International Journal of Technology, Leadership and Sciences (IJTLS)

ISSN: 3107-908
www.ijtls.com
Research Paper
DOI: https://doi.org/10.5281/zenodo.20766506

AN EXPLAINABLE AI FRAMEWORK FOR NETWORK INTRUSION DETECTION USING MULTI-LAYER PERCEPTRONS

RAHUL JONNADULA[1],P.SANTHI PRIYA[2], Dr.CHAVA HARI BABU[3],DR.VUNNAVA DINESH BABU[4], R.VAMSI KRISHNA[5], D.SRIDHAR[6]

[1]*M.Tech Student,RV Institute Of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.*

[2]*Assistant Professor,RV Institute Of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.*

[3]*Professor,RV Institute of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.*

[4]*Professor,RV Institute of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.*

[5]*Assistant Professor,RV Institute Of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.*

[6]*Assistant Professor,RV Institute Of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.*

ABSTRACT:

Intrusion detection systems (IDS) are critical for protecting network infrastructure from nefarious activity. The growing use of machine learning models, especially Multi-Layer Perceptrons (MLPs) in intrusion detection systems (IDS), highlights the essential need of transparency and interpretability. This work seeks to improve the interpretability of MLP-based IDS models by investigating the incorporation of two Explainable Artificial Intelligence (XAI) methodologies: LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations). We evaluate LIME and SHAP using conventional IDS datasets to ascertain their effectiveness in clarifying model predictions in an understandable way for individuals. To improve the reliability and accountability of AI-driven security systems, our findings suggest that both techniques provide significant insights into model decision-making. The results illustrate the practical use of explainable artificial intelligence (XAI) approaches to improve the visibility and interpretability of black-box neural network models for cybersecurity experts.

Keywords: Intrusion Detection System (IDS), Explainable Artificial Intelligence (XAI), Multi-Layer Perceptron (MLP), SHapley Additive exPlanations (SHAP), Machine Learning, Cybersecurity.

Received Date: 5 June 2026; **Accepted Date:** 15 June 2026; **Published Date:** 20 June 2026

I.INTRODUCTION

To protect networks from unwanted assaults, Intrusion Detection Systems (IDS) are an essential element of the evolving cybersecurity environment. Machine learning (ML) techniques, especially neural networks like the Multi-Layer Perceptron (MLP), have improved traditional intrusion detection systems (IDS) by recognizing intricate and non-linear data patterns. Notwithstanding their improved accuracy, the "black-box" design of these models fosters skepticism about trust, interpretability, and transparency, which is

particularly troubling in situations where security is critical.

The field of Explainable Artificial Intelligence (XAI) has arisen to tackle this challenge; its primary aim is to improve the interpretability and usefulness of machine learning (ML) models. SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are two leading model-agnostic explainable artificial intelligence approaches. These strategies provide insights into the predictive process, assisting analysts in assessing, trusting, and improving ML-based IDS models.

This paper investigates the use of LIME and SHAP to improve the understanding of decision-making processes in MLP-based intrusion detection systems. Our aim is to reconcile the disparity between transparent cybersecurity applications and high-performing machine learning algorithms by offering both local and global interpretations of model behavior.

II.LITERATURE REVIEW

Predictions from opaque machine learning models may be elucidated using LIME (Local Interpretable Model-agnostic Explanations), a technique first introduced by Marco Tulio Ribeiro and associates in 2016. To execute the notion, perturbed samples are produced around a target instance, and a principal interpretable model is trained to locally emulate the complex model. The experimental findings demonstrated that LIME enhances model interpretability and offers sufficient explanations for individual predictions. Alternatively, it may elucidate events inside a certain context and, depending upon the disruptions, may provide varied consequences. The proposed method utilizes LIME to clarify the specific intrusion detection predictions of the IDS model.

Scott Lundberg and Su-In Lee (2017) presented a cohesive framework for interpreting predictions from machine learning models, known as SHAP (SHapley Additive exPlanations). The method employs cooperative game theory, highlighting elements based on their influence on the ultimate prediction. The results demonstrated that SHAP consistently and reliably elucidates the behavior of both local and global models, grounded in a solid theoretical framework. However, the computational cost of SHAP may be rather burdensome when dealing with intricate models and large datasets. The proposed method utilizes SHAP to provide comprehensive explanations for IDS predictions.

