Journal of Digital Science

ISSN 2686-8296

Volume 7 Issue 1 June 2025

Institute of Cited Scientists (ICS)

CONTENTS

Explainable AI for Cybersecurity Applications: A Review Article on Techniques, Deployments, and Usability Challenges

Nipuna Sankalpa Thalpage[1][0009-0001-3374-1927]

Cardiff Metropolitan University, Wales, United Kingdom;
Institute of Cited Scientists, Agia Napa, Cyprus

https://doi.org/10.33847/2686-8296.7.1_1

Received 23.05.2025/Revised 23.06.2025/Accepted 25.06.2025/Published 25.06.2025

Abstract. The growing reliance on Artificial Intelligence (AI) in cybersecurity has elevated concerns about the interpretability and transparency of automated decision-making systems. In environments where trust, accountability, and real-time responsiveness are critical, the "black box" nature of many AI models poses significant barriers to their adoption and operational effectiveness. This systematic literature review examines recent developments in Explainable Artificial Intelligence (XAI) within the cybersecurity domain, focusing on its role in enhancing transparency, trust, and human-AI collaboration. A structured search was conducted across six major academic databases and preprint repositories, yielding nine peer-reviewed studies that met rigorous inclusion criteria. These studies were analyzed across five quality dimensions: relevance, clarity of XAI methods, empirical grounding, human factors consideration, and deployment realism. Findings reveal that while technical innovations—such as SHAP, LIME, Grad-CAM, and lightweight edge-based models—offer substantial gains in model transparency, these advances often fail to translate into actionable insights for end-users due to limitations in cognitive usability and system integration. The review identifies a recurring gap between the theoretical promise of XAI and its practical implementation in real-world security infrastructures. Studies highlight issues such as user disengagement, underutilization of explanation tools, and inadequate alignment with operational workflows. Emerging directions emphasize the need for user-centered design, co-explainability frameworks, and interdisciplinary approaches that incorporate cognitive science and human-computer interaction. In conclusion, the future of XAI in cybersecurity hinges on its ability to go beyond algorithmic transparency and embed interpretability within the social, cognitive, and organizational contexts in which security professionals operate. Bridging these gaps will be essential for realizing the full potential of explainable AI systems as trustworthy and effective tools in modern cybersecurity operations.

Keywords: Explainable AI, Cybersecurity, Artificial Intelligence.

1. Introduction

The increasing reliance on Artificial Intelligence (AI) in cybersecurity has introduced a paradox: while AI models can process massive volumes of data to detect threats with high accuracy, their often opaque decision-making processes can undermine user trust, accountability, and adoption. This has led to the emergence of Explainable Artificial Intelligence (XAI)—a suite of methods designed to make AI decisions interpretable and transparent to human users.

In high-stakes environments like cybersecurity, where decisions affect national security, data privacy, and organizational resilience, explainability is not merely a technical enhancement but a functional necessity. As AI systems are increasingly adopted in threat detection, intrusion response, and risk management workflows, stakeholders require insights not only into what the system predicts, but why. Recent literature underscores that explainability is foundational to trustworthy AI deployment across digital transformation initiatives [1].

To visualize this core challenge, Fig. 1 illustrates the trust gap between algorithmic prediction and human understanding, emphasizing the bridging role of XAI.

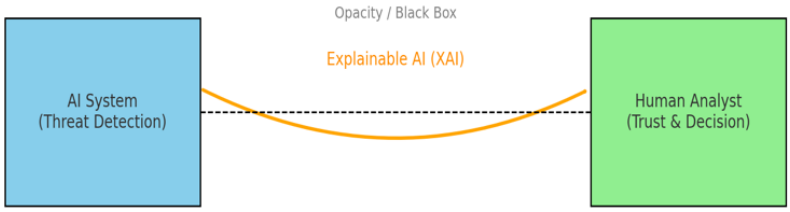

Fig. 1. Gap between algorithmic prediction and human understanding

While AI offers speed and pattern recognition, human analysts must still interpret and act on its outputs—often in high-pressure environments where transparency and timeliness are critical.

This paper conducts a **semi-systematic review** to explore recent developments in the integration of XAI within cybersecurity contexts. Unlike a full systematic review, which aims for exhaustive coverage, this approach emphasizes thematic synthesis of recent peer-reviewed studies that offer representative insights into technical innovations, deployment challenges, and human-centric design considerations. The goal is to identify patterns and gaps across both research and practice, particularly in relation to trust-building, usability, and real-world applicability of XAI methods in security operations.

