INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING &TECHNOLOGY (IJCET)

ISSN Print: 0976-6367
ISSN Online: 0976-6375

Publishers of High Quality Peer Reviewed Refereed Scientific, Engineering & Technology, Medicine and Management International Journals

PUBLISHED BY

IAEME

IAEME Publication

Chennai, India

https://iaeme.com/Home/journal/IJCET

International Journal of Computer Engineering and Technology (IJCET)
Volume 16, Issue 4, July-August 2025, pp. 170-182, Article ID: IJCET_16_04_012
Available online at https://iaeme.com/Home/issue/IJCET?Volume=16&Issue=4
ISSN Print: 0976-6367; ISSN Online: 0976-6375; Journal ID: 5751-5249
Impact Factor (2025): 18.59 (Based on Google Scholar Citation)
DOI: https://doi.org/10.34218/IJCET_16_04_012

TRANSFORMING CYBER DEFENSE THROUGH EXPLAINABLE AI: INTERPRETABILITY IN SECURITY CONTEXTS

Sri Ramya Deevi

USA.

ABSTRACT

Artificial Intelligence (AI) plays an increasingly vital role in modern cybersecurity, enabling faster detection of threats, automated responses, and adaptive defense mechanisms. Many AI models function as black boxes, lacking transparency and interpretability an issue that significantly limits their adoption in critical security contexts where accountability, trust, and human decision-making are essential. This paper investigates the transformative impact of Explainable AI (XAI) in cyber defense, focusing on how interpretability can enhance threat detection, support compliance, and empower analysts to make informed decisions. I provide a comprehensive overview of XAI techniques, including SHAP, LIME, counterfactual explanations, and saliency maps, and evaluate their effectiveness in applications such as intrusion detection, malware classification, and phishing detection. A novel framework is proposed for integrating XAI into existing security architectures, emphasizing user-centric explanations and real-time decision support. I demonstrate that incorporating XAI not only improves model transparency but also strengthens operational effectiveness. The paper concludes with a discussion on current challenges, such as adversarial risks and cognitive burden, and outlines future directions for research, policy, and governance.

My findings suggest that explainability is not just an enhancement, but a fundamental requirement for trustworthy and resilient cyber defense systems.

Keywords: Explainable AI (XAI), Cybersecurity, Interpretable Machine Learning, Threat Detection, Intrusion Detection Systems (IDS)

Cite this Article: Sri Ramya Deevi. (2025). Transforming Cyber Defense Through Explainable AI: Interpretability in Security Contexts. *International Journal of Computer Engineering and Technology (IJCET)*, 16(4), 170–182.

DOI: https://doi.org/10.34218/IJCET_16_04_012

1. Introduction

As cyber threats continue to evolve in scale, complexity, and sophistication, artificial intelligence (AI) has emerged as a critical asset in modern cyber defense. AI-powered systems offer capabilities such as real-time anomaly detection, automated response mechanisms, and adaptive threat modeling, making them indispensable in securing digital infrastructures [1]. Many of these systems particularly those based on deep learning and complex ensemble models operate as black boxes, providing little to no insight into how decisions are made. This opacity presents serious challenges in security-critical domains where trust, accountability, and regulatory compliance are paramount [2].

Explainable AI (XAI) seeks to bridge this gap by making the internal logic of AI systems transparent and interpretable to human stakeholders. While XAI has gained traction in fields like healthcare and finance, its application in cybersecurity remains underexplored and fragmented [3]. Cybersecurity analysts often face high-stakes, time-sensitive decisions that require clear justifications for AI-driven alerts or classifications. Without interpretability, human operators may disregard accurate warnings or, worse, trust flawed outputs blindly. This paper investigates the role of XAI in transforming cyber defense strategies. I explore existing interpretability techniques, assess their applicability to security use cases, and propose an integrated framework tailored for Security Operations Centers (SOCs). I argue that explainability is not merely a beneficial feature but a foundational requirement for trustworthy and effective cyber defense systems.

2. The Need for Explainability in Cyber Defense

The application of AI in cybersecurity offers significant potential for enhancing detection and response capabilities, yet the interpretability of these systems remains a major concern. In high-stakes environments such as national defense, critical infrastructure, and financial systems, the ability to understand and trust AI-generated outputs is not optional it is essential. Explainable AI (XAI) addresses this concern by providing human understandable justifications for AI decisions, enabling security analysts to verify, contest, or act on alerts with confidence.

