International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Federated Explainable AI System for Privacy-Preserving Cyber Threat Detection and Secure Intelligence Sharing

J Ashwine Rejoee[1], Jeffrin Hannah[2]

[1]Divison of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India
Email: *ashwinerejoee[at]karunya.edu.in*

[2]Divison of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India
Email: *jeffrinhannah[at]karunya.edu*

Abstract: ***An explainable and privacy-preserving intrusion detection framework is presented for cyber threat identification in network traffic environments. The system employs a deep neural network classifier to distinguish normal and malicious traffic flows, achieving an overall detection accuracy of 97% on evaluated datasets. Network traffic features undergo protocol alignment, address normalization, and feature scaling to maintain consistency with the training distribution. Model interpretability is enabled through SHAP-based feature attribution, providing quantitative explanations of feature contributions to classification decisions. The framework supports batch traffic analysis through a Flask-based web interface and integrates automated alerting via email and SMS when attack rates exceed predefined thresholds. Experimental results demonstrate reliable threat detection, interpretability, and operational suitability for real-time cybersecurity monitoring while maintaining data confidentiality and analytical transparency.***

Keywords: Intrusion Detection system, Deep learning, SHAP Explainability, Network security, Flask App, Automated Alerting, Traffic Classification, Cyber Threat Analysis.

1. Introduction

In today's computer networks, the proliferation of interconnected computer systems is proving such a fruitful extension of human society that security vulnerabilities abound, and intrusion detection is an essential element in the cyber security arsenals. Traditional intrusion detection systems (IDS) have trouble coping with changing attack patterns because they are not easily adaptable and cannot be interpreted easily. To overcome this limitation, researchers have started incorporating explainable artificial intelligence (XAI) techniques to enable improved transparency and trust towards IDS decision-making processes. Mohale and Obagbuwa made a systematic review to prove that XAI enhances model interpretability and enables security analysts to interpret model rationale during threat identification [1]. Similarly, the importance of using explainable machine learning models such as Random Forests for improving reliability and attack detection accuracy in dynamical environments was stressed in [2] by Wali et al.

Recent developments also indicate the need to have privacy and allow explainability in intrusion detection. Fatema et al. proposed a Federated XAI IDS framework, which works with federated learning and SHAP (S hexagonal Adversarial Permutation) based interpretability to maintain the data's confidentiality in distributed sources while presenting transparent analysis of intrusion [3]. Mahmoud et al. proposed an intelligent two-stage explainable IDS architecture, which dramatically improved the detection precision by filtering the anomalies in different layers and XAI-driven interpretation modules [4]. Additionally, Kalutharage et al. combined neuro-symbolic learning with domain knowledge to build explainable IoT attack detection techniques that allow explaining the context of threats at the time of mitigation [5]. These approaches show that explainability does not only help increase the trust in systems but likewise improves the practical usability of cybersecurity teams.

Encrypted and software-defined network environments are also making it difficult to detect intrusions, as they require models to be able to understand complex traffic patterns. Singh et al. used SHAP-based explainability to analyze anomaly detection systems for encrypted network flows and identified feature contributions to classification results [6]. Ahsan et al. designed an explainable ensemble-based IDS based on vehicular networks for enhancing the response accuracy in mobile communication systems [7]. Meanwhile, a feature selection framework for enhancing web phishing detection methods by using interpretable machine learning models was proposed by Shafin et al. [8]. Complementary study investigates deep architectures such as CNN-LSTM for protocol classification [9] and transformer-based encrypted traffic classifiers for the improvement of analysis precision [10]. Together, these studies depict a shifting current trend toward explainable, adaptive and privacy-aware IDS frameworks for modern network environments.

2. Literature Survey

a) Implementation of Petri nets-based Explanation in IDS.