2. Literature Review

The application of Explainable Artificial Intelligence (XAI) in cybersecurity has emerged as a critical research frontier, driven by the urgent need to balance model performance with interpretability and user trust. As AI-based decision systems become integral to security infrastructure, understanding how and why these systems make decisions is essential for ensuring accountability and facilitating human-AI collaboration.

[2]offer a foundational review of XAI's integration into cybersecurity, emphasizing the need for systematic frameworks to guide implementation and benchmarking effort (Capuano et al., 2022).Their survey outlines the key dimensions of explainability—including fidelity, robustness, and usability—and argues that without clear design principles, XAI applications risk becoming ad hoc and ineffective. Complementing this view, [3] deliver a more taxonomic perspective, dissecting XAI methods based on their applicability to automation, intelligence augmentation, and trust reinforcement in cybersecurity workflows [3]. Together, these works frame XAI not only as a technical tool but as a strategic enabler of secure and transparent AI systems.

However, empirical deployments highlight the disconnect between theoretical promise and operational adoption. [4] conducted a real-world pilot where analysts were provided with XAI-enhanced systems. Despite technical improvements, users often neglected explanation tools and saw limited gains in decision-making accuracy, revealing barriers in human factors such as cognitive load, trust calibration, and explainability literacy[4]. These findings underscore that XAI's success is contingent not only on technical performance but also on user engagement and organizational readiness.

Malware detection has served as a particularly active testbed for XAI integration. [5] developed an approach that integrates memory-based analysis with ML and DL classifiers, leveraging XAI techniques to provide semantic transparency in malware behavior detection. Similarly, [6]also elevated model trustworthiness among analysts. These use cases validate XAI's capacity to demystify opaque deep learning predictions in high-stakes environments.

Addressing the challenge of deployability in edge and constrained environments, [7] introduced an Explainable and Lightweight AI (ELAI) framework designed for real-time threat detection on edge networks. By optimizing for both interpretability and computational efficiency, ELAI demonstrates that XAI can be adapted to function in latency-sensitive and resource-limited settings, broadening its practical relevance.

Beyond technical enhancements, [8] highlight how XAI can be used to strengthen human-AI collaboration by fostering mutual transparency in mixed-initiative systems. Their proposed framework aligns with the broader trend of embedding XAI into design thinking, ensuring that interpretability is not retrofitted but built into AI systems from the outset.

Additionally, recent studies by [9] and [10] delve into the psychosocial dimensions of XAI deployment in cybersecurity. They argue that successful integration requires addressing user perception, cognitive models, and behavior dynamics, especially in high-stress operational environments. This pivot from algorithmic focus to user-centricity is critical for ensuring XAI's real-world effectiveness.

In summary, the evolving literature paints a nuanced picture: while XAI offers transformational benefits in transparency, auditability, and trust, its practical efficacy hinges on human, technical, and infrastructural readiness. Continued research must focus not only on refining explanation methods but also on standardizing evaluation, modeling user interaction, and supporting real-time, lightweight deployments.

3. Data and Methodology

3.1 Search Strategy

A structured search was conducted to identify relevant literature focusing on the application of Explainable Artificial Intelligence (XAI) in cybersecurity. The following digital databases were queried between year 2022 and 2025:

- IEEE Xplore
- ACM Digital Library
- SpringerLink
- ScienceDirect
- arXiv (for preprints)
- Google Scholar

Search terms were combined using Boolean operators and included variations of: ("Explainable AI" OR "XAI") AND ("cybersecurity" OR "malware detection" OR "edge computing" OR "trust in AI" OR "AI transparency").

This approach reflects established practices in systematic review research where structured database searches and transparent inclusion logic are used to ensure replicability and thematic rigour [11].

3.2 Inclusion Criteria

To ensure relevance and rigor, studies were included if they:

- Were published between January 2022 and May 2025
- Focused specifically on the application or evaluation of XAI techniques in a cybersecurity context
- Provided empirical evidence, theoretical frameworks, or human-centered analysis
- Employed or discussed recognized XAI techniques (e.g., SHAP, LIME, Grad-CAM, counterfactuals)
- Were peer-reviewed or part of recognized preprint repositories (e.g., arXiv)

3.3 Exclusion Criteria

Studies were excluded if they:

- Addressed explainability in non-cybersecurity domains (e.g., healthcare, finance)
- Focused exclusively on general AI or ML techniques without linking to interpretability or transparency
- Were non-English publications or lacked accessible full-text
- Consisted solely of opinion pieces or short workshop abstracts without substantial methodological or empirical depth

3.4. Screening Process

All identified records were first screened by title and abstract. Duplicates were removed. Full-text reviews were then conducted on shortlisted studies. In total, nine key papers were selected based on their thematic alignment with the goals of this literature review, which emphasize trust, transparency, human factors, and edge deployment in XAI for cybersecurity.

