East African Scholars Journal of Engineering and Computer Sciences
Abbreviated Key Title: East African Scholars J Eng Comput Sci
ISSN: 2617-4480 (Print) & ISSN: 2663-0346 (Online)
Published By East African Scholars Publisher, Kenya
Volume-9 | Issue-2 | Mar-Apr-2026 | DOI: https://doi.org/10.36349/easjecs.2026.v09i02.001

OPEN ACCESS

Original Research Article

Demystifying the "Black Box" of Deep Learning in Cybersecurity: Integrating Explainable AI (XAI) to Augment the Reliability of Intrusion Detection Systems

Nguyễn Kiên Trung[1]*, Ngô Văn Nam[1], Nguyễn Trung Kiên[1]
[1]Hung Vuong University – Phu Tho Province

Article History
Received: 30.01.2026
Accepted: 26.03.2026
Published: 30.03.2026

Journal homepage:
https://www.easpublisher.com

Quick Response Code

Abstract: The rapid and continuous rise of sophisticated cyberattacks has driven the adoption of Deep Learning models in Intrusion Detection Systems (IDS). While these models deliver superior detection performance compared to traditional approaches, their "black-box" nature poses a significant barrier to real-world deployment. Network administrators often struggle to trust system-generated alerts when they cannot understand the reasoning behind them. This paper proposes a comprehensive solution to address this challenge by integrating Explainable Artificial Intelligence (XAI) techniques into IDS. Specifically, we apply SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) to Deep Neural Network (DNN) models trained on benchmark datasets, namely NSL-KDD and UNSW-NB15. The results demonstrate not only the model's high effectiveness in detecting various types of cyberattacks but also its ability to provide detailed explanations at both global and local levels. These insights enable network administrators to better analyze the root causes of alerts, thereby improving the reliability and transparency of cybersecurity defense systems.

Keywords: Intrusion Detection Systems (IDS), Deep Learning; Explainable AI (XAI).

1. INTRODUCTION

In the era of digital transformation and the pervasive expansion of the Internet of Things (IoT), cybersecurity has emerged as one of the most critical and pressing challenges confronting organizations and governments alike. Cyber threats are becoming increasingly sophisticated, heterogeneous, and unpredictable, thereby necessitating the continual advancement of defensive mechanisms. Intrusion Detection Systems (IDS) function as a pivotal line of defense, systematically monitoring network traffic to identify anomalous activities or violations of established security policies [1].

Conventional IDS approaches, primarily signature-based and rule-based, have been extensively adopted in practice. However, these methods exhibit significant limitations when confronted with zero-day exploits and emerging attack variants. To overcome these shortcomings, the research community has progressively shifted toward Machine Learning (ML) and, more prominently, Deep Learning (DL)–driven paradigms. Advanced Deep Learning architectures, including Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN), possess the capability to autonomously extract intricate feature representations from large-scale network traffic data, thereby achieving superior detection performance [2].

Notwithstanding these notable advancements, the real-world deployment of Deep Learning–based IDS is impeded by a fundamental challenge: the "black-box" phenomenon. Characterized by high-dimensional parameter spaces, Deep Learning models operate as opaque systems whose internal decision-making processes remain largely inscrutable to human observers. Consequently, when an IDS generates an alert, it often fails to provide a comprehensible rationale for classifying specific traffic as malicious. This lack of interpretability erodes the confidence of security analysts, particularly in mission-critical scenarios that demand rapid and precise decision-making [3]. Moreover, the absence of transparency significantly complicates root

***Corresponding Author:** Nguyễn Kiên Trung
Hung Vuong University – Phu Tho Province

cause analysis, thereby limiting the system's practical utility.

To address this limitation, Explainable Artificial Intelligence (XAI) has emerged as a paradigm-shifting research frontier. XAI encompasses a comprehensive suite of methodologies aimed at enhancing the interpretability and transparency of complex AI models [4]. Within the cybersecurity domain, XAI not only elucidates the underlying rationale behind model predictions but also quantifies the contribution of individual network features to those decisions. This capability equips security practitioners with actionable insights, enabling them to validate system outputs, identify latent model vulnerabilities, and implement informed mitigation strategies.

In this study, we present a rigorous investigation into the integration of XAI within IDS frameworks. We develop a DNN-based classification model for detecting cyberattacks and incorporate two state-of-the-art XAI techniques, SHAP and LIME, to interpret its predictive behavior. Experimental evaluations are conducted on widely recognized benchmark datasets, namely NSL-KDD and UNSW-NB15. The central objective of this work is to demonstrate that the incorporation of XAI not only preserves the detection efficacy of IDS but also substantially enhances system trustworthiness, interpretability, and analytical utility in cybersecurity operations.