From a regulatory standpoint, frameworks such as the European Union's General Data Protection Regulation (GDPR) mandate the "right to explanation" for decisions made by automated systems, including those used in cybersecurity [4]. The U.S. National Institute of Standards and Technology (NIST) emphasizes explainability as a core principle in its AI Risk Management Framework [5].

Cyber analysts are often overwhelmed with alert fatigue due to the high volume of false positives from detection systems. A lack of context or reasoning behind AI decisions exacerbates this issue, leading to mistrust or underutilization of advanced detection tools [6]. Explainability not only helps reduce response time by clarifying why an alert is raised, but it also enhances collaboration between humans and machines creating a more effective defense posture.

Adversaries increasingly exploit AI models through evasion and poisoning attacks. Interpretable models help security teams understand vulnerabilities in their defenses and improve the robustness of threat detection systems [7]. In this context, explainability functions as both a technical and strategic asset, empowering organizations to operate with resilience, transparency, and accountability in an increasingly hostile digital environment.

3. Overview of Explainable AI (XAI)

Explainable Artificial Intelligence (XAI) refers to a suite of methods and tools designed to make the behavior, decisions, and predictions of AI models comprehensible to humans. In contrast to black-box models particularly deep neural networks XAI aims to provide transparency, foster trust, and support effective human-machine collaboration in high-risk domains such as cybersecurity. XAI techniques are generally classified into two categories: intrinsic and post hoc explainability. Intrinsic methods involve designing interpretable models

from the outset, such as decision trees, linear models, or rule-based systems. These models are inherently more transparent but may suffer from reduced accuracy when handling complex patterns. In contrast, post hoc methods aim to interpret black-box models after training, using techniques like Local Interpretable Model-agnostic Explanations (LIME) [8], SHapley Additive exPlanations (SHAP) [9], saliency maps, and counterfactual reasoning [10].

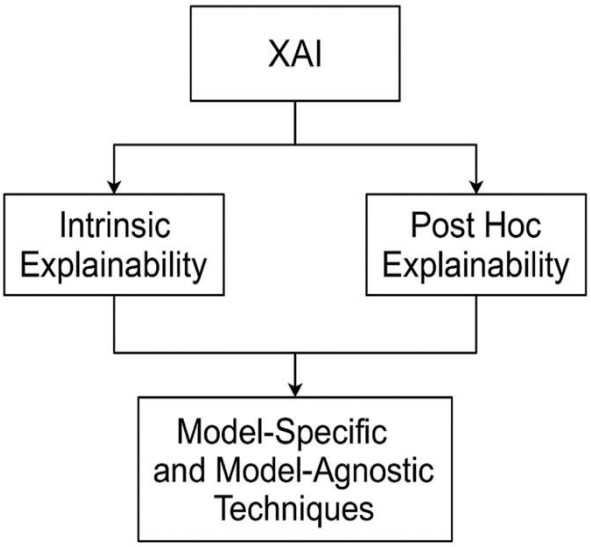

Figure 1. Overview of Explainable AI (XAI)

Another key distinction is between model-specific and model-agnostic approaches. Model-specific techniques rely on internal access to the model structure and gradients saliency maps in convolutional neural networks, while model-agnostic methods operate as wrappers that can be applied to any classifier or regressor. Despite their advantages, XAI methods present trade-offs between accuracy, fidelity, stability, and computational overhead. In security contexts, where the cost of false positives or missed detections is high, balancing interpretability with performance is essential.

The utility of an explanation varies based on the user developers may seek technical insights into model internals, while security analysts prioritize actionable insights that guide response. Thus, effective XAI design must be context-aware, user-centric, and adaptable to evolving threat landscapes.

4. XAI in Security Applications: Use Cases and Techniques

The deployment of Explainable AI (XAI) in cybersecurity operations provides security professionals with critical visibility into AI-driven threat assessments. This section highlights key use cases intrusion detection, malware classification, and phishing detection and explores XAI techniques applied within each.

