Responsible AI in Action: Human-in-the-Loop Explainable NIDS

TA Hassan[1], MS Rashid[3*], MMJ Ayan[2*], HMM Jamil[3], M Islam[2], LA Amin[4], RK Das[5], F Quader[4]

[1]Dept. of CSE, UAP, [2]Dept. of CSE, UIU, [3]Dept. of EEE, IUT, Bangladesh, [4]Dept. of IS, UMBC, [5]Dept. of IST, PSU, USA
[*]denotes that these authors contributed equally to this work
tazzinaafroze1998@gmail.com, shahriarrashid96@gmail.com, mayan201439@bscse.uiu.ac.bd, mubashshirjamil99@gmail.com
mislam201011@mscse.uiu.ac.bd, alamin1@cognitivelinks.llc, rjd6099@psu.edu, fquader1@umbc.edu

I. Introduction

This work presents a Network Intrusion Detection System (NIDS) that integrates deep learning (DL) methods Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) with Explainable AI (XAI) model SHapley Additive exPlanations (SHAP) [1] and a human-in-the-loop interface. The design attempts to follow the principle pillars of Responsible AI, making the system not just accurate but also transparent, fair, and trustworthy to cybersecurity experts.

II. Problem Statement

DL–based NIDS achieves high accuracy but faces some major challenges such as lack of transparency [2], as their decisions are difficult for analysts to interpret, and limited trust and adaptability, since security professionals are reluctant to rely on AI-driven systems without human involvement, risking false positives, compliance gaps, and reduced usability. Addressing this problem requires integrating Responsible AI principles. Responsible AI principles are human-centered design that empowers analysts through decision involvement, feedback incorporation, and informed consent; fairness that identifies and mitigates bias to advance equity; explainability that provides human-readable insights into model predictions; security measures that protect data and ensure controlled access; reliability through continuous monitoring, audit trails, and data quality assurance; and compliance that ensures ethical data use and adherence to regulatory standards. By combining high detection performance with these Responsible AI safeguards, the proposed framework enhances transparency, fosters analyst trust, and ensures practical usability in real-world cybersecurity operations.

III. Solution and Results

The proposed system (figure 1) employs a framework combining CNN and LSTM models for network intrusion detection. SHAP was integrated to provide interpretability, allowing users to understand and trust model predictions. A structured User Interface (UI) facilitated interaction, feedback collection, and evaluation of trust, reliability, and usability among cybersecurity specialists. The UI was tested using Cronbach's Alpha test. The results demonstrate that both CNN and LSTM models achieved high overall accuracy of 99 percent on the NSL-KDD dataset as shown in table I. User evaluations indicated high levels of trust, reliability, and usability, reflecting the effectiveness of the XAI integration and the intuitive UI design as shown in table II.

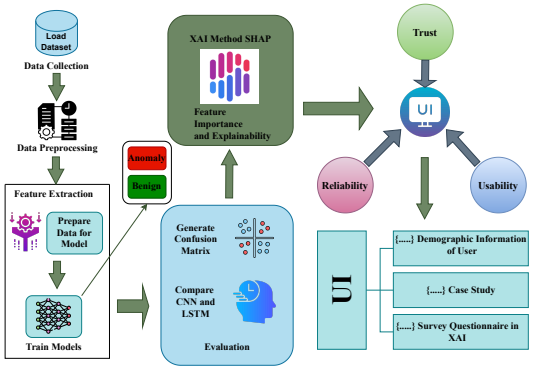

Fig. 1: Our proposed responsible AI framework for explainable deep learning-based NIDS with expert verification UI

TABLE I: Proposed NIDS system's DL models performance based on accuracy, and F1-score

Models	Accuracy	Macro Avg F1 Score	Weight Avg F1 Score
CNN	0.99	0.86	0.98
LSTM	0.99	0.93	0.99

TABLE II: UI performance via Cronbach's Alpha Test

User Experience Metrics	**Cronbach's Alpha Score**
Trust	0.90
Reliability	0.90
System Usability	0.60

IV. Conclusion

This study demonstrates that a lightweight, interpretable NIDS using SHAP can achieve high accuracy while providing transparency into model decisions. The system is effective, trusted, and usable, showing that XAI combined with a responsible AI focused UI enables robust and interpretable intrusion detection in practical settings.

References

[1] Y.-W. Chen, S.-Y. Chien, and F. Yu, "An overview of XAI Algorithms," in *2023 International Automatic Control Conference (CACS)*, pp. 1–5, IEEE, 2023.

[2] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. e1424, 2021.