The Multi-Layer Perceptron (MLP) is based on concepts articulated by Geoffrey Hinton and colleagues in 1986, emphasizing neural network learning. The method identifies complex nonlinear relationships in data via interconnected neurons organized in several hidden layers. The findings indicated that pattern recognition tasks exhibited improved categorizing skills. The enigmatic characteristics of neural networks make them challenging to understand. The primary intrusion detection model proposed for the system is MLP.

Long Short-Term Memory (LSTM) is a deep learning architecture developed by Sepp Hochreiter and Jürgen Schmidhuber in 1997. The objective is to identify enduring temporal correlations in sequential data. The method employs gating mechanisms and memory cells to maintain and regulate data flow. The outcomes in time-dependent categorization and sequence prediction were significantly improved. The model requires substantial training and significant computational resources. The proposed method employs LSTM to enhance the accuracy of intrusion detection via the examination of successive network events.

XGBoost, an efficient and scalable gradient boosting framework for machine learning, was introduced by Tianqi Chen and Carlos Guestrin in 2016. This technique enhances predictive accuracy and reduces bias by integrating several decision trees with gradient boosting. Experimental results indicate that XGBoost has considerable accuracy and robustness in classification tasks. It requires meticulous calibration of computational parameters. The proposed approach utilizes XGBoost, a supplementary model, to enhance IDS efficacy.

III.EXISTING SYSTEM

With the increasing complexity and frequency of cybersecurity attacks in today's digital

environment, Intrusion Detection Systems (IDS) have emerged as an essential element of contemporary network security architectures. Conventional Intrusion Detection System (IDS) models, especially those using machine learning, have shown considerable efficacy in identifying suspicious behaviors and illegal access attempts. The Multi-Layer Perceptron (MLP) and other neural network architectures are esteemed for their exceptional pattern recognition capabilities. Notwithstanding their remarkable accuracy, these models are deficient in interpretability due to their "black box" characteristics. Implementing AI-driven intrusion detection systems in actual, mission-critical contexts is challenging owing to security experts' difficulties in understanding the reasoning behind the model's suggestions. A rising need exists to increase the openness and clarity of current Intrusion Detection Systems to provide better human supervision and compliance with regulatory requirements, notwithstanding the effectiveness of these systems in threat detection.

DISADVANTAGES

- Security professionals have difficulties in identifying the origin of a particular alarm activation.
- Improvement of the model via error analysis is unattainable if it stays opaque.

IV.PROPOSED SYSTEM

This study suggests that MLP-based intrusion detection models can improve their interpretability by incorporating Explainable AI (XAI) methodologies, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), to mitigate the interpretability shortcomings in AI-driven IDSs. Cybersecurity researchers may understand the reasoning behind danger classifications with XAI algorithms, which provide post-hoc explanations for specific forecasts. SHAP use game theory to determine the impact of each input characteristic on the ultimate prediction, whereas LIME utilizes more straightforward, interpretable models to locally mimic the behavior of the intricate MLP model. The proposed approach employs these methodologies to provide significant insights into the decision-making processes of MLP models while preserving their exceptional detection accuracy. This hybrid methodology signifies a considerable improvement over prior IDS frameworks, augmenting confidence and usability while facilitating the advancement of security measures and accelerating incident response.

ADVANTAGES

- In comprehending the decision-making process, security professionals will demonstrate more trust in the IDS.
- It enables more effective debugging and model enhancement via clear feedback.

