilitate acceptance.

5. **Evaluation Metrics for XAI**: Nearby is a lack of homogenous evaluation metrics for assessing the efficiency of explainable models in cybersecurity situations, which complicates comparisons stuck between different approaches.

Solution: Improve a set of standardized assessment metrics that address both presentation and interpretability, such as operator comprehension tests, trust scores, and operative effectiveness in real-world situations. Collaborate with cybersecurity practitioners to recognize key evaluation criteria.

6. **Handling Evasion Techniques**: Research is imperfect on how XAI can excellently counter advanced evasion systems that adversaries may usage to bypass traditional detection classifications while still maintaining interpretability.

Solution: Investigation adversarial training techniques that definitely enhance XAI models' resilience to

circumvention tactics. This can include generating synthetic data that pretends evasion attempts and incorporating it interested in the training process.

7. **User-Centric Explanations**: Existing models often do not satisfactorily consider the end-user's perspective apropos the types of explanations that would stand most useful aimed at cybersecurity professionals, indicating a gap now user-centric XAI design.

Solution: Behaviour user studies to gather feedback proceeding the types of elucidations that cybersecurity professionals find maximum valuable. Develop customizable explanation agendas that allow users to excellent the level of detail and category of explanation (e.g., visual, textual) that costumes their needs.

8. **Addressing Adversarial Attacks on XAI**: Nearby is a need to discover the vulnerabilities of explainable models themselves toward adversarial attacks, which can challenge the trustworthiness of AI-based structures in cybersecurity.

Solution: Appliance robustness checks and adversarial exercise to fortify XAI models in contradiction of potential manipulation. Researching detection machineries for adversarial attacks on enlightenments can also improve the general trustworthiness of XAI systems.

9. **Comprehensive Taxonomy of XAI Techniques**: Though some papers propose taxonomies, nearby is still a lack of a wide-ranging taxonomy that imprisonments all existing XAI techniques and their applicability in the direction of various cybersecurity domains.

Solution: Collaborate transversely research institutions to collect a comprehensive taxonomy of XAI systems specific to cybersecurity. This can stay achieved through methodical literature reviews and expert panels that recognize and categorize existing approaches.

10. **Longitudinal Studies on Trust and Adoption**: Here is a scarcity of longitudinal trainings investigating how the overview of XAI in cybersecurity inspirations trust and adoption among cybersecurity practitioners in excess of time, which could deliver insights into user approval and practical implementation challenges.

Solution: Stranger longitudinal studies involving real-world distributions of XAI systems in cybersecurity atmospheres. These studies can pathway user experiences, trust levels, and the impression of XAI on decision-making developments over time, provided that valuable insights for future implementations.

These gaps highpoint opportunities for further investigation and development in the pitch of XAI within cybersecurity, suggestive of pathways for future research that can principal to more operative and user-friendly solutions. Implementing these solutions can help address the existing gaps in research and practice, ultimately leading to more effective, reliable, and user-friendly XAI systems in cybersecurity.

III. Conclusion and Future Work

Now conclusion, the integration of Explainable Artificial Intelligence (XAI) addicted to cybersecurity represents an important advancement in enhancing threat discovery and response mechanisms. This literature review has illumined how interpretability cutting-edge AI models not individual improves the accuracy of risk detection but also equips cybersecurity specialists with crucial insights into model behaviour, permitting them to authenticate predictions and proactively address vulnerabilities. The search of several XAI methodologies underscores their probable to transition traditional "black box" structures into transparent, reliable tools that adoptive trust and accountability in AI-driven sanctuary solutions. Though, the review also reveals imperative limitations and challenges associated by the deployment of XAI in real-world applications. Problems such as scalability, computational anxieties, and the necessity for continuous adaptation to developing cyber threats must be addressed to maximize the efficiency of XAI in cybersecurity. Additionally, while the developments are promising, continuing research is required to improve these methodologies, expand their applicability, and increase their interpretability. Observing ahead, future research should attention on developing XAI frameworks that can impeccably integrate with present cybersecurity infrastructures and adapt to emerging fears. By prioritizing transparency and interpretability, the cybersecurity communal can influence the full potential of AI knowledges while fostering superior trust among stakeholders. Ultimately, the development of XAI in cybersecurity is not simply a technical enhancement; it is a needed paradigm shift that aligns with the increasing

imperative for ethical AI practices, robust sanctuary actions, and resilient defines strategies in a progressively complex digital landscape.