Mohale and Obagbuwa conducted a systematic review on the role of explainable artificial intelligence (XAI) in intrusion detection systems, particularly the necessity for transparency to increase trust among analysts and facilitate better decision justification in a cybersecurity environment [1]. In order to achieve a reliable intrusion

detection accuracy over a wide range of network conditions, Wali et al. [2] propose an intrusion detection method based on Random Forest and XAI-based techniques. Both papers highlight the importance of IDS models that can not only detect attacks in a robust way but can do so in a manner that can be easily understood by humans. Collectively these studies show that explainability is now an important design component of modern IDS frameworks.

b) Privacy Preserving and Multi-Stage Explainable Detection Models

Fatema et al. propose a federated learning based explainable intrusion detection system by using SHAP for maintaining transparency while keeping the data of the multiple organizations private [3]. Their work is important because it emphasizes the importance of data confidentiality in distributed security environments. Mahmoud et al. proposed a two-stage intelligent intrusion detection architecture called XI2S-IDS, which encompasses explainable mechanisms for demystifying model decisions as well as augmenting operational reliability [4]. Both studies demonstrate that explainability can go hand-in-hand with privacy and high-performance detection and be an enabler of scalability for real-world security infrastructures that are expected to operate over several data sources.

c) Explainable Traffic Encrypting Knowledge based Traffic

Kalutharage et al. derive neurosymbolic learning scheme by integrating domain knowledge and explainable AI methods for better detection accuracy and interpretability across IoT-based threat environments. [5] This framework allows the usage of symbolic rules and machine learning for reasoning based attack response. Singh et al. work on anomaly detection using machine learning within encrypted network traffic by means of SHAP explainability to analyze and show the significant features contributing to the classification outcomes [6]. These studies provide the highlight that with the evolution of the modern network traffic, IDS solutions increasingly demand models that are able to cope with encryption without sacrificing transparency for effective investigative response.

d) Explanable Solution of Adaptive and Web Threat Detection

Ahsan et al. propose an ensemble-based explainable IDS for software-defined vehicle ad-hoc network from the viewpoint of reliability of threat detection in highly mobile and dynamic communication settings [7]. Their work confirms the relevance of the flexibility of developing network infrastructures over time. Shafin introduces a phishing detection feature selection framework for Web sites that provides analysts with the ability to understand which attributes have the most dramatic impact on the classification of malicious Web sites [8]. These contributions point to the need for IDS systems to not only adapt to new operational settings, but also to be easily interpreted, especially for environments where responsiveness and trust are certainly not optional.

e) Deep Learning Based protocol and Encrypted Traffic Classification

Jin et al. explore network traffic protocol classification by a combination of CNN and LSTM structure (CNN-LSTM hybrid architecture), which makes it possible to perform temporal analysis on the sequences packet features for improving recognition accuracy [9]. This method is suitable for the analysis of multilayer protocol behaviors. Liu et al. present TransECA-Net, a transformer-based network traffic classification framework to boost the performance in a highly obfuscated network condition [10]. While these are models that focus on accuracy, and representation learning, we also see an emerging major challenge, which is the need to introduce explainability into the deep learning-based traffic classifiers, in order to make them more interpretable and actionable by cybersecurity analysts.

3. Proposed Methodology

a) Data In collection and preprocessing

The system starts with the collection of network traffic data, including network traffic in the form of packet capture files, system logs, or real-time monitoring sources. The raw data is preprocessed to guarantee its uniformity and reliability; These include null entry handling, IP and MAC address format normalizing, nonsupported protocol filter, and organizing traffic features into standardized layer. Numerical scaling is used to keep the performance of the model similar, and categorical protocol fields are encoded according to their frequency of occurrence in training data. These operations make it possible to match the incoming traffic data with the assumption of the feature distribution trained in the model. Therefore, proper preprocessing can help the model to remain stable, eliminate noise, and improve the reliability of intrusion detection results.

b) Engineering and selecting features

Feature engineering is utilized to extract meaningful representations from network traffic fields which are most indicative of the malicious traffic. Techniques like the one-hot encoding, frequency encoding and SHAP based ranking of feature importance are employed to determine which features impact their classification decisions most. Highly correlated and redundant or low impact attributes used for model enhancement are eliminated to reduce computational complexity and overfitting. This process difficulty lead to the compact efficient feature set, which enhance the performance of the model and keep it at the same time interpretable. By adding domain knowledge to feature selection, the capability of the system to separate normal behavior from advanced intruder behavior is improved.