V.SYSTEM MODEL

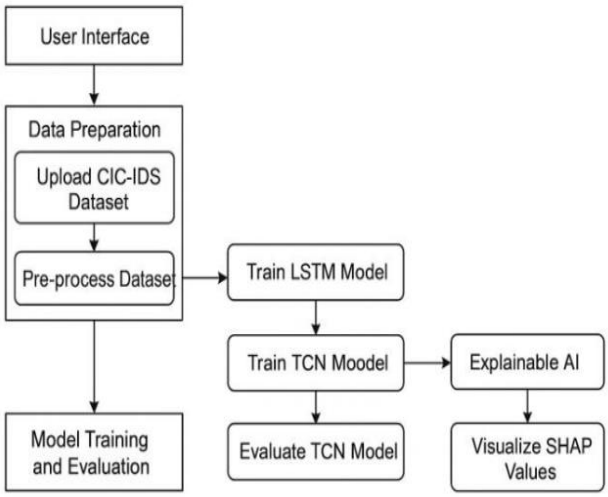


Fig 1.System Model

This research promotes the incorporation of Explainable Artificial Intelligence (XAI) into a deep learning-based intrusion detection system to rectify the shortcomings of current models. The research utilizes the CICIDS dataset, which comprises records of both benign and harmful network activity. To improve the model's efficacy, the dataset is subjected to preprocessing, which includes the elimination of missing values, standardization, and the conversion of non-numerical data into numerical format. The last stage is dividing the dataset into two pieces, according to an 80:20 ratio, so forming a training set and a testing set.

The dataset is used to train two deep learning models—Temporal Convolutional Network (TCN) and Long Short-Term Memory (LSTM)—for the identification of malicious activity in network traffic. Metrics like as F1-score, recall, accuracy, and precision are used to assess the performance of these models. The assessment becomes more precise and

demanding by including perturbations into the test data.

Additionally, feature selection and interpretation of model predictions are performed using SHAP-based Explainable Artificial Intelligence (XAI). The relative importance of various features in developing intrusion detection evaluations may thus be more thoroughly comprehended. A Graphical User Interface (GUI) is then created to facilitate user interaction with the model and provide the prediction findings along with their justifications.

VI.MODULES

•**Dataset Collection**-Obtains the CICIDS network traffic dataset, including both normal and attack records.

•**Preprocessing**–Rectifies missing values, converts non-numeric data, and normalizes attributes.

•**Model Training**–Trains Long Short-Term Memory (LSTM) and Temporal Convolutional Network (TCN) models for the purpose of intrusion detection.

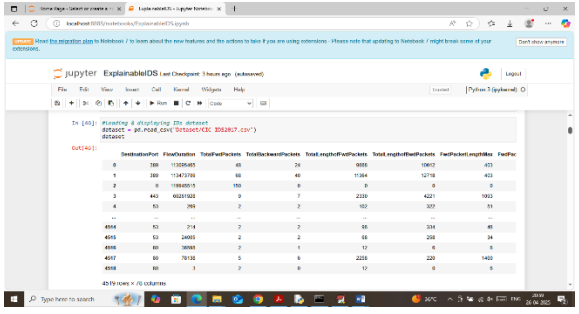
•**Evaluation**-Assesses accuracy, precision, recall, and F1-score.

•**Perturbation Testing**–Incorporates variations into test data for robustness assessment.

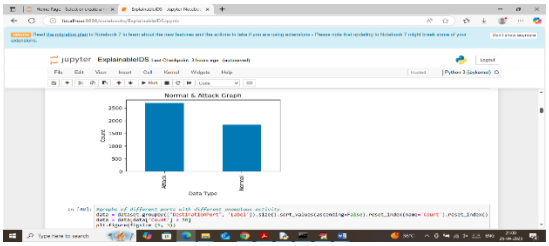
•**XAI**–Employs SHAP for feature selection and clarification of forecasts.

•**Web Application**–Provides interactive representations of detection results and their associated explanations.

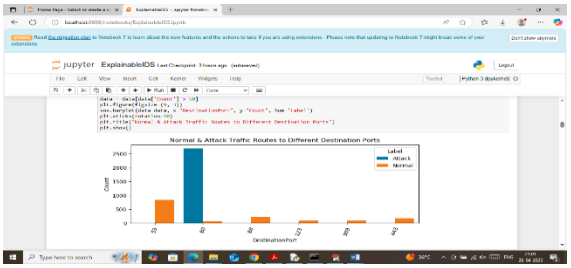
VII.SCREENSHOTS



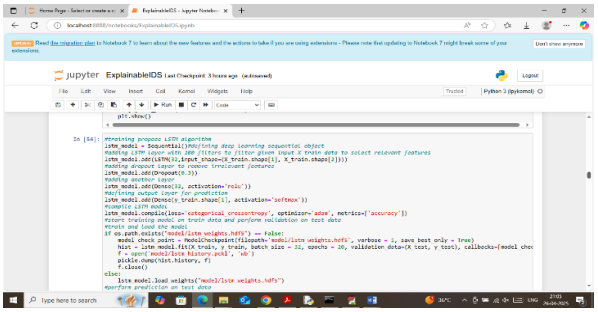
The IDS intrusion dataset has been imported and is shown on the screen above.



