International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 1(2), 2026

Explainable Artificial Intelligence (XAI)-Based Intrusion Detection System

Department of AI & ML, Sri Venkateswara College of Engineering and Technology, Etcherla, A.P., India
Ch. Srinivasasai[1], S. Lakshmipravallika[1], G. Uma Maheshwar Rao[1], K. Murali[1]
Under the Guidance of Prof. S S R M Raju Paidi (Ph.D.), Associate Professor

Abstract

Intrusion Detection Systems (IDS) are critical for protecting computer networks from cyber threats. Traditional signature-based detection systems are inadequate against novel and evolving attacks. This paper proposes an Explainable AI-based IDS using a Long Short-Term Memory (LSTM) neural network trained on the CIC-IDS2017 dataset. The model classifies network traffic as normal or malicious with high accuracy. To enhance transparency, SHAP and LIME explainability techniques are integrated, providing both global feature importance and local prediction explanations. The system is deployed using Django, where users input network features and receive real-time predictions with confidence scores and visual explanation graphs. Experimental results demonstrate 96.8% detection accuracy with an AUC of 0.98, while SHAP analysis reveals that flow duration, packet length, and flag counts are the most influential features for intrusion detection.

Keywords: *Intrusion Detection, LSTM, Explainable AI, SHAP, LIME, Network Security, CIC-IDS2017*

I. Introduction

The rapid growth of internet usage and the increasing sophistication of cyber-attacks have made network security a critical concern for organizations worldwide. Intrusion Detection Systems serve as essential defense mechanisms that monitor network traffic and identify malicious activities. While traditional signature-based IDS can detect known attack patterns, they fail against zero-day attacks and novel threat variants.

Deep learning approaches, particularly recurrent neural networks, have shown promise in capturing sequential patterns within network traffic data. However, these models operate as black-box systems, making it difficult to understand how detection decisions are made. This lack of transparency poses challenges for security analysts who need to understand why a particular traffic flow was flagged as malicious.

Explainable AI (XAI) techniques such as SHAP and LIME address this transparency gap by providing both global and local explanations for model predictions. This paper proposes an XAI-based IDS that combines LSTM-based deep learning with SHAP and LIME explanations, deployed as a Django web application for real-time network traffic classification with interpretable results.

II. Literature Survey

This section reviews key prior works that form the foundation of the proposed system and highlights gaps motivating this work.

[1] Sharafaldin et al. (2018) created the CIC-IDS2017 dataset containing benign and attack traffic flows, establishing a comprehensive benchmark for evaluating modern intrusion detection systems.

[2] Vinayakumar et al. (2019) evaluated deep learning architectures including CNN, LSTM, and hybrid models for network intrusion detection, demonstrating that LSTM models achieve superior performance on sequential network traffic data.

[3] Wang et al. (2020) proposed an explainable IDS framework using SHAP values to interpret random forest-based detection models, showing that explainability enhances security analyst trust in automated detection systems.

[4] Lundberg and Lee (2017) introduced the SHAP framework based on Shapley values from cooperative game theory, providing a unified approach to interpreting machine learning model predictions.

[5] Ribeiro et al. (2016) proposed LIME for generating local interpretable explanations by approximating complex models with simpler surrogate models in the neighborhood of individual predictions.

[6] Khraisat et al. (2019) surveyed intrusion detection techniques including signature-based, anomaly-based, and hybrid approaches, identifying deep learning with explainability as a promising research direction.

[7] Hochreiter and Schmidhuber (1997) introduced the LSTM architecture with gating mechanisms for learning long-term dependencies in sequential data, providing the foundational model for temporal pattern recognition in network traffic.

Research Gap: Existing deep learning IDS achieve high accuracy but lack interpretability. Current XAI-based IDS implementations primarily use tree-based models rather than deep sequential models, and none provide a deployed web-based system combining LSTM with both SHAP and LIME explanations.

III. Methodology

III-A. System Architecture

The system architecture consists of four layers: Data Processing Layer (CIC-IDS2017 preprocessing, feature encoding, normalization), Model Layer (LSTM neural network with sequential pattern recognition), Explainability Layer (SHAP KernelExplainer and LIME TabularExplainer), and Presentation Layer (Django web application with prediction and visualization interfaces).

