Explainable AI IDS using SHAP/LIME on MLP

Varshini B R[1], S.Thippareddy[2]

[1]M. Tech Student, Department of Computer science and engineering, Golden valley integrated campus, Kadiri Road, Angallu Post, Madanapalli, Chittoor, Andhra Pradesh 517326

[2]Assistant Professor, Department of Computer science and engineering, Golden valley integrated campus, Kadiri Road, Angallu Post, Madanapalli, Chittoor, Andhra Pradesh 517326

Abstract Intrusion Detection Systems (IDS) play a critical role in securing computer networks by identifying malicious activities and cyberattacks. Traditional machine learning-based IDS models, particularly Multi-Layer Perceptron (MLP) neural networks, achieve high detection accuracy but suffer from a major limitation: lack of interpretability. This makes it difficult for security analysts to understand why a specific decision (benign or malicious) was made.

This study proposes an Explainable Artificial Intelligence (XAI)-based IDS using MLP integrated with SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations). The MLP model is trained on network traffic data to classify normal and attack patterns. To enhance transparency, SHAP is used to provide global feature importance, identifying which features most influence overall model predictions, while LIME offers local explanations for individual predictions, clarifying why a specific network instance is flagged as an attack or normal.

The combination of SHAP and LIME improves trust, interpretability, and usability of deep learning models in cybersecurity environments. Experimental results indicate that the proposed approach maintains high detection accuracy while significantly improving explainability for security analysts. This makes the system more suitable for real-world deployment in critical infrastructures such as cloud networks, IoT systems, and enterprise environments.

1. Introduction

With the rapid growth of internet usage, cloud computing, and interconnected devices, cybersecurity has become a major concern for individuals and organizations. Cyberattacks such as denial-of-service (DoS), phishing, malware injection, and unauthorized access are increasing in both frequency and complexity. To protect networks from such threats, Intrusion Detection Systems (IDS) are widely used to monitor network traffic and identify malicious activities.

Traditional IDS approaches, such as signature-based detection, are effective only for known attacks and fail to detect new or evolving threats. To overcome this limitation, machine learning (ML) and deep learning techniques have been introduced, which can learn patterns from data and detect both known and unknown attacks. Among these models, Multi-Layer Perceptron (MLP), a type of artificial neural network, has shown strong performance in classifying network traffic with high accuracy.

However, a major drawback of MLP and other deep learning models is their lack of interpretability. These models are often considered “black boxes,” making it difficult for cybersecurity analysts to understand how decisions are made. In critical security applications, this lack of transparency reduces trust and limits real-world adoption.

To address this issue, Explainable Artificial Intelligence (XAI) techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) are used.

These methods help interpret the predictions of complex models by explaining feature importance globally and locally. SHAP provides overall insight into which features influence model decisions, while LIME explains individual predictions in a human-understandable manner.

This project proposes an Explainable AI-based Intrusion Detection System using MLP combined with SHAP and LIME to achieve both high detection accuracy and improved transparency, making it more reliable for practical cybersecurity applications.

2. Existing system

The existing Intrusion Detection Systems (IDS) mainly rely on two traditional approaches: signature-based detection and conventional machine learning-based detection. Signature-based IDS works by comparing network traffic with a database of known attack patterns or signatures. If a match is found, the system raises an alert. This method is highly effective for detecting previously known attacks but fails to identify new, unknown, or modified attack patterns, making it less suitable for modern dynamic cyber threats.

To overcome this limitation, machine learning techniques have been introduced in IDS. Algorithms such as Support Vector Machines (SVM), Decision Trees, Random Forest, and Multi-Layer Perceptron (MLP) are widely used to classify network traffic into normal and malicious categories. These models improve detection accuracy and can identify some previously unseen attacks by learning patterns from historical data.

Among these, MLP-based IDS provides better performance due to its ability to learn complex nonlinear relationships in network data. However, despite its high accuracy, the system behaves like a "black box," meaning it does not provide explanations for its predictions. Security analysts cannot easily understand why a particular network activity is classified as an attack or normal.

Additionally, existing machine learning-based IDS systems lack interpretability and transparency, which are crucial in real-time security environments. They also require significant computational resources and careful tuning of parameters to achieve optimal performance. Without explanation mechanisms, trust in automated detection systems remains limited, especially in critical domains such as banking, healthcare, and cloud security.

Therefore, there is a strong need for an IDS that not only provides high accuracy but also offers interpretability and transparency in its decision-making process. This limitation motivates the development of an Explainable AI-based IDS using SHAP and LIME with MLP.