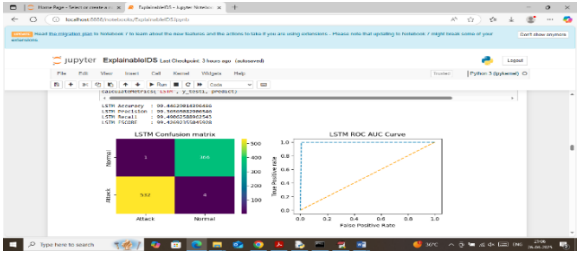
The graph depicts the total counts of normal and attack packets identified in the dataset, with the x-axis denoting packet classifications and the y-axis reflecting their corresponding quantities.



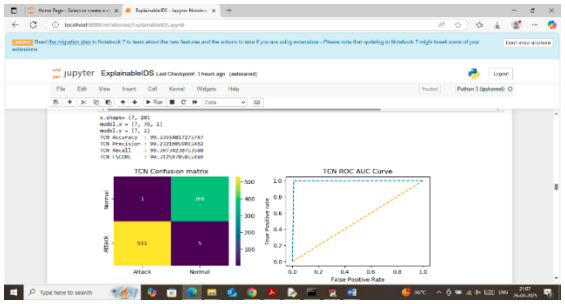
The graph above depicts the different destination ports for regular and attack packets. The blue bar in the following graph signifies "Attack," whilst the orange bar indicates "Normal port. The x-axis signifies the number of ports, while the y-axis indicates the total count. A plethora of both legitimate and harmful packets use port 80 as their first entry point, as seen in the following graph.



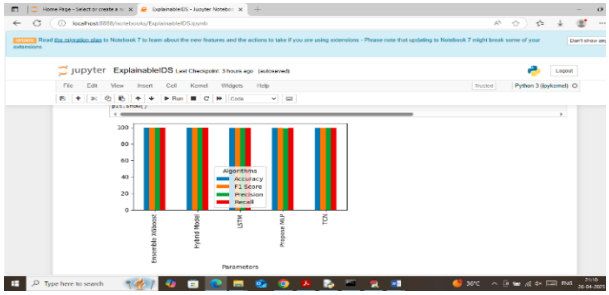
The LSTM algorithm is trained using the training data shown on the upper screen, and predictions are produced using the test data.



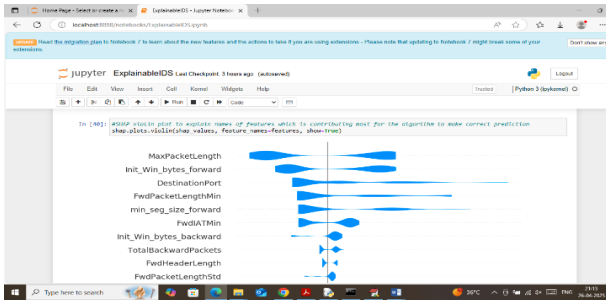
An accuracy of 99.44% was achieved using LSTM, as seen in the previous screen. Supplementary metrics like as recall, accuracy, and F-score are also included. The diagonal yellow and green boxes denote the count of accurate predictions, while the blue boxes signify the minimal number of inaccurate guesses in the confusion matrix graph, which use the x-axis for predicted labels and the y-axis for actual labels. The ROC curve depicts the True Positive Rate on the Y-axis and the False Positive Rate on the X-axis. Accurate forecasts are shown when the blue line crosses over the orange line, but erroneous predictions are depicted when the blue line falls below the orange line.



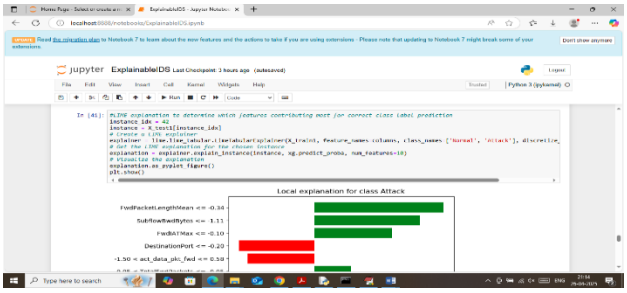
Among the previously specified parameters, TCN achieved an accuracy rating of 99.33%.



