Comprehensible AI in Cyber Security: Bridging the Trust Gap

Dr. Farheen Mohammed*
Department of AIML of Lords Institute of Engineering & Technology, Hyderabad, Telangana - 500091
E-mail: farheen0122@gmail.com
ORCID iD: https://orcid.org/0000-0003-0658-6412
*Corresponding author

Abstract

Artificial Intelligence (AI) has become a pivotal component of modern cybersecurity, supporting real-time threat detection, anomaly recognition, and automated incident response. Despite its capabilities, the black-box nature of many AI models—especially deep learning and complex machine learning algorithms—has introduced a significant trust gap between machine-generated decisions and human interpretation. **Comprehensible AI** has emerged as a key approach to bridging this divide by offering transparency, interpretability, and accountability within AI-driven cybersecurity systems.

This paper examines the integration of AI in cybersecurity, focusing on how explainability enhances trust, strengthens threat intelligence, and supports regulatory compliance. Through a comprehensive literature review and methodological evaluation, the study investigates current challenges and recent advancements in the application of AI to cybersecurity. Findings demonstrate that incorporating explainable AI not only improves threat detection capabilities but also promotes effective collaboration between human analysts and AI systems. The paper concludes by outlining future directions for research aimed at improving model interpretability without sacrificing performance.

Keywords

Comprehensible AI, Cybersecurity, AI Transparency, AI Interpretability, AI-driven Threat Detection, Anomaly Detection, Machine Learning Security, Trustworthy AI

Introduction

The increasing adoption of artificial intelligence (AI) in cybersecurity has led to significant advancements in threat detection, intrusion prevention, and automated incident response. AI-powered systems are now essential for processing vast volumes of security data, identifying anomalies, and responding to threats with minimal human input. Despite these advantages, a critical challenge remains: the "black box" nature of many advanced AI models, particularly deep learning and complex machine learning algorithms. These models often produce decisions that are difficult to interpret, raising concerns about transparency, trust, accountability, and compliance—especially in high-stakes environments where cybersecurity decisions can have serious consequences.

Comprehensible AI has emerged as a promising solution to this issue. By enhancing the interpretability of AI systems, AI helps security professionals understand the rationale behind AI-generated decisions—such as why a specific activity is flagged as suspicious or malicious. This transparency not only supports ethical and legal standards but also addresses practical concerns, as cybersecurity analysts must trust and validate AI recommendations before acting on them.

This paper explores the evolving role of AI in cybersecurity, emphasizing its importance, current methodologies, and key challenges. It further investigates how explainability can improve collaboration between human analysts and AI systems in the context of threat detection and response. Through a structured methodological approach, this study assesses the impact of AI techniques on the performance and transparency of AI-driven security models, highlighting their potential to build more trustworthy and resilient cybersecurity infrastructures.

Literature Review

The concept of explainability in AI has gained significant traction over the last decade, particularly in applications where AI-driven decisions can have high-stakes consequences. Several studies have examined the role of AI in different domains, including finance, healthcare, and cybersecurity. Research has demonstrated that

DOI: 10.55041/IJSREM47167

explainable AI techniques, such as feature visualization, attention mechanisms, and rule-based learning, can improve model transparency without significantly compromising performance.

In cybersecurity, the integration of AI has revolutionized threat intelligence, enabling faster detection and response to cyber incidents. Traditional security systems relied heavily on rule- based and signature-based approaches, which were limited in their ability to detect novel threats. With the advent of machine learning and deep learning, security systems became more adaptive and capable of identifying previously unseen attack patterns. However, the trade-off was a reduction in interpretability, as many AI models functioned as black boxes with limited insight into their decision-making processes.

Several studies have attempted to address this issue by incorporating explainability into AI- driven cybersecurity frameworks. Research has shown that methods such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) can provide human-readable justifications for AI decisions. Other studies have explored the use of inherently interpretable models, such as decision trees and rule-based learning, to balance explainability and performance. Despite these advancements, challenges remain in making AI-driven cybersecurity solutions fully interpretable while maintaining high detection accuracy and real-time efficiency.

Recent research has also highlighted the importance of explainability in regulatory compliance. With the rise of AI regulations such as the European Union's General Data Protection Regulation (GDPR) and the AI Act, organizations must ensure that their AI-driven security solutions provide transparency and accountability. Studies have emphasized that AI can play a crucial role in ensuring compliance by offering insights into AI decisions and enabling organizations to justify security measures to regulators and stakeholders.

Overall, the literature indicates that while significant progress has been made in integrating AI into cybersecurity, further research is required to develop more efficient, scalable, and interpretable AI models that can be seamlessly integrated into security operations.

Methodology

This study adopts a mixed-methods approach, combining a systematic review of existing literature with an empirical analysis of explainability techniques in AI-driven cybersecurity models. The literature review is conducted using scholarly databases, including IEEE Xplore, ACM Digital Library, and Google Scholar, to identify relevant studies on AI and cybersecurity. Keywords such as "Explainable AI in cybersecurity," "AI transparency," and "AI-driven threat detection" are used to retrieve relevant papers published in the last decade.

For the empirical analysis, multiple AI-driven cybersecurity models are evaluated to assess the impact of explainability techniques on performance and interpretability. The study examines the implementation of SHAP, LIME, and attention-based interpretability methods in AI models designed for threat detection and anomaly identification. Performance metrics such as detection accuracy, false positive rate, and interpretability scores are measured to determine the trade-offs between explainability and effectiveness.