3.5. Data Extraction and Synthesis

For each study, information was extracted regarding:

- Authors and publication year
- Cybersecurity focus area
- XAI techniques employed
- Empirical findings or proposed frameworks
- Insights on usability, trust, and system performance

3.6 . Study Quality Assessment

To ensure a nuanced understanding of each selected study, a structured quality evaluation was conducted using five dimensions:

1. Relevance to Cybersecurity – alignment of study focus with XAI applications in security domains.
2. Clarity of XAI Methods – transparency in describing the explainability techniques used.

3. Empirical Evidence – presence of experiments, deployments, or evaluations.
4. Human Factors Consideration – integration of trust, cognitive modeling, or usability in the design.
5. Deployment Realism – realism of setting, such as field trials or edge computing integration.

Each study was scored qualitatively across these dimensions (✓ = present, ⚪ = limited/absent), resulting in a composite quality impression.

Table 1. Quality assessment of the papers

Study	Relevance	XAI Clarity	Empirical Evidence	Human Factors	Deployment Realism	Quality Score (out of 5)
Capuano et al., 2022	✓ High	✓ Clear	⚪ Theoretical	⚪ Not addressed	⚪ Conceptual	3
Sarker et al., 2024	✓ High	✓ Detailed	⚪ Review-based	⚪ Partial	⚪ Conceptual	3
Nyre-Yu et al., 2022	✓ High	✓ Applied (LIME)	✓ Pilot study	✓Key focus	✓ Real-world	5
Ravikumar, C. *et al.,* 2024	✓ High	✓ Applied	✓ Experimental	⚪ Limited	⚪ Lab setting	4
Nazim et al., 2025	✓ High	✓ Visual methods	✓ Comparative testing	✓ User feedback	⚪ Simulated	4
Rahmati, 2025	✓ High	✓ Lightweight	✓ Prototyping	⚪ Minimal	✓ Edge deployment	4
Desai et al., 2024	✓ High	✓ UI design	⚪ Design framework	✓Core theme	⚪ Conceptual	4
Pan et al., 2023	✓ High	✓ Cognitive model	⚪ User analysis	✓ Psychology focus	⚪ No deployment	4
Barletta et al., 2023	✓ High	✓ Behavioral framing	⚪ Design-oriented	✓ Key focus	⚪ No empirical test	4

4. Discussion and Findings

The literature reveals a dynamic and interdisciplinary landscape where Explainable Artificial Intelligence (XAI) is increasingly being adopted to enhance cybersecurity systems. A total of nine studies met the inclusion criteria and were thematically analyzed. These studies span both theoretical frameworks and empirical deployments, encompassing aspects such as malware detection, real-time edge computing, trust in AI systems, and human-AI collaboration.

4.1 Thematic Insights

1. Framework Development and Taxonomies
   [2] and [3] offer foundational models and taxonomies that categorize XAI approaches based on their utility in cybersecurity. These contributions emphasize the need for systematic integration rather than ad hoc use of explainability techniques.

2. Usability and Trust
   [4] conducted one of the few field deployments, revealing a significant gap between technical explainability and human comprehension. Despite tool

availability, analyst trust and engagement did not measurably improve, highlighting a need for cognitive alignment and interface design.

3. Malware Detection Use Cases
[5] and [6]illustrate how XAI enhances transparency in malware classification. Their studies show that visual and memory-based explainability methods can support both model validation and operational clarity.

4. Edge and Real-Time Environments
[7] introduces a lightweight, explainable model tailored for resource-constrained environments. This reflects a growing interest in deploying XAI beyond centralized systems into edge networks where both speed and clarity are critical.

5. Human-Centric Design
[8], [9], and [10] push the discourse toward human factors, exploring how cognitive ergonomics, behavior modeling, and co-explainability foster effective human-AI synergy in high-stress environments.

A qualitative synthesis was conducted to compare and contrast methodological approaches, reported outcomes, and practical implications. A summary table was included to highlight recurring themes and gaps in current research.