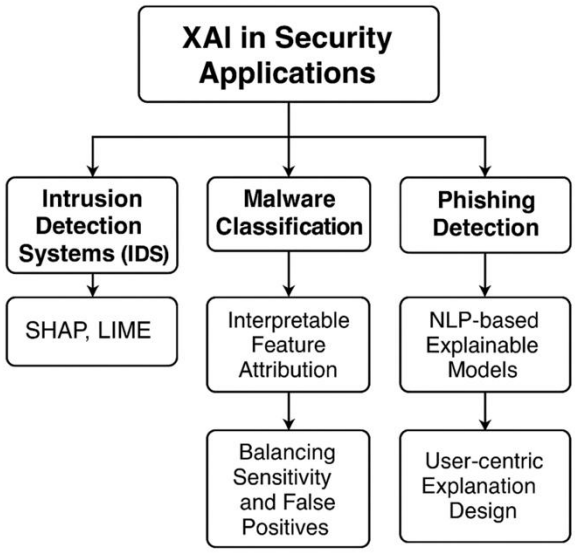

Figure 2. XAI Security Applications

Intrusion Detection Systems (IDS)

Intrusion Detection Systems benefit significantly from interpretability, as analysts must evaluate large volumes of alerts with minimal context. XAI methods such as SHAP and LIME have been used to explain model predictions in IDS, providing clarity on which network features unusual port activity, packet rates contributed most to classifying an event as malicious [11]. Researchers have demonstrated that incorporating feature attribution improves analyst confidence and reduces false positive triage times [12].

Malware Classification

Deep learning classifiers for malware detection often lack transparency, making them difficult to trust. XAI techniques like gradient-based saliency maps and layer-wise relevance propagation (LRP) are employed to visualize which byte sequences or API calls most influenced a malware prediction [13]. These insights help reverse engineers and forensic analysts understand model reasoning and identify novel attack patterns.

Phishing and Social Engineering Detection

Phishing detection increasingly uses natural language processing (NLP) models to analyze emails and URLs. Model-agnostic XAI methods can highlight specific textual indicators urgent phrases, misspellings, suspicious domains that led to a phishing classification [14]. This transparency enhances training for end users and helps security teams develop targeted response strategies.

These applications illustrate the dual role of XAI in improving detection performance and fostering human-AI collaboration. By making model outputs interpretable, security teams can better prioritize risks, investigate anomalies, and adapt defense mechanisms.

5. Proposed Framework for XAI-Enabled Cyber Defense

To operationalize Explainable AI (XAI) in cybersecurity, I propose a modular framework that integrates interpretable machine learning components into existing security infrastructures such as Security Information and Event Management (SIEM) and Security Orchestration, Automation, and Response (SOAR) systems. The framework is designed to facilitate actionable, transparent, and real-time threat assessments while maintaining system scalability and adaptability.

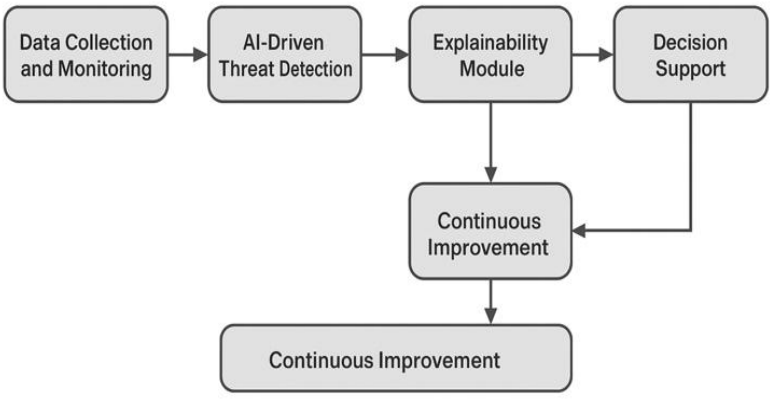

Figure 3. Framework for XAI-Enabled Cyber Defense

Architectural Components

Data Ingestion Layer: This component aggregates heterogeneous data sources network traffic, system logs, endpoint telemetry and performs normalization and preprocessing. Ensuring high-quality, labeled input data is foundational to effective XAI [15].