3. Proposed system

The proposed system introduces an Explainable Artificial Intelligence (XAI)-based Intrusion Detection System (IDS) using a Multi-Layer Perceptron (MLP) model integrated with SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations). The main objective of this system is to achieve high detection accuracy while improving transparency and interpretability of the model's decisions.

In this system, the MLP model is trained using network traffic data to classify activities as either normal or malicious. The trained model learns complex patterns and relationships in the dataset to detect various types of cyberattacks effectively. However, unlike traditional approaches, the system does not stop at prediction.

To enhance explainability, SHAP is used to provide global explanations by identifying the overall importance of each feature in the dataset. It helps in understanding which network attributes contribute most to intrusion detection across all predictions. LIME is used to generate local explanations for individual predictions, explaining why a specific instance is classified as an attack or normal.

By combining MLP with SHAP and LIME, the proposed system ensures both high performance and interpretability. This helps cybersecurity analysts to better understand model behavior, increase trust in

automated decisions, and respond effectively to potential threats. The system is designed to be suitable for real-time security environments such as cloud networks, IoT systems, and enterprise-level infrastructures.

4. Results and analysis

All machine or deep learning algorithms get trained on past dataset and then perform prediction on test data. Algorithm performance is evaluated based on predicted accuracy without performing any black box testing. Black box technique helps in knowing how algorithm predicted particular class and the predicted class label is True Positive or False Positive. Black box technique will also in explaining what features in dataset contribute most for particular class label prediction. Black box techniques help developers in knowing best features and then they can trained models with best features to get better accuracy.

In propose paper author utilizing deep learning MLP algorithm for training IDS dataset and then employing two different explainable algorithms such as SHAP and LIME for Black box testing.

LIME provides local explanations by creating a simplified, interpretable model around a specific instance, while SHAP uses game theory to attribute feature contributions to predictions, offering both local and global explanations.

LIME is best suited for explaining individual predictions and is computationally less expensive. LIME generates a set of perturbed samples around the instance you want to explain. This involves slightly modifying the input features of the original instance. These perturbed samples are then fed into the complex, "black-box" model you want to explain, and their predictions are recorded. LIME will internally perturbed different features values and then call SYSTEM function to record how False Positive and True positive prediction percentage will change based on data perturbed.

SHAP provides both local and global explanations, offering a more comprehensive understanding of the model's behaviour and is based on a more rigorous mathematical foundation.

In short both algorithms will explain how particular model utilize training features to predict particular class labels.

In propose paper author has used ADFA-LD dataset for MLP training and for explanation but this dataset not available on internet so we have CIC dataset to train IDS dataset and then perform explanation with SHAP and LIME.

Algorithms Implementation

In propose work author has used only MLP deep learning algorithm and to further enhance prediction accuracy we have experimented with multiple algorithms such as LSTM (long short term memory), TCN (temporal convolution neural network and MLP (multilayer perceptron). Each algorithm performance is evaluated in terms of accuracy, precision, recall and FSCORE. Among all propose algorithms LSTM got high accuracy.

Extension Algorithms

In propose work author has concentrate only on model explanation but has not done any experiment with different algorithms which can enhance prediction accuracy and this enhance model can help explanation algorithms to decode or explain predicted class with best features.

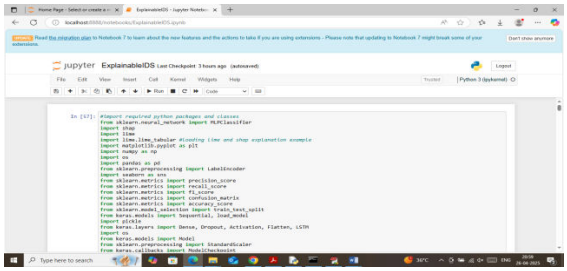
To enhance prediction accuracy we have experimented with many algorithms but accuracy got increased with Ensemble XGBOOST and Hybrid algorithm by combining multiple algorithms using Voting classifier.

Extension 1: XGBoost (eXtreme Gradient Boosting) is a powerful, open-source machine learning algorithm that utilizes gradient boosting on decision trees. It's known for its high performance, efficiency, and ability to handle large datasets. XGBoost is an ensemble learning method, combining the predictions of multiple weak learners (usually decision trees) to create a strong predictive model. Multiple decision tree help XGBOOST in enhancing accuracy.

