International Journal of Information Security and Privacy
Volume 20 • Issue 1 • January-December 2026

Machine Learning and Explainable Artificial Intelligence for Network Intrusion Detection

Ibidun Christiana Obagbuwa
http://orcid.org/0000-0002-7965-2823
Walter Sisulu University, South Africa

Madison N. Ngafeeson
http://orcid.org/0000-0002-7805-3968
Palm Beach Atlantic University, USA

Oluwatimileyin Favour Obagbuwa
Business Systems Group, South Africa

Anthony Tsetse
Northern Kentucky University, USA
Accepted: February 18th, 2026

ABSTRACT

The growing sophistication of cyber threats demands adaptive security mechanisms beyond traditional Intrusion Detection Systems (IDS). This paper explores integrating Machine Learning (ML) and Explainable Artificial Intelligence (XAI) to enhance Network Intrusion Detection Systems (NIDS). Using the CICIDS2017 dataset, the authors evaluate ML models including Convolutional Neural Networks (CNN), Random Forest, and XGBoost, balancing detection performance with interpretability. Results show XGBoost achieves the highest accuracy with minimal misclassifications, underscoring its robustness for intrusion detection. To address the black-box challenge of deep learning, SHapley Additive exPlanations (SHAP) is applied to interpret predictions. Key features such as Destination Port, Flow Duration, and Packet Length emerged as critical, improving trust, reducing false positives, and aiding investigation. The authors highlight the necessity of coupling high-performing ML with XAI frameworks for transparency. Finally, challenges in scalability, robustness, and dataset generalizability are discussed.

KEYWORDS

Cybersecurity, SHAP Analysis, Machine Learning, Network Traffic, Anomaly Detection, Artificial Intelligence, Explainable Artificial Intelligence, and Intrusion Detection

INTRODUCTION

The rapid proliferation of digital technologies has led to an exponential increase in the volume and complexity of network traffic, thereby expanding the attack surface for cyber threats. Traditional security measures, such as signature-based Intrusion Detection Systems (IDS), are increasingly inadequate in detecting novel or sophisticated attacks due to their reliance on predefined patterns and rules (Abdullahi et al., 2022; El Houda, Brik, & Senouci, 2022). This limitation has spurred the adoption of Machine Learning (ML) techniques, which offer the capability to identify previously unknown threats by learning from data patterns (Ali et al., 2024).

DOI: 10.4018/IJISP.402900

Recent advancements in ML have significantly enhanced the efficacy of IDS. Deep learning models, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have demonstrated superior performance in capturing complex patterns in network traffic, leading to improved detection rates for various types of intrusions (Zhao, Fok, & Thing, 2024). Additionally, the integration of Generative Adversarial Networks (GANs) has addressed challenges related to data imbalance by generating synthetic samples that augment training datasets, thereby enhancing model robustness and accuracy (Zhao, Fok, & Thing, 2024).

However, the deployment of ML-based IDS introduces new challenges, primarily concerning the interpretability of model decisions. The "black box" nature of many ML models complicates the understanding of how decisions are made, which is critical in cybersecurity contexts where transparency is essential for trust and accountability (Ribeiro, Singh, & Guestrin, 2016). To address this issue, Explainable Artificial Intelligence (XAI) has emerged as a pivotal field, focusing on developing models and techniques that provide human-understandable explanations for their predictions (Mohale & Obagbuwa, 2025).

XAI methods, such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), have been applied to IDS to elucidate the reasoning behind model decisions (Mohale & Obagbuwa, 2025; Ribeiro, Singh, & Guestrin, 2016). These techniques not only enhance the transparency of ML models but also empower security analysts to validate and trust automated alerts, facilitating more effective response strategies (El Houda, Brik, & Senouci, 2022).

Despite the promising developments, several challenges persist in the integration of ML and XAI into IDS. Issues such as the scalability of explainability methods, the trade-off between model accuracy and interpretability, and the need for real-time analysis in dynamic network environments require ongoing research and innovation (Abdullahi et al., 2022; Ali et al., 2024).

This paper aims to explore the intersection of ML and XAI in the realm of network intrusion detection. We review recent advancements in ML algorithms applied to IDS, examine the role of XAI in enhancing model interpretability, and discuss the challenges and future directions in this interdisciplinary domain. By synthesizing current research, this paper seeks to provide a comprehensive understanding of how ML and XAI can collaboratively advance the field of network intrusion detection, ensuring both effective threat detection and transparent decision-making processes.

LITERATURE REVIEW

The field of network intrusion detection has witnessed a paradigm shift with the advent of machine learning (ML) and, more recently, explainable artificial intelligence (XAI). This section critically reviews key developments in ML-based intrusion detection, followed by the emergence of explainability tools to enhance trust, interpretability, and transparency in cybersecurity applications. This section synthesizes prior work by explicitly outlining what has been achieved, where interpretability limitations persist, and how the present study addresses these gaps.

Advances in ML-Based Intrusion Detection

Traditional and ML-Based NIDS

Early NIDS relied heavily on signature-based detection tools such as Snort and Suricata, which effectively identify known attacks but fail against zero-day threats due to their dependence on predefined rules (Axelsson, 2000). To overcome this limitation, researchers shifted toward anomaly-based detection using ML models capable of learning behavioural patterns in network traffic.

Supervised and unsupervised ML algorithms including Decision Trees, SVM, Naive Bayes, k-NN, Random Forest, and Gradient Boosting have demonstrated strong performance in detecting a variety of intrusions (Sommer & Paxson, 2010; Tsai et al., 2009). Deep learning approaches such as CNNs, RNNs, LSTMs, and DBNs further improved detection accuracy by capturing complex temporal and