Model Layer with Integrated XAI: The core analytics engine includes ML/DL models with embedded or post hoc explainability methods like LIME, SHAP, attention mechanisms. Hybrid models may combine rule-based systems with deep learning to balance interpretability and performance [16].

Explanation Interface Layer: A visual interface presents explanations tailored to different stakeholders. Analysts receive feature importance scores or saliency maps, while executive dashboards may include natural language justifications [17].

Feedback Loop and Adaptive Learning: A human-in-the-loop mechanism captures user feedback on explanations and detection outcomes, enabling model retraining and explanation refinement, thus supporting continual learning [18].

Evaluation and Metrics

The framework integrates interpretability-specific metrics such as fidelity, stability, and cognitive load alongside traditional performance metrics like accuracy, precision, and recall. Evaluating user trust, explanation usability, and response time improvements is essential for validating effectiveness in live environments [19].

Integration with SOC Workflows

The proposed architecture is designed for plug-and-play integration into SOC workflows. It supports alert prioritization, root-cause analysis, and incident triage with transparency. The framework aligns with NIST AI RMF guidelines to support risk-based governance in AI-driven systems [20].

6. Challenges and Limitations

Despite the promise of Explainable AI (XAI) in enhancing cyber defense, several critical challenges and limitations remain that hinder its widespread adoption and operational maturity.

Scalability of XAI Methods

Many popular XAI methods such as SHAP and LIME are computationally expensive, especially when applied to large-scale, high-dimensional datasets typical of cybersecurity environments. Real-time applications, such as network intrusion detection and endpoint monitoring, demand low-latency responses, which can be incompatible with the time-intensive nature of many explanation algorithms [21].

Adversarial Manipulation of Explanations

Recent research highlights that explanations themselves can be manipulated by adversaries to mislead security analysts or obscure malicious behavior. Attackers may craft inputs that not only evade detection but also produce misleading, benign-looking explanations compromising the trustworthiness of the entire defense system [22].

Ambiguity and Inconsistency in Explanation Quality

There is no universally accepted metric for evaluating the quality of explanations. Metrics such as fidelity, comprehensibility, and stability often yield conflicting evaluations across models and tasks. Explanations that are technically accurate may still be unusable for non-expert analysts due to cognitive overload or lack of contextual relevance [23].

Cognitive Load on Analysts

Overly detailed or abstract explanations can overwhelm users, especially during high-pressure incidents where time-sensitive decisions are required. Analysts require concise, relevant, and contextual insights that support rapid interpretation and action, which many current XAI systems are not optimized to deliver [24].

Lack of Standardization and Regulatory Alignment

Although frameworks such as NIST AI RMF provide general guidance, there is limited standardization specific to cybersecurity XAI applications. A lack of domain-specific regulatory standards complicates the integration of explainability into governance, auditing, and certification workflows [25].

Overcoming these challenges requires multidisciplinary collaboration between AI researchers, cybersecurity practitioners, human-computer interaction experts, and regulatory bodies. Addressing these limitations will be essential for achieving operational-grade XAI systems that are scalable, trustworthy, and usable in real-world defense environments.

7. Future Directions

As cybersecurity threats evolve and the complexity of AI models increases, the development of next-generation Explainable AI (XAI) systems for cyber defense must be guided by interdisciplinary innovation, adaptive design, and human-centered priorities. The following future directions offer promising pathways to address current limitations and enhance operational resilience.

Hybrid Human-AI Threat Hunting: Future XAI systems will likely support collaborative threat hunting, combining machine speed with human intuition. By embedding real-time explainability into security operations, analysts can rapidly validate AI-generated insights, identify false positives, and uncover sophisticated multi-stage attacks [26]. Active learning paradigms, where the system queries humans for ambiguous cases, can further enhance detection accuracy while refining the model over time.

Explainability in Federated and Edge AI Systems: As cybersecurity workloads shift toward federated learning and edge computing, explainability must follow. Designing lightweight, privacy-preserving XAI methods that function under resource constraints will be essential. Such techniques must also account for distributed data sources and heterogeneous devices, ensuring consistent interpretability across varied environments [27].