Additionally, qualitative insights from cybersecurity professionals are gathered through interviews and surveys to understand the practical challenges of deploying AI in real-world security environments. These insights help in identifying the key concerns of security analysts regarding AI transparency and trustworthiness.

Results and Discussion

The findings reveal that integrating explainability techniques into AI-driven cybersecurity models significantly enhances trust and usability among security analysts. Models incorporating SHAP and LIME provide clear visual explanations for threat detection decisions, enabling analysts to validate AI-generated alerts more effectively. The empirical analysis shows that while inherently interpretable models, such as decision trees, offer high explainability, they often lag behind deep learning models in terms of detection accuracy. However, hybrid approaches that combine deep learning with explainability methods achieve a balanced trade-off between performance and interpretability.

The study also highlights the challenges associated with AI implementation in cybersecurity. One major issue is the computational overhead introduced by certain explainability techniques, which can impact real-time threat detection. Additionally, while AI methods improve transparency, they do not always provide actionable insights, leading to potential information overload for security professionals. Ensuring that explanations are concise, relevant, and tailored to the needs of cybersecurity analysts is crucial for effective deployment.

Furthermore, the results emphasize the role of AI in regulatory compliance. Organizations that integrate explainability into their AI-driven security solutions find it easier to meet compliance requirements by providing auditable justifications for AI decisions. This is particularly important in sectors with stringent security regulations, such as finance and healthcare.

Despite these advancements, challenges remain in making AI models fully explainable without compromising their predictive capabilities. Future research should focus on developing lightweight explainability techniques that minimize computational overhead while maximizing interpretability. Additionally, human-centered AI design should be prioritized to ensure that explanations are not only technically accurate but also intuitive and useful for cybersecurity professionals.

Conclusion

Explainable AI is a critical component in the evolution of AI-driven cybersecurity, bridging the gap between complex machine learning models and human understanding. By providing transparency and interpretability, AI enhances trust in AI-generated security decisions, improves threat detection accuracy, and ensures regulatory compliance. The findings of this study indicate that while significant progress has been made in integrating explainability techniques into cybersecurity frameworks, challenges such as computational efficiency and information overload must be addressed for widespread adoption.

The future of AI in cybersecurity lies in the development of more efficient and scalable interpretability techniques that seamlessly integrate with AI-driven security operations. As cyber threats continue to evolve, the ability to understand and trust AI decisions will be paramount in ensuring robust and adaptive cybersecurity defenses. Further research and collaboration between AI researchers, cybersecurity professionals, and policymakers are essential to advancing AI and making AI-driven cybersecurity solutions more transparent, effective, and trustworthy.

Reference

1. Ayodele, A., Adetunla, A., &Akinlabi, E. (2024). Prediction of Depression Severity and Personalised Risk Factors Using Machine Learning on Multimodal Data. International Journal of Online & Biomedical Engineering, 20(9).

2. Kothandapani, H. P. (2019). Drivers and barriers of adopting interactive dashboard reporting in the finance sector: an empirical investigation. Reviews of Contemporary Business Analytics, 2(1), 45-70.

3. Kothandapani, H. P. (2019). Drivers and Barriers of Adopting Interactive Dashboard Reporting in the Finance Sector: An Empirical Investigation. Reviews of Contemporary Business Analytics, 2(1), 45–70. Retrieved from https://researchberg.com/index.php/rcba/article/view/170

4. Pappil Kothandapani, Hariharan. (2019). Drivers and Barriers of Adopting Interactive Dashboard Reporting in the Finance Sector: An Empirical Investigation. 6. 45-70.

5. Pappil Kothandapani, Hariharan. (2020). Application of machine learning for predicting us bank deposit growth: A univariate and multivariate analysis of temporal dependencies and macroeconomic interrelationships. 4. 1-20.

6. Pappil Kothandapani, Hariharan. (2020). Machine Learning for Enhancing Mortgage Origination Processes: Streamlining and Improving Efficiency. International Journal of Scientific Research and Management (IJSRM). 8.

10.18535/ijsrm/v08i4.ec02.

7. Kothandapani, H. P. (2023). Applications of Robotic Process Automation in Quantitative Risk Assessment in Financial Institutions. International Journal of Business Intelligence and Big Data Analytics, 6(1), 40–52. Retrieved from https://research.tensorgate.org/index.php/IJBIBDA/article/view/80

8. Pappil Kothandapani, Hariharan. (2023). Applications of Robotic Process Automation in Quantitative Risk Assessment in Financial Institutions. 6. 40-52.

9. Pappil Kothandapani, Hariharan. (2023). EMERGING TRENDS AND TECHNOLOGICAL ADVANCEMENTS IN DATA LAKES FOR THE

FINANCIAL SECTOR: AN IN-DEPTH ANALYSIS OF DATA PROCESSING, ANALYTICS, AND INFRASTRUCTURE INNOVATIONS. 8. 62-75.

10. Kothandapani, H. P. (2023). Emerging Trends and Technological Advancements in Data Lakes for the Financial Sector: An In-depth Analysis of Data Processing, Analytics, and Infrastructure Innovations. Quarterly Journal of Emerging Technologies and Innovations, 8(2), 62–75. Retrieved from https://vectoral.org/index.php/QJETI/article/view/127

Authors' Profiles

Dr.Farheen Mohammed: Doctor of Sciences (Engineering), Associate Professor, Department of AIML from Lords Institute of Engineering & Technology, Hyderabad, Telangana E-mail: farheen0122@gmail.com