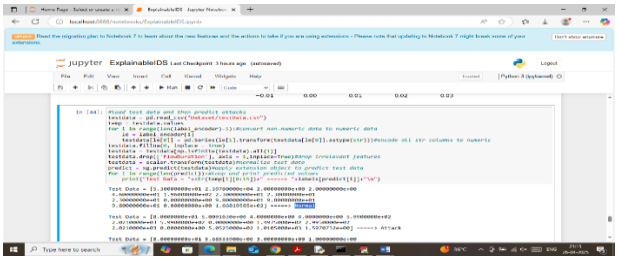
The method names are shown on the x-axis, with various bar colors indicating metric values on the y-axis; Hybrid extension2 exhibited notable accuracy overall.



The SHAP Violin plot, which presents same information on the model and its attributes, is shown in the screen above.



The next screen provides an explanation of the LIME visualization language, outlining the model and its significant features.



After importing the test data, evaluating it, and using the XGBOOST method to forecast the class label, the following screen is shown. The test data values are shown inside square brackets in the output above, with the anticipated label denoted as Normal or Attack after the = arrow sign.

VIII.CONCLUSION

A study was conducted to evaluate the effectiveness of Explainable AI (XAI) techniques, specifically SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), in improving the interpretability of Multi-Layer Perceptron (MLP) models employed in intrusion detection systems (IDS). The experimental findings indicated that LIME and SHAP substantially improve comprehension of the MLP classifier's internal decision-making mechanisms, hence providing end-users and security analysts with more trust and transparency. In contrast to SHAP's uniform global and local attributions highlighting the significance of each feature in the model's predictions, LIME offered unique local explanations for specific cases.

Incorporating explainability into Intrusion Detection System frameworks is crucial, as shown by the results. This applies to both theoretical frameworks and actual applications, where trust, transparency, and responsibility are paramount. Analysts may improve their ability to detect false positives, comprehend model behavior, and guarantee adherence to cybersecurity legislation and ethical standards via the use of these XAI methodologies.

IX.FUTURE ENHANCEMENTS

This study mainly examines the interpretability of MLP-based IDS using LIME and SHAP; nevertheless, further undiscovered pathways need future examination. A possible direction is to expand this system to include more sophisticated and real-time models, such RNNs, CNNs, or hybrid deep learning architectures. Another approach to examine the robustness of explanations is to analyze the performance of XAI tools against multifarious assaults.

Domain experts assess the effectiveness of LIME and SHAP explanations in real-world decision-making scenarios; more study should include user-centered assessments of explainability. Online learning and federated learning Intrusion Detection Systems (IDS) with explainable components may mitigate data privacy and continuous adaption issues. To improve the comparability and application of future research in this field, it is crucial to establish unambiguous metrics for assessing the validity and usefulness of XAI approaches in cybersecurity situations.

X.REFERENCES

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 4765–4774.

[3] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," Expert Systems with Applications, vol. 41, no. 4, pp. 1690–1700, 2014.

[4] E. Hodo, X. Bellekens, A. Hamilton, J. Dubouchaud, and E. Iorkyase, "Threat detection based on artificial neural networks," in 2016 Int. Symp. Networks, Computers and Communications (ISNCC), 2016, pp. 1–6.

[5] B. Shapira, L. Rokach, and S. Freilikhman, "Feature selection for anomaly detection in intrusion detection systems," Cyber Security and Information Systems Journal, vol. 1, no. 2, pp. 1–13, 2013.

[6] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[7] M. A. Rassam, A. Zainal, and M. A. Maarof, "A survey of intrusion detection systems in cloud computing," Journal of Network and Computer Applications, vol. 36, no. 1, pp. 25–41, 2013.

[8] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016.

[9] A. Das and S. Sengupta, "Explainable artificial intelligence for intrusion detection systems," in IEEE Symp. Computers and Communications (ISCC), 2020, pp. 1–6.

[10] T. Yadav and S. Selvakumar, "Detection of application layer DDoS attack by feature learning using stacked autoencoder," Neurocomputing, vol. 172, pp. 385–393, 2015.