RADemics

Machine Learning Fundamentals for Intrusion Detection and Threat Classification

G.Sangeetha, S.Kalaivani, Vishal Vijay Chahare
SRM Valliammai Engineering College, B.S. Abdur Rahman Crescent Institute of science and technology, M.S.P. Mandal's Deogiri Institute Of Engineering & Management Studies

Machine Learning Fundamentals for Intrusion Detection and Threat Classification

[1]G. Sangeetha, Assistant Professor, Computer Science and Engineering, SRM Valliammai Engineering College, Kattankulathur, sangeethag.cse@srmugavalliammai.ac.in.

[2]S. Kalaivani, Assistant Professor, Computer Applications, B.S. Abdur Rahman Crescent Institute of science and technology, Vandalur, kalaivani@crecent.education,

[3]Vishal Vijay Chahare, Assistant ProfessorMechanical Engineering, M.S.P. Mandal's Deogiri Institute Of Engineering & Management Studies, Railway Station Road, CH. Sambhajinagar, vishalvijaychahare@dietms.org,

Abstract

Machine learning (ML) has revolutionized the field of intrusion detection systems (IDS), offering enhanced capabilities for detecting and mitigating cyber threats in real-time. However, the adoption of ML-driven IDS models is often hindered by concerns over their explainability, trustworthiness, and robustness, which are crucial for their effective deployment in security-critical environments. This chapter explores the fundamental challenges and strategies for ensuring the reliability of ML-based IDS models, focusing on techniques for enhancing explainability, overcoming user skepticism, and ensuring long-term model robustness. Key areas addressed include the black-box nature of deep learning models, the scalability of explainability techniques for large-scale networks, and the critical role of certification and auditing in maintaining system integrity. Additionally, the chapter highlights emerging trends in the field, such as the integration of adversarial defense mechanisms and the need for transparent, explainable AI solutions in cybersecurity. By examining these topics, this chapter aims to provide a comprehensive framework for developing more effective, trustworthy, and interpretable ML-driven IDS solutions, paving the way for their broader acceptance and application in modern cybersecurity practices.

Keywords: Intrusion Detection Systems, Machine Learning, Explainability, Robustness, Trustworthiness, Certification

Introduction

Machine learning (ML) has made significant strides in revolutionizing intrusion detection systems (IDS), offering a more adaptive, efficient, and scalable solution compared to traditional rule-based methods [1]. Traditional IDS models, primarily signature-based, are often limited by their reliance on predefined patterns and signatures, which struggle to detect novel, zero-day attacks [2]. The increasing sophistication of cyberattacks, coupled with the dynamic nature of network traffic, necessitates more advanced systems capable of identifying complex, previously unseen threats [3]. ML-powered IDS models leverage vast amounts of data to identify patterns, anomalies, and malicious activities, providing the ability to detect threats in real-time [4]. While these systems offer enhanced capabilities, they come with their own set of challenges, particularly in terms of explainability, trust, and robustness, which are essential for their adoption in high-security environments [5].

The "black-box" nature of many ML models poses a significant challenge to their integration into intrusion detection systems [6]. Deep learning models, for example, often involve complex architectures with multiple layers of neurons, making it difficult to trace the decision-making process [7]. This lack of transparency can lead to a lack of trust in the model's outputs, especially when these models are tasked with identifying potential threats in critical infrastructure [8]. Cybersecurity professionals and system administrators need to understand the reasoning behind an IDS's decision to flag network traffic as benign or malicious [9]. Without this understanding, the system's effectiveness is undermined, as users may hesitate to trust the model's predictions, opting instead for manual validation. Thus, achieving transparency and interpretability in ML-based IDS is crucial for enhancing user confidence and ensuring widespread adoption [10].

Scalability is another critical challenge when it comes to implementing machine learning in IDS [11]. As the volume of data in modern network infrastructures grows exponentially, so does the complexity of processing and analyzing that data [12]. ML models in IDS are expected to handle large-scale environments with vast amounts of network traffic, which introduces significant computational challenges [13]. The need for real-time decision-making, particularly in fast-evolving cyber threats, further amplifies these issues. Ensuring that explainability techniques can scale with the increasing volume and complexity of data is vital for the success of ML-based IDS [14]. While techniques such as LIME and SHAP offer valuable insights into model decisions, they can become computationally expensive when applied to large datasets. Consequently, finding scalable, efficient solutions for maintaining model transparency while processing massive amounts of data remains a significant challenge that must be addressed to ensure the real-time viability of ML-driven IDS [15].