References

[1]. H. C. and P. J. M. P., "An Explainable and Optimized Network Intrusion Detection Model using Deep Learning", International Journal of Advanced Computer Science and Applications, Vol. 15 (1), pp. 1-7, 2024.

[2]. I. H. Sarker, H. Janicke, A. Mohsin, A. Gill and L. Maglaras, "Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects", The Korean Institute of Communications and Information Sciences, ICT Express, Elsevier,

[3]. pp. 1-24, 2024.

[4]. S. B. A. Maricar, A. Anoop, B. E. Samuel, A. Appukuttan, and K. H. Alsinjlawi, "An Improved Explainable Artificial for Intrusion Detection System", International Journal of Intelligent Systems and Applications in Engineering, Vol. 12 (14), pp. 108-115, 2024.

[5]. B. T. Familoni, "Cybersecurity Challenges in The Age of AI: Theoretical Approaches and Practical Solutions", Computer Science & IT Research Journal, Vol. 5 (3), pp. 703-724, 2024.

[6]. N. Katiyar, S. Tripathi, P. Kumar, S. Verma, A. K. Sahu and S. Saxena, "AI and Cyber Security: Enhancing Threat Detection and Response with machine learning", Educational Administration: Theory and Practice, Vol. 30 (4), pp. 6273-6282, 2024.

[7]. O. Arreche, T. Guntur and M. Abdallah, "XAI-IDS: Toward Proposing and Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems", Applied Sciences, MDPI, Vol. 14 (4170), pp. 1-41, 2024.

[8]. P. P. Kundu, T. T. Huu, L. Chen, L. Zhou and D. G. Teo, "Detection and Classification of Botnet Traffic using Deep Learning with Model Explanation", IEEE Transactions on Dependable and Secure Computing, pp. 1-15, 2024.

[9]. M. Binhammad, S. Alqaydi, A. Othman and L. H. Abuljadayel, "The Role of AI in Cyber Security: Safeguarding Digital Identity", Journal of Information Security, Vol. 15, pp. 245-278, 2024.

[10]. I. H. Sarker, H. Janicke, M. A. Ferrag and A. Abuadbba, "Multi-aspect rule -based AI: Methods, taxonomy, challenges and directions towards automation, intelligence and transparent cybersecurity modelling for critical infrastructures", Internet of Things, Elsevier, Vol. 25, pp. 1-24, 2024.

[11]. L. Almuqren, M. S. Maashi, M. Alamgeer, H. Mohsen, M. A. Hamza and A. A. Abdelmageed, "Explainable Artificial Enabled Intrusion Detection Technique for Secure Cyber Physical Systems", Applied Sciences, MDPI, Vol. 13 (3081), pp. 1-17, 2023.

[12]. A. Oseni, N. Moustafa, G. Creech, N. Sohrabi, A. Strelzoff, Z. Tari and I. Linkov, "An Explainable Deep Learning Framework for Resilient Intrusion Detection in IoT-Enabled Transportation Networks", IEEE Transactions on Intelligent Transportation Systems, Vol. 24 (1), pp. 1-15, 2023.

[13]. N. N. tai, H. D. Hoang, P. T. Duy and V. H. Pham, "An Interpretable Approach for Trustworthy Intrusion Detection Systems against Evasion Samples", CTU Journal of Innovation and Sustainable Development, Vol. 15, pp. 12-19, 2023.

[14]. F. Ullah, A… Alsirhani, M. M. Alshahrani, A. Alomari, H. Naeem and S. A. Shah, "Explainable Malware Detection System using Transformers based Transfer Learning and Multi Model Visual Representation", Sensors, MDPI, Vol. 22 (6766), pp. 1-22, 2023.

[15]. T. T. H. Le, H. Kim, H. Kang and H. Kim, "Classification and Explanation for Intrusion Detection System based on Ensemble Trees and SHAP Method", Sensors, MDPI, Vol. 22 (1154), pp. 1-28, 2023.

[16]. T. Zebin, S. Rezvy and Y. Luo, "An Explainable AI based Intrusion Detection System for DNS over HTTPS (DoH) Attacks", IEEE Transactions on Information Forensics and Security, pp. 1-12, 2022.

[17]. X. Yuan, "Adversarial examples: Attacks and defenses for deep learning", IEEE Transactions on Neural Networks and Learning Systems, Vol. 30(9), pp. 2805-2824, 2019.