Cognitive Modeling and Adaptive Explanations: A one-size-fits-all explanation paradigm is insufficient. Future work must explore adaptive explanation systems that tailor outputs to users cognitive profiles, roles, and situational context. SOC analysts may require detailed causal chains, while CISOs might prefer risk summaries in natural language [28]. Incorporating user modeling and context-aware interfaces will significantly enhance interpretability and usability.

Integration with Governance, Ethics, and Auditing: Explainability will play a pivotal role in AI governance, especially in regulated sectors. Future frameworks must support auditable XAI pipelines, offering logs of model behavior, reasoning paths, and feedback mechanisms. Such capabilities are essential for compliance with emerging AI accountability policies like the EU AI Act and U.S. Executive Order on Safe AI [29].

8. Potential Uses

This scholarly article serves as a foundational resource for researchers, cybersecurity professionals, policymakers, and system architects seeking to integrate Explainable AI (XAI) into cyber defense strategies. It provides a conceptual and technical roadmap for enhancing threat detection and decision-making transparency in security operations centers (SOCs). By detailing XAI methods such as SHAP, LIME, and saliency maps across use cases like intrusion detection and malware classification, the article aids practitioners in selecting appropriate interpretability techniques tailored to operational needs.

The article can be used in graduate-level coursework on cybersecurity, AI ethics, or human-computer interaction, offering a multidisciplinary perspective that bridges machine learning with security engineering. Government agencies and regulatory bodies may also use the insights to inform guidelines and risk management frameworks that demand transparency and accountability in AI-driven systems.

This work can support vendors and developers of AI-based security tools in designing human-centric interfaces and adaptive explanation pipelines that align with compliance requirements such as the NIST AI RMF and EU AI Act. Overall, this article contributes to shaping the next generation of trustworthy, explainable, and resilient cyber defense systems.

9. Conclusion

As cyber threats grow in complexity and volume, the integration of Explainable AI (XAI) into cybersecurity has become imperative not just for improving detection capabilities, but for fostering transparency, accountability, and human trust. This article explored the critical role of interpretability in security contexts, reviewed state-of-the-art XAI techniques, and analyzed their application across key use cases such as intrusion detection, malware classification, and phishing detection. I proposed a modular, user-centric framework for embedding explainability into cyber defense architectures and identified key challenges such as scalability, cognitive load, and adversarial manipulation.

While XAI presents clear advantages, its effectiveness depends on alignment with operational workflows and user needs. Future progress requires collaborative research that spans technical, human, and policy domains. Tailored explanations, adaptive interfaces, and governance-ready designs will be essential to ensuring responsible and effective AI use in cybersecurity. XAI is not a luxury or an afterthought it is a fundamental enabler of trust and resilience in modern cyber defense. As AI continues to evolve, so too must my approaches to making its decisions understandable and actionable in the face of ever-changing digital threats.

References

[1] U. Fiore et al., "Using generative adversarial networks for improving classification effectiveness in credit card fraud detection," Information Sciences, vol. 479, pp. 448–455, 2019.

[2] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[3] J. Lin et al., "Explainable AI: A survey on techniques and challenges in cybersecurity," IEEE Access, vol. 10, pp. 9927–9946, 2022.

[4] S. Wachter, B. Mittelstadt, and L. Floridi, “Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation,” International Data Privacy Law, vol. 7, no. 2, pp. 76–99, 2017.

[5] National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework (AI RMF) 1.0,” NIST, Jan. 2023. [Online]. Available: [https://www.nist.gov/itl/ai-risk-management-framework]

[6] R. Sommer and V. Paxson, “Outside the closed world: On using machine learning for network intrusion detection,” in Proc. IEEE Symposium on Security and Privacy (SP), pp. 305–316, 2010.

[7] B. Biggio and F. Roli, “Wild patterns: Ten years after the rise of adversarial machine learning,” Pattern Recognition, vol. 84, pp. 317–331, 2018.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?” Explaining the predictions of any classifier,” in Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), pp. 1135–1144, 2016.

[9] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774, 2017.

[10] C. Wachter, S. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the GDPR,” Harvard Journal of Law & Technology, vol. 31, no. 2, pp. 841–887, 2018.

[11] H. Xu, C. Liu, and M. Zhang, “Explainable machine learning for intrusion detection: A case study with LIME and SHAP,” in Proc. 2021 IEEE Conf. on Dependable and Secure Computing (DSC), pp. 157–164, 2021.