4.2 Synthesis of Literature

Table 2. Literature Synthesis from papers

Author(s)	Topic	XAI Methods	Key Insight
Capuano et al., 2022	Survey/Frameworks	General XAI taxonomy	Calls for standardized frameworks and evaluation in cybersecurity XAI
Sarker et al., 2024	Method taxonomy & challenges	SHAP, LIME, visualizations	Reviews XAI methods by application (trust, automation, intelligence)
Nyre-Yu et al., 2022	Field study (pilot deployment)	LIME, XAI dashboards	Found XAI tools underutilized by analysts; trust did not increase
Ravikumar, C. et al. (2024)	Malware detection	ML/DL + explainable output	Enhanced classification using memory analysis + XAI
Nazim et al., 2025	Malware image classification	SHAP, LIME, Grad-CAM	Visual XAI improves deep learning model transparency
Rahmati, 2025	Real-time edge security	Lightweight XAI (ELAI)	Combines speed + explainability for edge networks
Desai et al., 2024	Human-AI collaboration	XAI-aware interfaces	Emphasizes co-explainability to enhance human-AI synergy
Pan et al., 2023	Human factors in AI	Cognitive models	Studies how users perceive and engage with XAI outputs
Barletta et al., 2023	Behavioral UX design	Human-centric XAI	Suggests aligning XAI with cognitive ergonomics

Further, to illustrate the distribution of XAI methods across the reviewed studies, Figure 1 summarises the frequency with which key techniques such as SHAP, LIME, Grad-CAM, and custom models were employed.

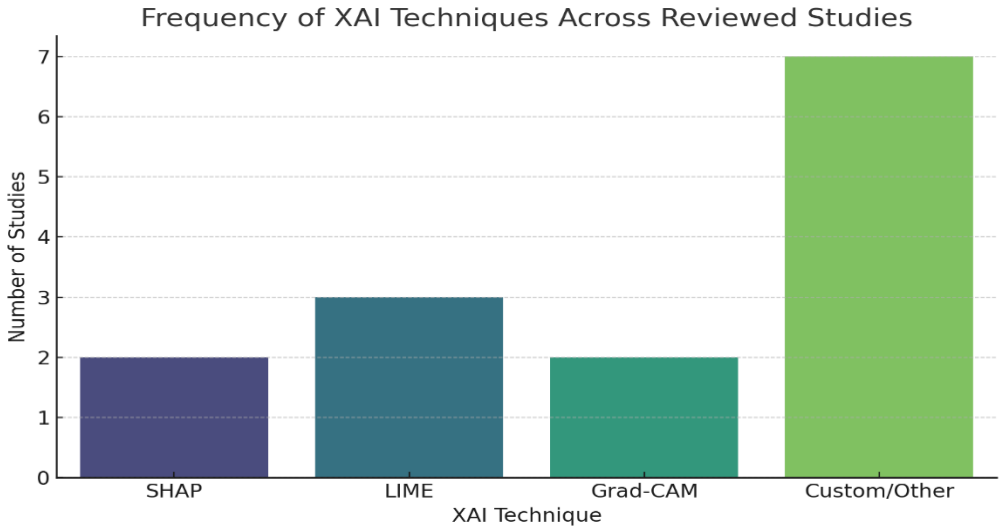

Fig. 2. Distribution of XAI methods across the studies

5. Conclusion

This review underscores that while Explainable Artificial Intelligence (XAI) offers considerable potential to advance cybersecurity systems, its practical impact in real-world settings is shaped by more than just algorithmic sophistication or model transparency. Across the reviewed literature, a consistent theme emerges: the success of XAI in cybersecurity is not merely a function of its ability to explain, but of its ability to be understood, trusted, and effectively used by human operators in dynamic, high-stakes environments.

The selected studies demonstrate promising progress in technical innovation—ranging from the use of visual interpretability methods such as SHAP and Grad-CAM to the development of lightweight, real-time explainable frameworks optimized for edge computing environments. These contributions reflect a broader evolution in the field: a shift from viewing explainability as a post hoc add-on, toward designing AI systems that embed interpretability into their architecture from the ground up. In particular, research exploring co-explainability frameworks, which aim to foster mutual understanding between human analysts and AI systems, signals an encouraging move toward user-centered and collaborative design principles.

However, this optimism is tempered by persistent challenges that limit the operational effectiveness of XAI. Multiple field and simulation studies highlight that even when explainability tools are available, analysts often overlook or underutilize them, pointing to a crucial disconnect between technical explainability and cognitive usability. This disconnect is exacerbated in high-pressure cybersecurity operations where users are inundated with information, under time constraints, and often lack the training or cognitive bandwidth to interpret complex AI outputs. As a result, XAI frequently fails to translate from a desirable technical feature into a meaningful decision support tool.

Furthermore, the review identifies significant gaps in the current research ecosystem. Many studies still operate within controlled or theoretical settings, with limited real-world deployment or user-centered evaluation. Few works directly address organizational factors, such as workflow integration, team dynamics, or training ecosystems, which are critical for the sustained and effective use of XAI tools

in security operations. There is also a notable underrepresentation of behavioral and psychological research that could inform how users perceive, engage with, and calibrate trust in AI systems.