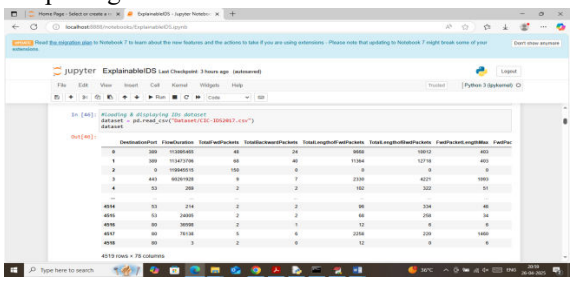
Extension 2: The voting classifier is an ensemble learning method that combines several base models

to produce the final optimum solution. The base model can independently use different algorithms such as KNN, Random forests, Regression, etc., to predict individual outputs. Combining multiple algorithms can help in finding model with best accuracy without individually experimenting with different algorithms.

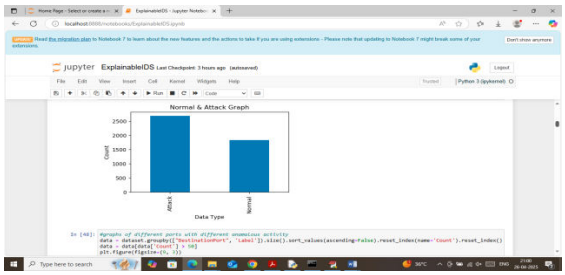
We have coded this project using JUPYTER notebook and below are the code and output screens with blue colour comments



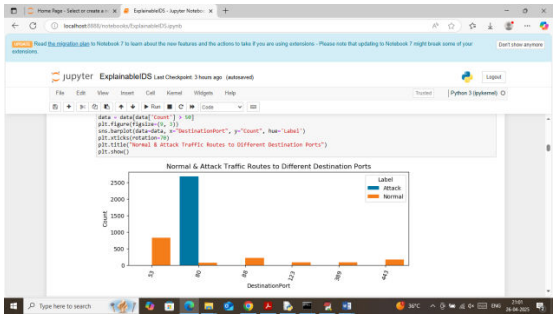
In above screen importing required python classes and packages



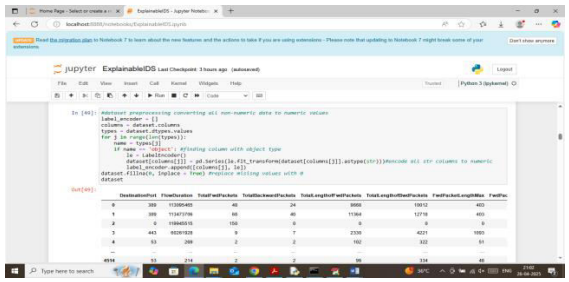
In above screen loading and displaying IDS intrusion dataset



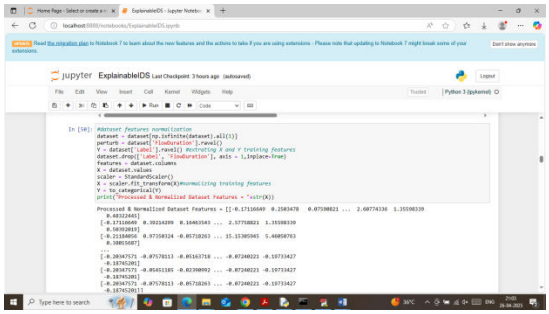
In above screen visualizing graph of number of normal and attacks packets found in dataset where x-axis represents packets type and y-axis represents count



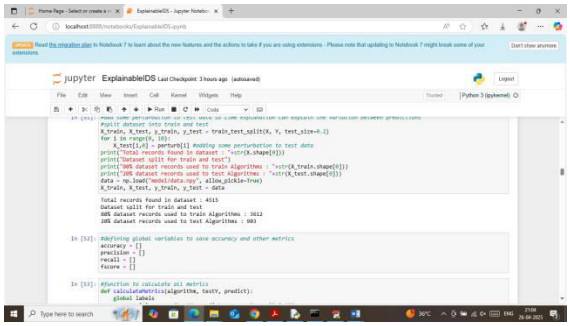
In above screen visualizing graph of different destination ports which route normal and attacks packets. In above graph x-axis represents PORT No and y-axis represents count and then blue bar represents 'Attack' and orange bar represents Normal port. In above graph can see many number of normal and attacks packets route from port 80.