2. LITERATURE REVIEW

The convergence of Deep Learning, cybersecurity, and Explainable Artificial Intelligence (XAI) has garnered substantial attention from the academic community in recent years. This section surveys seminal studies on the application of Deep Learning in Intrusion Detection Systems (IDS) and the evolution of XAI methodologies.

2.1. Deep Learning in Intrusion Detection Systems

The adoption of Deep Learning for IDS has consistently demonstrated superiority over traditional Machine Learning algorithms such as Support Vector Machines (SVM) and Random Forests. Deep neural architectures are particularly adept at learning complex, non-linear feature representations from raw network data. Numerous studies have successfully leveraged Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) models to process temporal network traffic data, achieving high accuracy on benchmark datasets such as KDD99 and NSL-KDD [2].

However, the majority of these studies predominantly emphasize optimizing performance metrics—such as Accuracy and F1-Score—while largely neglecting the critical aspect of model interpretability.

2.2. The Evolution of Explainable Artificial Intelligence (XAI)

XAI has witnessed rapid advancement and widespread adoption across high-stakes domains, including healthcare, finance, and autonomous driving, where erroneous decisions may lead to severe consequences [5]. XAI techniques are broadly categorized into two principal classes: intrinsic methods and post-hoc methods.

Intrinsic methods rely on inherently interpretable models, such as Decision Trees and Linear Regression. While these approaches offer transparency, they often fall short in achieving competitive performance on complex, high-dimensional datasets. In contrast, post-hoc methods are applied after the training of a "black-box" model to elucidate its behavior.

Among the most prominent and powerful post-hoc techniques are LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations). LIME operates by constructing a locally faithful, interpretable surrogate model—typically linear—around a specific prediction instance, thereby approximating the behavior of the complex model in that locality [6]. Conversely, SHAP is grounded in cooperative game theory, leveraging Shapley values to systematically quantify the contribution of each feature to the final prediction. SHAP offers a theoretically rigorous and unified framework for both local and global interpretability [7].

2.3. XAI in Cybersecurity

Despite its maturity in several domains, the application of XAI in cybersecurity—particularly within IDS—remains in its nascent stages. Recent studies have begun to explore the potential of XAI in interpreting models for malware detection and network anomaly detection [8]. Nevertheless, there remains a paucity of comprehensive empirical evaluations that apply state-of-the-art XAI techniques to diverse and complex IDS datasets such as UNSW-NB15.

Furthermore, translating quantitative outputs from XAI methods, such as SHAP values or LIME explanations, into actionable insights for network administrators continues to pose a significant challenge. This paper seeks to bridge this gap by proposing a structured framework for integrating and interpreting XAI outputs within the context of root cause analysis for cyberattacks, thereby enhancing both the practical utility and trustworthiness of IDS.

3. METHODOLOGY

To mitigate the "black-box" limitation in Intrusion Detection Systems (IDS), we propose a tightly integrated architectural framework that synergizes a high-capacity Deep Learning model with state-of-the-art Explainable Artificial Intelligence (XAI) techniques. This section delineates the core components of the

proposed system, including dataset selection, model architecture, and interpretability methodologies.

3.1. Proposed System Architecture

The proposed framework is structured into four fundamental stages:

- Data Preprocessing: Raw network traffic data undergoes rigorous cleansing, normalization, and encoding to ensure compatibility with neural network architectures.
- Model Training: A Deep Neural Network (DNN) is trained to perform fine-grained classification of network flows, distinguishing benign traffic from multiple categories of cyberattacks.
- XAI Integration: Advanced interpretability techniques, namely SHAP and LIME, are employed to extract explanatory insights from the trained model's predictions.
- Explanation Interface: The derived explanations are systematically visualized and presented through an interpretable interface, enabling network administrators to conduct efficient root cause analysis.

3.2. Datasets

To ensure robustness, objectivity, and generalizability, we utilize two extensively validated benchmark datasets in IDS research:

- NSL-KDD: An improved variant of the canonical KDD99 dataset, NSL-KDD mitigates issues related to redundant and biased records, thereby facilitating more equitable model evaluation. It comprises 41 features characterizing network connections and categorizes traffic into normal behavior and four principal attack classes: DoS (Denial of Service), Probe, R2L (Remote to Local), and U2R (User to Root) [9].
- UNSW-NB15: Curated by the University of New South Wales, this dataset encapsulates contemporary and heterogeneous network traffic patterns. It consists of 49 features extracted from 100GB of raw network data, encompassing benign traffic alongside nine modern attack categories, including Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms [10]. Its diversity renders it particularly suitable for evaluating both Deep Learning and XAI methodologies under realistic conditions.