To bridge these gaps, future work must expand beyond the algorithmic domain and embrace a holistic, interdisciplinary approach. This includes incorporating insights from human-computer interaction (HCI), cognitive science, organizational psychology, and systems engineering. Interface design should prioritize intuitive, context-aware visualizations, while system architectures should allow for interactive, dialogic forms of explanation rather than static outputs. Evaluation metrics should also evolve to include user trust, situational awareness, and decision quality, rather than model performance alone.

In sum, the path forward for XAI in cybersecurity demands a synthesis of technical excellence and human-centered design. Only by addressing the cognitive, behavioral, and operational dimensions of explainability can we ensure that XAI systems are not only interpretable in theory but also empowering, actionable, and impactful in practice.

References

1. Barletta, V.S. *et al.* (2023) 'Serious Games for Cybersecurity: How to Improve Perception and Human Factors', in *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*. IEEE, pp. 1110–1115. Available at: https://doi.org/10.1109/MetroXRAINE58569.2023.10405607.
2. Capuano, N. *et al.* (2022) 'Explainable Artificial Intelligence in CyberSecurity: A Survey', *IEEE Access*, 10, pp. 93575–93600. Available at: https://doi.org/10.1109/ACCESS.2022.3204171.
3. Desai, B. *et al.* (2024) 'Explainable AI in Cybersecurity: A Comprehensive Framework for enhancing transparency, trust, and Human-AI Collaboration', in *2024 International Seminar on Application for Technology of Information and Communication (iSemantic)*. Semarang, Indonesia: IEEE, pp. 135–150. Available at: https://doi.org/10.1109/iSemantic63362.2024.10762690.
4. Milad Rahmati (2025) 'Towards Explainable and Lightweight AI for Real-Time Cyber Threat Hunting in Edge Networks'.
https://doi.org/10.48550/arXiv.2504.16118

5. Nazim, S. *et al.* (2025) 'Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM', *PLOS One*, 20(5), p. e0318542. Available at: https://doi.org/10.1371/journal.pone.0318542.
6. Nyre-Yu, M. *et al.* (2022) 'Explainable AI in Cybersecurity Operations: Lessons Learned from xAI Tool Deployment', in *Proceedings 2022 Symposium on Usable Security*. Reston, VA: Internet Society. Available at: https://doi.org/10.14722/usec.2022.23014.
7. Pan, Z. and Mishra, P. (2023) *Explainable AI for Cybersecurity*. Cham: Springer Nature Switzerland. Available at: https://doi.org/10.1007/978-3-031-46479-9.
8. Ravikumar, C. *et al.* (2024) 'Advancing Malware Detection Using Memory Analysis and Explainable AI Approach', in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*. IEEE, pp. 518–523. Available at: https://doi.org/10.1109/ICoICI62503.2024.10696406.
9. Sarker, I.H. *et al.* (2024) 'Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects', *ICT Express*, 10(4), pp. 935–958. Available at: https://doi.org/10.1016/j.icte.2024.05.007.
10. Thalpage, N. (2024) 'The Integration of Machine Learning and Explainable AI and Business Digitization: Unleashing the Power of Data - A Review', *Journal of Digital Science*, 6(1), pp. 18–27. Available at: https://doi.org/10.33847/2686-8296.6.1_2.
11. Thalpage, N.S. and Nisansala, T.A.D. (2023) 'Exploring the Opportunities of Applying Digital Twins for Intrusion Detection in Industrial Control Systems of Production and Manufacturing – A Systematic Review', in *Data Protection in a Post-Pandemic Society*. Cham: Springer International Publishing, pp. 113–143. Available at: https://doi.org/10.1007/978-3-031-34006-2_4.

Aims and Objectives

Published online by Institute of Cited Scientists, Cyprus, two times a year, Journal of Digital Science (JDS) is an international peer-reviewed journal which aims at the latest ideas, innovations, trends, experiences and concerns in the field of digital science covering all areas of the scholarly literature of the sciences, social sciences and arts & humanities.

The principal ambition of this periodical is the efficacious propagation of original insights and outcomes derived from human cerebral activity and exemplified in scholarly treatises through the utilisation of contemporary information and digital technology.

The main topics currently covered include: Artificial Intelligence in Cybersecurity and Decision-making process, Cryptocurrency as Digital Assets, Digital decoding of language, Digital Technology in Health Care.

Contact information

Journal URL: https://ics.events/journal-of-digital-science/
Email: conf@ics.evnets
Printed online from the original layout under the imprint at:
1, Vlachou, Nicosia, The Republic of Cyprus

The picture on JDS cover was generated by Andrei Stepanov.