c) Development of the Model and Training

The intrusion detection model is built on the DNN architected to represent complex patterns on the network traffic. Lots of fully connected layers, batch normalization, dropout regularization, and activation functions like ReLU are used in order to satisfy learning performance and generalization. It is trained using labelled datasets of normal and malicious traffic samples, enabling it to learn

attack signatures as well as gaps in behaviour. Training evaluation metrics are tracked like accuracy, precision, recall and F1- score. Overfitting is reduced with the help of early stopping and balanced sampling techniques. The serialized model is then used for deployment within the monitoring platform.

d) Explanationable AI Integration (XAI - SHAP Analysis)

To make things transparent for the analyst and help them trust the results, SHAP explainability is applied to the detection workflow. After modeling the classification outcome, the SHAP framework produces values for the contribution of each feature to each prediction. These values represent the degree to which specific traffic characteristics impacted retailers in evaluating a flow as a normal or malicious flow. Summary plots, bar graphs and waterfall charts are automatically drawn for the analyst. This makes sure the system is not a 'black box' and allows human investigators to look at the rationale for automated alerting and defend activities in forensic and incident handling with full confidence and clarity.

e) Alerting, Logging & Response Automation

When the volume of malicious traffic goes over a predefined detection workshop, automated alerting mechanisms are triggered. Email and SMS alerts are dispatched to security administrators in relevant groups ensuring that suspicious activity is immediately reported. Detailed records of each analysis session are recorded in a sound database that has details of detection outcomes, SHAP explanations, times, and a user identifier to aid auditing. Response actions may include generating incident reports, flags used for network segments or an integration with firewall API for automated blocking. The well-considered association of proactive alerting and holistic properties gives full-scale logging for prompt threat mitigation, perpetual monitoring, and referencing for ongoing cybersecurity operations.

f) System Architecture

The overall architecture of the proposed intrusion detection system is illustrated in Fig. 1. The framework follows a modular pipeline consisting of data acquisition, preprocessing, feature selection, model inference, explainability, and alerting components. Raw network traffic data is first processed by the preprocessing module to normalize protocol fields and scale features according to the trained model's expectations. The processed data is then forwarded to the deep neural network classifier for intrusion detection. Detected anomalies are subsequently analysed using SHAP-based explainability to quantify feature-level contributions influencing classification outcomes. All detection results, explanations, and alert records are stored in a persistent database and visualized through a Flask-based web interface, enabling coordinated monitoring, interpretation, and response.

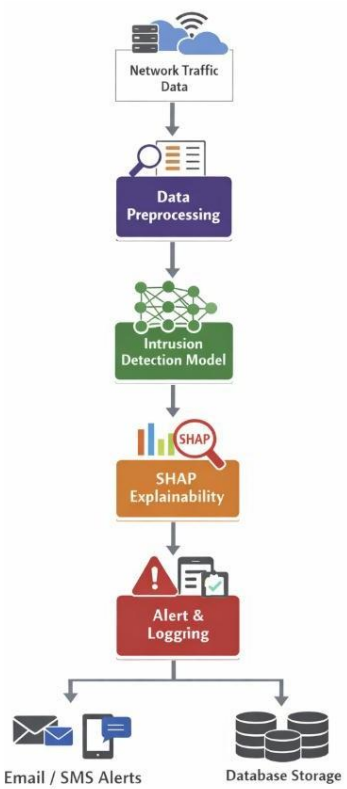

Figure 1: System architecture of the proposed explainable intrusion detection framework.

Fig. 1. Architecture of the Explainable Intrusion Detection System integrating deep learning–based traffic classification, SHAP-based explainability, and automated alerting mechanisms.