3.3. Deep Learning Model

We design a multi-layer Deep Neural Network (DNN), specifically a Multi-Layer Perceptron (MLP), tailored for multi-class classification tasks. The architectural configuration includes:

- Input Layer: The dimensionality of this layer corresponds to the feature space of the dataset following one-hot encoding of categorical variables.
- Hidden Layers: Three fully connected layers with progressively diminishing neuron counts (e.g., 128, 64, 32) are employed to capture hierarchical and abstract feature representations. The ReLU (Rectified Linear Unit) activation function is utilized to alleviate the vanishing gradient problem, while Dropout regularization is systematically incorporated to mitigate overfitting.
- Output Layer: A Softmax activation function is applied to generate probabilistic class distributions across benign and attack categories. Model optimization is conducted using the Categorical Cross-Entropy loss function in conjunction with the Adam optimizer.

3.4. Applied XAI Techniques

To enhance the interpretability of the DNN, we adopt two model-agnostic, post-hoc XAI methodologies:

SHAP (SHapley Additive exPlanations): SHAP conceptualizes the explanation process through the formalism of cooperative game theory, wherein each input feature is treated as a "player" contributing to the aggregate "payout" (i.e., the model's prediction). Shapley values are computed to quantify the average marginal contribution of each feature across all possible feature coalitions. This mathematically rigorous framework guarantees properties such as fairness and consistency, thereby enabling robust estimation of feature importance at both global (model-level) and local (instance-level) perspectives [7].

LIME (Local Interpretable Model-Agnostic Explanations): In contrast, LIME emphasizes localized interpretability. For a given prediction instance (e.g., a suspicious network packet), LIME synthesizes perturbed samples within the vicinity of the original data point and evaluates the corresponding responses of the black-box model. Subsequently, it fits an interpretable surrogate model—typically a weighted linear model—where weights are assigned based on proximity to the original instance. This localized approximation faithfully captures the behavior of the complex model in the neighborhood of the instance, thereby elucidating the most influential features driving the prediction [6].

4. Experiments and Evaluation

This section presents the experimental results, including the performance of the DNN model and detailed analyses based on interpretability results from SHAP and LIME.

4.1. Experimental Setup

The experiments were conducted using Python, leveraging the TensorFlow/Keras library to construct the DNN model and the shap and lime libraries for XAI analysis. The datasets were split into 80% training and 20% testing sets. Standard evaluation metrics—

Accuracy, Precision, Recall, and F1-Score—were employed to assess the classification performance.

4.2. Performance of the Deep Learning Model

Table 1: Summary of DNN Model Performance on Two Datasets

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
NSL-KDD	98.45	98.21	98.50	98.35
UNSW-NB15	95.12	94.85	95.30	95.07

Table 1: Classification performance of the DNN model on NSL-KDD and UNSW-NB15 datasets.

The results indicate that the DNN model achieves exceptionally high performance, particularly on the NSL-KDD dataset. Although UNSW-NB15 is more complex, encompassing a broader range of contemporary attack types, the model still maintains an accuracy exceeding 95%. This underscores the effectiveness of Deep Learning in detecting network threats. However, as previously discussed, these performance metrics alone do not inspire confidence without accompanying interpretability.

4.3. Explanation Analysis with XAI

This section represents the core of the study, where XAI is leveraged to interpret and elucidate the decisions of the DNN model.

4.3.1. Global Explanations with SHAP

Global explanations provide network administrators with a holistic understanding of the model's decision-making: which features the model relies on to distinguish between benign traffic and attacks. By computing the average SHAP values across all samples in the test set, we can rank features according to their overall importance.

For the NSL-KDD dataset, SHAP analysis indicates that features such as src_bytes (bytes sent from source to destination), dst_bytes (bytes sent from destination to source), and flag (connection status) are the most influential. This aligns with domain knowledge: DoS attacks often exhibit anomalous data volume patterns, while Probe attacks tend to trigger abnormal connection flags (e.g., incomplete connections).

On the UNSW-NB15 dataset, features related to packet lifetimes (sttl, dttl) and payload sizes (sload, dload) emerge as the most critical. This reflects the nature of modern attacks in UNSW-NB15, where adversaries manipulate protocol parameters more subtly.

4.3.2. Local Explanations and Root Cause Analysis

Local explanations are particularly valuable for network administrators when confronted with a specific alert. Suppose the IDS issues a DoS alert: the administrator needs to understand precisely why this packet was flagged.

Using LIME for a specific DoS instance in NSL-KDD, the local explanation may show: the model predicts a DoS attack with 99% probability. The features most strongly driving this decision are count > 250 (excessive number of connections to the same server within 2 seconds) and serror_rate > 0.8 (high SYN error rate). Conversely, duration = 0 (short connection time) slightly reduces the predicted probability but does not alter the final classification.