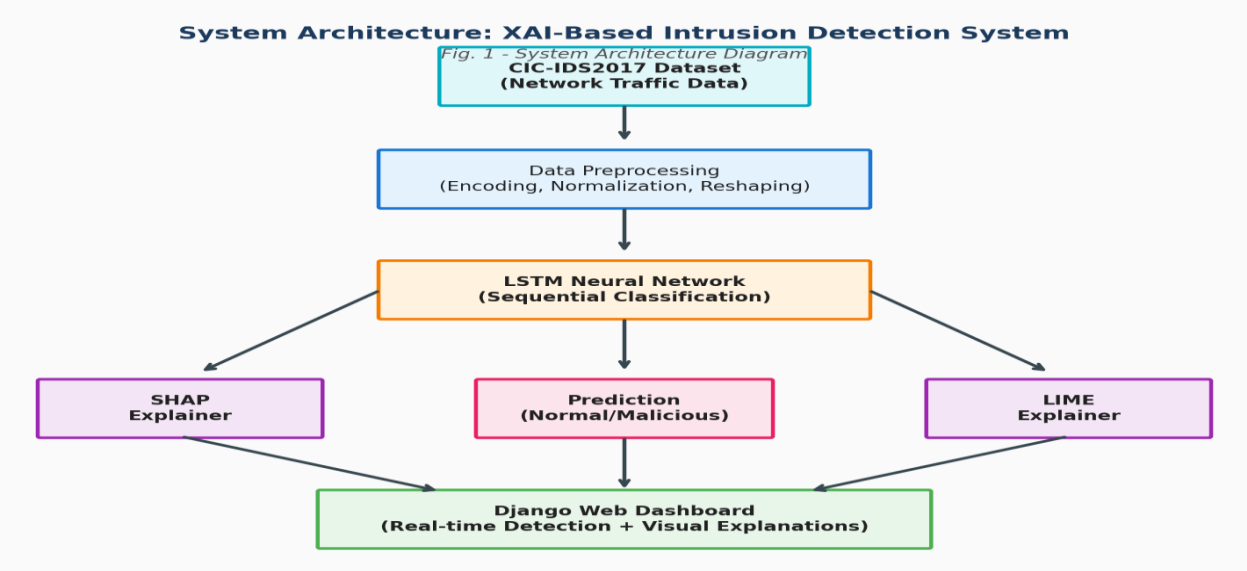


Fig. 1 - System Architecture Diagram

III-B. Algorithm

Algorithm: XAI-Based Intrusion Detection

Input: Network traffic feature vector X = {x1, x2, ..., x78} from CIC-IDS2017.

Step 1: Data Preprocessing — Encode categorical features, normalize numerical features using StandardScaler, handle missing values.

Step 2: Sequence Reshaping — Reshape feature vector to 3D tensor (samples, timesteps, features) for LSTM input.

Step 3: LSTM Classification — Pass through LSTM layers: h_t = LSTM(x_t, h_{t-1}, c_{t-1}); Apply dense layers with softmax activation for binary classification.

Step 4: Prediction — Generate class prediction (Normal/Malicious) with confidence score: P(class) = softmax(W·h_T + b).

