RADemics

Explainable AI (XAI) for Cybersecurity Decision-Making Using SHAP and LIME for Transparent Threat Detection

Shantha Visalakshi Upendran, Karthiyayini.S, Dinesh Vijay Jamthe

ETHIRAJ COLLEGE FOR WOMEN (AUTONOMOUS), VEL TECH RANGARAJAN DR. SAGUNTHALA R&D INSTITUTE OF SCIENCE AND TECHNOLOGY, PRIYADARSHINI BHAGWATI COLLEGE OF ENGINEERING

Explainable AI (XAI) for Cybersecurity Decision-Making Using SHAP and LIME for Transparent Threat Detection

[1]Shantha Visalakshi Upendran, Associate Professor, MCA, Ethiraj College for Women (Autonomous), Chennai. profdrusv@gmail.com

[2]Karthiyayini.S, Assistant Professor Senior Grade, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Tamil Nadu, Chennai. Mail id: drkarthiyayinis@veltech.edu.in

[3]Dinesh Vijay Jamthe, Assistant Professor, Computer Science & Engineering, Priyadarshini Bhagwati College of Engineering, Harpur Nagar, Nagpur (MH). dineshjamthe@gmail.com

Abstract

The increasing complexity and sophistication of cyber threats have necessitated the integration of Explainable Artificial Intelligence (XAI) into cybersecurity frameworks to enhance transparency, trust, and decision-making. Traditional black-box machine learning models, despite their high accuracy, pose significant challenges in understanding threat detection mechanisms, leading to reduced interpretability and limited adoption in critical security applications. This book chapter explores the role of XAI techniques, specifically Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), in improving the explainability of AI-driven cyber defense systems. A detailed analysis of computational efficiency, real-time applicability, and scalability challenges associated with SHAP and LIME in large-scale cybersecurity environments is provided., the chapter introduces hardware-accelerated approaches, such as FPGA-based optimization, to mitigate computational overhead while ensuring rapid and interpretable threat detection. reinforcement learning-based optimization for explainability is examined to enhance adaptive security mechanisms in dynamic threat landscapes. The integration of XAI-driven security information and event management (SIEM) systems is also discussed to bridge the gap between automated cyber threat detection and human-centric decision-making. This chapter provides a comprehensive exploration of state-of-the-art methodologies, challenges, and future research directions in the domain of XAI for cybersecurity, with a focus on balancing detection accuracy, computational efficiency, and interpretability.

Keywords: Explainable AI (XAI), SHAP, LIME, Cybersecurity, FPGA Acceleration, Reinforcement Learning.

Introduction

The increasing sophistication of cyber threats, including advanced persistent threats (APTs), ransomware, and polymorphic malware, has necessitated the deployment of AI-driven cybersecurity frameworks. Traditional rule-based security systems struggle to adapt to the dynamic nature of cyberattacks, making machine learning (ML) and deep learning (DL) essential

for threat detection, anomaly identification, and risk assessment. However, a major challenge associated with these AI-driven security mechanisms is their black-box nature, which limits interpretability and transparency. Security professionals require clear justifications for AI-driven decisions, especially in high-stakes environments where false positives and false negatives can lead to severe financial and operational consequences. The lack of explainability in cybersecurity AI models raises concerns regarding trust, regulatory compliance, and human-in-the-loop decision-making. To address this, Explainable AI (XAI) has emerged as a critical field, focusing on making AI predictions more interpretable, transparent, and actionable for security analysts.

Explainability in AI-driven cybersecurity is essential for improving threat response, incident investigation, and compliance with regulatory frameworks such as the General Data Protection Regulation (GDPR) and the National Institute of Standards and Technology (NIST) cybersecurity guidelines. Among the most effective XAI techniques are Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), which provide insights into how AI models classify threats and detect anomalies. These methods help analysts understand feature importance, enabling them to verify AI decisions and adjust security policies accordingly. However, the implementation of SHAP and LIME in cybersecurity presents computational challenges, particularly in real-time applications where rapid detection and response are crucial. Addressing these limitations requires optimization techniques that enhance the efficiency and scalability of XAI while preserving interpretability.

One of the primary bottlenecks in integrating XAI into cybersecurity systems is the computational overhead associated with explainability techniques. SHAP, for instance, computes feature importance by considering all possible coalitions of input features, making it computationally expensive for high-dimensional datasets commonly encountered in network security and intrusion detection systems (IDS). LIME, on the other hand, perturbs input data and builds local surrogate models to approximate black-box AI decisions, which can be time-intensive for large-scale security logs and streaming data. The need for real-time, low-latency explainability solutions is paramount in cybersecurity operations, where delays in decision-making can expose systems to persistent and evolving threats. Therefore, research into optimizing XAI techniques through algorithmic improvements, approximation methods, and parallel computing is essential for their practical deployment in cybersecurity environments.