This information enables the administrator not only to recognize an ongoing DoS attack but also to understand its mechanics: the attacker is sending a high volume of connection requests (SYN flood) in a short timeframe. This constitutes XAI-assisted root cause analysis. Rather than sifting through thousands of log entries, administrators are immediately provided with actionable evidence, facilitating confident decisions such as blocking the offending IP or reconfiguring firewalls promptly.

The combination of SHAP (for global model behavior insight) and LIME (for fine-grained instance-level analysis) creates an IDS that is not only intelligent but also transparent and trustworthy.

5. DISCUSSION

Integrating XAI into Deep Learning–based IDS offers substantial benefits but also presents several challenges for future research.

5.1. Benefits of XAI for Network Administrators

Enhanced Trust Calibration:

By providing clear and comprehensible explanations, XAI helps administrators overcome skepticism toward black-box models. When they can verify that decisions are based on meaningful network features rather than noise, confidence in deploying automated defense mechanisms significantly increases.

Support for Root Cause Analysis:

As demonstrated in the experiments, XAI transforms an alert into a detailed analytical report. Accurate identification of anomalous features shortens incident response time, allowing security experts to quickly isolate threats and patch vulnerabilities.

Bias Detection and Mitigation:

Global SHAP analysis can reveal if the model is relying on shortcuts or biased by imbalanced training data. For instance, if the model overly depends on source IP addresses rather than packet behavior, administrators can retrain the model with more diverse data.

5.2. Challenges and Limitations

Computational Overhead:

XAI algorithms, particularly SHAP, are computationally intensive. Exact Shapley value computation has exponential complexity. Although approximations (e.g., KernelSHAP, DeepSHAP) exist, real-time application in high-speed networks (Gigabit/s) remains a significant technical challenge.

Accuracy-Interpretability Trade-off:

While post-hoc methods like LIME and SHAP do not modify the original model, their explanations may not always fully capture the model's actual behavior. Ensuring high fidelity of XAI explanations remains an ongoing research topic.

Vulnerability to Adversarial Attacks:

Recent studies indicate that XAI algorithms themselves can be targeted. Adversaries may craft examples that mislead both the classifier and the explanation output, potentially steering network administrators toward incorrect conclusions [11].

6. CONCLUSION AND FUTURE DIRECTIONS

This paper presents a comprehensive study on applying Explainable AI (XAI) to address the black-box nature of Deep Learning models in Intrusion Detection Systems (IDS). By integrating SHAP and LIME into a Deep Neural Network (DNN) framework and evaluating on standard datasets (NSL-KDD, UNSW-NB15), we demonstrate that XAI not only preserves high detection performance but also provides deep insights at both global and local levels. These explanations are crucial for building administrator trust and facilitating effective root cause analysis of security incidents.

Looking forward, intelligent cybersecurity systems must prioritize not only accuracy but also transparency and interpretability. Future research will focus on optimizing the computational efficiency of XAI algorithms to meet the real-time processing demands of large-scale networks. Additionally, developing XAI methodologies tailored for time-series network data and designing defenses against adversarial attacks targeting XAI (Adversarial XAI) are top priorities to ensure IDS are both secure and reliable.

REFERENCES

1. S. Ali et al., "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. 99, p. 101805, 2023.
2. N. Khan et al., "Explainable AI-Based Intrusion Detection Systems for Industry 5.0 and Adversarial XAI: A Systematic Review," *Information*, vol. 16, no. 12, p. 1036, 2025.
3. S. AL and S. Sagiroglu, "Explainable artificial intelligence models in intrusion detection systems," *Engineering Applications of Artificial Intelligence*, vol. 144, p. 110145, 2025.
4. P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.
5. B. H. M. van der Velden et al., "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, p. 102470, 2022.
6. M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135-1144.
7. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765-4774.
8. S. Mane and D. Rao, "Explaining Network Intrusion Detection System Using Explainable AI Framework," *arXiv preprint arXiv:2103.07110*, 2021.
9. M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009, pp. 1-6.
10. N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*, 2015, pp. 1-6.
11. X. Zheng et al., "F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI," *arXiv preprint arXiv:2410.02970*, 2024.

Cite This Article: Nguyễn Kiên Trung, Ngô Văn Nam, Nguyễn Trung Kiên (2026). Demystifying the "Black Box" of Deep Learning in Cybersecurity: Integrating Explainable AI (XAI) to Augment the Reliability of Intrusion Detection Systems. *East African Scholars J Eng Comput Sci, 9*(2), 15-19.