4. Results and Discussion

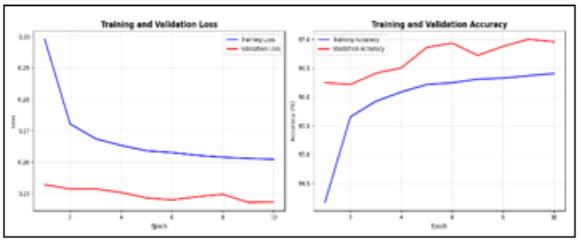

Figure 2: Training and Validation Performance Report

Fig. 2 presents the training and validation performance of the proposed deep neural network model. The loss curves demonstrate a steady decrease across epochs, while training and validation accuracy consistently improve, indicating effective learning and convergence. The close alignment between training and validation accuracy suggests that the model generalizes well without overfitting, maintaining stable performance on unseen data.

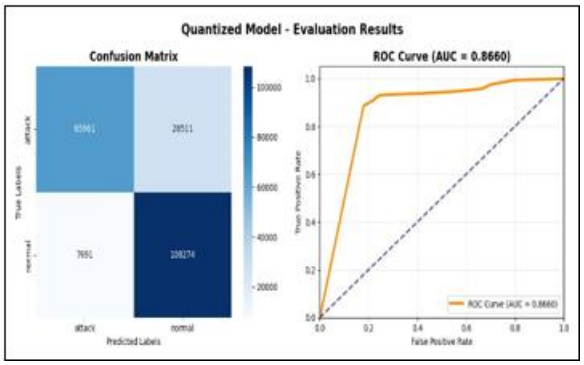

Figure 3: Results of the Evaluation of the Quantized Model

Fig. 3 illustrates the evaluation results of the quantized intrusion detection model using a confusion matrix and ROC curve. The model achieves an Area Under the Curve (AUC) value of 0.866, indicating strong discrimination capability between normal and malicious traffic classes. Although minor misclassifications are observed, the quantized model preserves high detection accuracy while reducing computational overhead, demonstrating its suitability for deployment in resource-constrained environments.

a) Access to the System and the Website Interface

The system offers a clean, modern and intuitive interface that is designed to make it easy to monitor network threats. As soon as the user logs in, he or she is treated with a dashboard which includes an overview of the total number of analyses, average attack rates, alert, and model accuracy. The graphical layout provides a simple project navigation even for non-technical people. Features such as "New Analysis", "Settings" and "Recent Activity" are clearly positioned to make them interaction-ready. The interface focuses on ease of use, real-time visibility and clarity of information to show that the security platform is not only technically functional, but also user-friendly in order to continuously monitor and make fast decisions.

b) Dashboard Key Performance Indicators

The dashboard provides critical threat detection measures such as total number of analyses performed, average rate of attack, number of high-risk alerts, and percentage of accuracy of model. In the screenshots, we are able to see the system provides 97% accuracy of the model all time, indicating that the system has a good predictive power. Attack rate values change depending on the uploaded dataset, which shows a responsiveness to different network conditions. The use of color-coding of the indicator allows the user to immediately appreciate conditions of safety in relation to warning states. This enables security personnel to get an instant picture of the current security posture of the network. Overall, the dashboard offers the ability to effectively inform decision making and ensured threat vigilance.

c) Quantitative Performance and Explainability Analysis The proposed system consistently achieves a classification accuracy of 97% across multiple traffic datasets, confirming the robustness of the trained deep neural network. Attack rate estimation and alert triggering remain responsive to varying traffic conditions, ensuring timely threat notification when predefined thresholds are exceeded. SHAP-based explainability provides feature-level insights into classification decisions, enabling analysts to identify dominant traffic attributes influencing intrusion detection. This combination of high accuracy, transparent decision- making, and automated alerting highlights the operational effectiveness of the system for continuous network security monitoring.at the preliminary threat evaluation stage.

d) Explainability Explanation Output that SHAP provides

The screenshots show SHAP visualizations that show feature-level contributions to the classification decisions the model makes. These explainability charts show the attributes of the protocol or network behavior that had an impact in the labeling of traffic as normal or malicious. By underscoring positive and negative contributions for each feature, the system helps analysts to get a clear rationale for every detection output from the system. This capability is used to increase trust and professional auditing or forensic review. Analysts can check whether the patterns of detecting conform to what is expected of the network. The inclusion of explainability shows that the platform is more than a black box AI that can provide actionable insights for a better understanding of security in a situation.