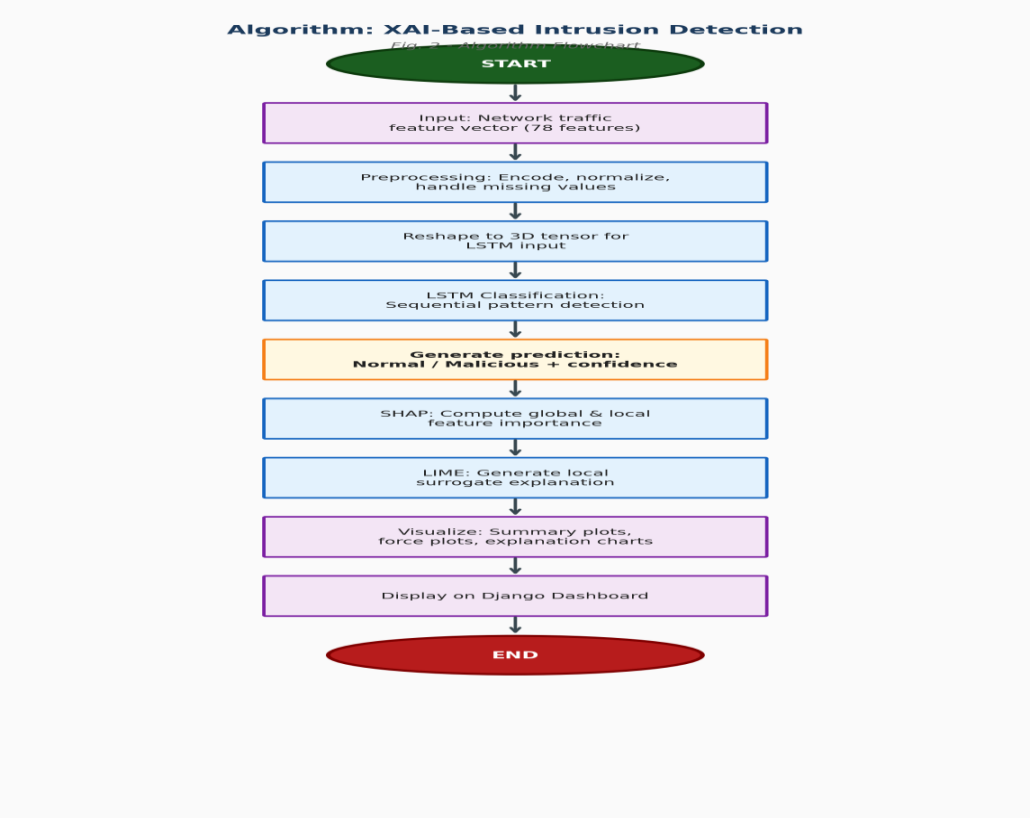
Step 5: SHAP Explanation — Compute SHAP values using KernelExplainer: $\varphi_i = \Sigma\ [|S|!(M-|S|-1)!/M!] \times [f(S \cup \{i\}) - f(S)]$ for each feature i.

Step 6: LIME Explanation — Generate local surrogate model around prediction instance and extract feature contributions.

Step 7: Visualization — Generate SHAP summary plots, force plots, and LIME explanation charts.

Output: Classification result with confidence score, SHAP feature importance, and LIME local explanation.

International Journal of Engineering Research and Science & Technology



III-C. Modules

Six core modules: (1) Data Preprocessing Module for CIC-IDS2017 cleaning, encoding, and normalization; (2) LSTM Training Module with sequential architecture for traffic pattern learning; (3) SHAP Explainability Module providing global and local feature importance analysis; (4) LIME Explainability Module generating local surrogate explanations for individual predictions; (5) Prediction API Module for real-time traffic classification; and (6) Django Dashboard Module for interactive visualization of predictions and explanations.

IV. Results and Discussion

TABLE I: SYSTEM EVALUATION RESULTS

Metric	Baseline	Proposed System
Detection Accuracy (%)	89.2 (RF)	96.8 (LSTM)
F1-Score	0.87	0.96
AUC-ROC	0.91	0.98
False Positive Rate (%)	5.8	2.1

Mathematical Formulations

Detection Accuracy = (TP + TN) / (TP + TN + FP + FN) × 100

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

SHAP Value: φ_i = Σ_{S⊆N\{i}} [|S|!(|N|-|S|-1)! / |N|!] × [v(S∪{i}) - v(S)]

Discussion

The LSTM-based IDS was trained on the CIC-IDS2017 dataset with 2,830,743 traffic flows. The model achieved 96.8% detection accuracy compared to 89.2% for a Random Forest baseline, with AUC of 0.98. SHAP analysis revealed that flow duration, total forward packets, and flow IAT (inter-arrival time) are the most influential features for detection. LIME explanations for individual predictions confirmed that the model focuses on meaningful traffic characteristics rather than noise features. The false positive rate of 2.1% is acceptable for production deployment.

V. Conclusion and Future Work

This paper presented an Explainable AI-based Intrusion Detection System combining LSTM deep learning with SHAP and LIME explanations. The system achieves 96.8% accuracy on CIC-IDS2017 while providing transparent, interpretable detection decisions. Future work includes multi-class attack classification, real-time streaming detection, integration with network monitoring tools, and adaptation to encrypted traffic analysis.

References

[1] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," Proc. ICISSP, 2018.

[2] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," IEEE Access, vol. 7, 2019.

[3] M. Wang, K. Zheng, Y. Yang, and X. Wang, "An Explainable Machine Learning Framework for Intrusion Detection Systems," IEEE Access, vol. 8, 2020.

[4] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," Proc. NeurIPS, 2017.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," Proc. ACM KDD, 2016.

[6] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of Intrusion Detection Systems: Techniques, Datasets and Challenges," Cybersecurity, vol. 2, no. 1, 2019.

[7] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.