[12] Y. Wang et al., “XID: Explainable intrusion detection using attention-based deep neural networks,” IEEE Access, vol. 9, pp. 34132–34145, 2021.

[13] F. K. Ghaffary and M. Abadi, “Visualizing deep learning decisions for malware detection using LRP and saliency maps,” in Proc. 2022 IEEE Int. Symp. on Technologies for Homeland Security (HST), pp. 1–7, 2022.

[14] D. Ribeiro, P. Cerqueira, and M. Gonçalves, “Explaining phishing detection with interpretable machine learning,” Computers & Security, vol. 116, 102638, 2022.

[15] H. Kim, A. Oh, and B. Kim, “On the interpretability of detection systems for cybersecurity,” IEEE Trans. on Information Forensics and Security, vol. 16, pp. 2460–2475, 2021.

[16] A. Kaur and N. Kumar, "Hybrid interpretable models for real-time threat detection in edge-cloud environments," Future Generation Computer Systems, vol. 135, pp. 171–182, 2023.

[17] Y. Wang et al., "Human-centric XAI interfaces for cybersecurity: Design principles and case studies," in Proc. IEEE VIS Workshop on Explainable AI (XAI), pp. 1–6, 2022.

[18] M. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," IEEE Trans. on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021.

[19] M. Arya, S. Chatterjee, and A. Joshi, "Evaluating XAI effectiveness in cybersecurity: A user-centered study," in Proc. ACM Conf. on Human Factors in Computing Systems (CHI), pp. 1–13, 2023.

[20] National Institute of Standards and Technology, "AI Risk Management Framework (AI RMF) 1.0," Jan. 2023. [Online]. Available: [https://www.nist.gov/itl/ai-risk-management-framework]

[21] A. Slack, J. Kim, and A. Moosavi-Dezfooli, "Scalable explainable AI for high-dimensional cybersecurity data," in Proc. IEEE Int. Conf. on Big Data Security on Cloud, pp. 47–56, 2022.

[22] Y. Zhang et al., "Manipulating explanations to fool deep learning detectors: Adversarial attacks on model interpretability," IEEE Trans. on Dependable and Secure Computing, early access, doi: 10.1109/TDSC.2023.3248795.

[23] M. Chandrasekaran and T. Nguyen, "Trust but verify: On the (in)stability of explainable AI methods in security," in Proc. ACM Workshop on Artificial Intelligence and Security (AISec), pp. 89–98, 2023.

[24] C. Abdul et al., "Cognitive load in security explainability: A usability study of XAI interfaces for analysts," ACM Transactions on Interactive Intelligent Systems, vol. 13, no. 1, pp. 1–25, 2024.

[25] S. Bhatt, A. Fix, and R. Liao, "Aligning explainability with risk: A policy perspective for cybersecurity AI," in Proc. IEEE Symposium on Security and Privacy Workshops (SPW), pp. 210–219, 2023.

[26] L. Li et al., "Human-AI collaboration for cyber threat hunting: Challenges and research directions," IEEE Trans. on Human-Machine Systems, vol. 54, no. 1, pp. 102–115, 2024.

[27] M. Lin, X. He, and A. Ghosh, "Federated explainable AI for distributed cyber defense," in Proc. IEEE Int. Conf. on Distributed Computing Systems (ICDCS), pp. 243–252, 2023.

[28] E. Ribeiro and S. Williams, "Adaptive XAI for security analysts: Contextualizing explanations in SOC environments," Journal of Cybersecurity and Privacy, vol. 5, no. 2, pp. 67–84, 2024.

[29] A. D. Narayanan and J. Tran, "Auditing AI: Legal and technical requirements for explainability in cyber systems," in Proc. IEEE Symposium on Security and Privacy Workshops (SPW), pp. 193–200, 2023.

Citation: Sri Ramya Deevi. (2025). Transforming Cyber Defense Through Explainable AI: Interpretability in Security Contexts. International Journal of Computer Engineering and Technology (IJCET), 16(4), 170–182.

Abstract Link: https://iaeme.com/Home/article_id/IJCET_16_04_012

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_16_ISSUE_4/IJCET_16_04_012.pdf

✉ **editor@iaeme.com**