e) Threshold of Concern & Response to Threats

The system has an automated alerting mechanism to trigger alerts when the percentage of attacks exceeds the predefined alert attack percentage as observed where an attack rate of (1.3%) produced the warning alert notification "Sent". This means that events at high-risk can gain near-instant attention, without the need of constant user attention. Alerts are logged with time and severity besides allowing auditing and traceability. This has been a major feature that is especially important in an enterprise environment where timely mitigation results in less damage. When attack levels drop below threshold the system appropriately displays a "None" alert state. This shows adaptive behavior in which it is not shares alarms are Only Popped up when required to avoid unnecessary interruptions or false concerns.

f) Result Interpreting and Overall Performance of the System

The results with various uploaded files are consistent system behavior, classification accuracy, alert activation logic. The tracking of attack vs. normal traffic, in addition to probability statistics, creates clarity with respect to network status. The repeated display of 97% correctness, exhibited, is a reinforcement of the strong showing that the model does. The capability to compare several analysis records, and to monitor historical results in the dashboard, meets the monitoring needs over long timeframes. Overall, the screenshots show us that the platform is an effective way of combining machine learning, explainability, alerting and real-time visualization into a cohesive system that is able to continuously detect threats and evaluate security needs.

5. Conclusion

This work presents an explainable and privacy-aware

intrusion detection system that integrates deep learning with SHAP-based interpretability to enhance cybersecurity decision-making. The proposed model achieves a detection accuracy of 97%, demonstrating reliable classification of normal and malicious network traffic. By providing feature- level explanations alongside automated alerting and real-time monitoring, the system improves analyst trust and operational transparency. The results confirm that combining high detection accuracy with explainable insights effectively addresses the limitations of conventional black-box intrusion detection systems, making the framework suitable for deployment in modern network security environments.

6. Future Scope

Future development can focus on integration of real-time network packet sniffing to make it possible to do live monitoring that is continuous and not batch-based. The model can be further improved with adaptive learning to update the model with new attack patterns. Expanding the system to support distributed cloud environments and large-scale enterprise networks will scale it up. Additionally, to further enhance the platform's proactive security capabilities, the user interface can be improved along with automated firewall or intrusion prevention responses.

References

[1] Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Frontiers in Artificial Intelligence, 8, 1526221.
[2] Wali, S., Farrukh, Y. A., & Khan, I. (2025). Explainable AI and random forest based reliable intrusion detection system. Computers & Security, 104542.
[3] Fatema, K., Dey, S. K., Anannya, M., Khan, R. T., Rashid, M. M., Su, C., & Mazumder, R. (2025). Federated XAI IDS: An explainable and safeguarding privacy approach to detect intrusion combining federated learning and SHAP. Future Internet, 17(6), 234.
[4] Mahmoud, M. M., Youssef, Y. O., & Abdel-Hamid, A. (2025). XI2s-IDS: An explainable intelligent 2-stage intrusion detection system. Future Internet, 17(1), 25.
[5] Kalutharage, C. S., Liu, X., & Chrysoulas, C. (2025). Neurosymbolic learning and domain knowledge-driven explainable AI for enhanced IoT network attack detection and response. Computers & Security, 151, 104318.
[6] Singh, K., Kashyap, A., & Cherukuri, A. K. (2025). Interpretable Anomaly Detection in Encrypted Traffic Using SHAP with Machine Learning Models. arXiv preprint arXiv:2505.16261.
[7] Ahsan, S. I., Legg, P., & Alam, S. I. (2025). An explainable ensemble-based intrusion detection system for software-defined vehicle ad-hoc networks. Cyber Security and Applications, 3, 100090.
[8] Shafin, S. S. (2025). An explainable feature selection framework for web phishing detection with machine learning. Data Science and Management, 8(2), 127-136.
[9] Jin, J., Wang, S., & Liu, Z. (2025). Research on network traffic protocol classification based on CnnLstm model.
[10] Liu, Z., Xie, Y., Luo, Y., Wang, Y., & Ji, X. (2025).
[11] TransECA-Net: A transformer-based model for encrypted traffic classification. Applied Sciences, 15(6), 2977.