IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.67

Volume 5, Issue 1, March 2025

Explainable AI (XAI) for Cyber Defense: Enhancing Transparency and Trust in AI-Driven Security Solutions

Giriraj Agarwal

Sr. Manager - Projects – Cognizant,

https://orcid.org/0009-0006-1042-6568

Abstract*:* *The increasing reliance on Artificial Intelligence (AI) for cyber defense has led to the development of advanced detection systems capable of identifying complex threats in real time. However, many of these systems function as "black boxes," offering little insight into their decision-making processes. Explainable AI (XAI) seeks to address this limitation by providing transparent, interpretable outputs that empower security analysts to understand, validate, and trust AI-generated decisions. This paper reviews the state-of-the-art in XAI methodologies applied to cybersecurity, discusses key challenges such as balancing interpretability with model performance, and proposes a hybrid framework that integrates explainability into AI-based cyber defense systems. Through simulation-based benchmarking and case studies, we illustrate how XAI can enhance threat detection accuracy, streamline incident response, and ultimately foster greater trust in automated security solutions. Future research directions include the development of standardized XAI metrics for cybersecurity and the integration of real-time explanation engines into operational environments.*

Keywords: Explainable AI, Cyber Défense, Cybersecurity, Interpretability, Transparency, Trust, Hybrid AI Models

DOI: 10.48175/IJARSCT-23624