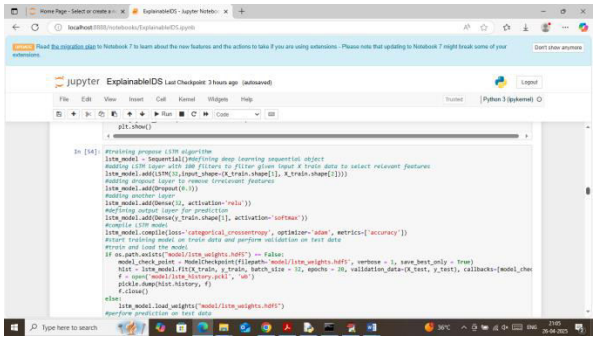
In above screen applying dataset processing to convert non-numeric values to numeric values and then replacing missing values with 0 and then displaying processed dataset values.



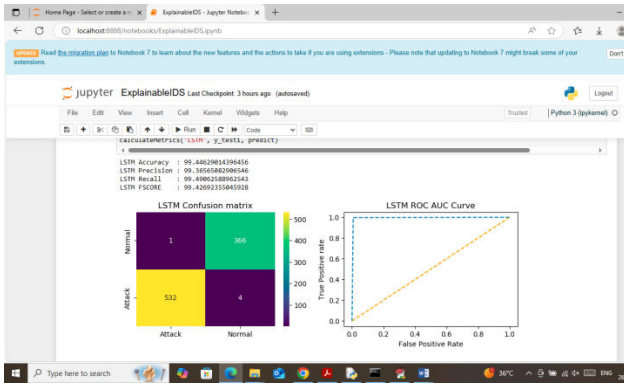
In above screen extracting X training features and Y target label from dataset and then performing some data perturbation on test data and then normalizing and displaying all features



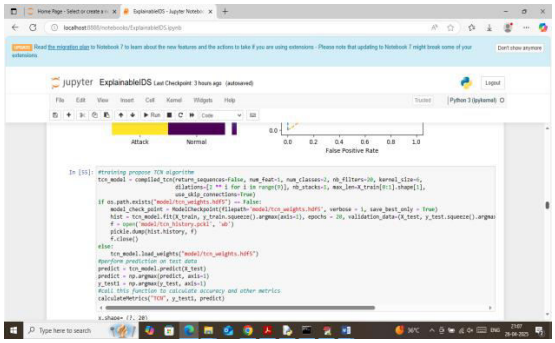
In above screen splitting dataset into train and test where application using 80% dataset for training and 20% for testing and then defining blocks to calculate accuracy and other metrics



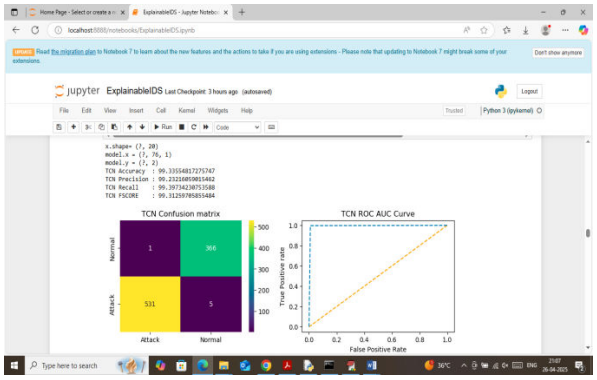
In above screen training propose LSTM algorithm on train data and then performing prediction on test data to get below output



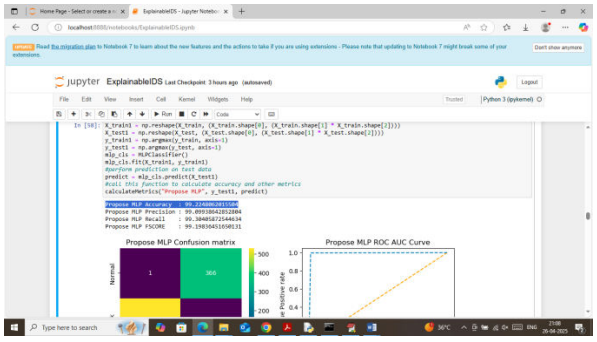
In above screen with LSTM we got 99.44% accuracy and we can see other metrics like precision, recall and FSCORE. In confusion matric graph x-axis represents Predicted Labels and y-axis represents True Labels and then yellow and green colour boxes in diagonal represents correct prediction and remaining blue boxes contains incorrect prediction count which are very few. In ROC curve graph x-axis represents False Positive Rate and Y-axis represents True Positive Rate and if blue line comes on top of orange line then all predictions are correct and if goes below orange line then predictions are incorrect.



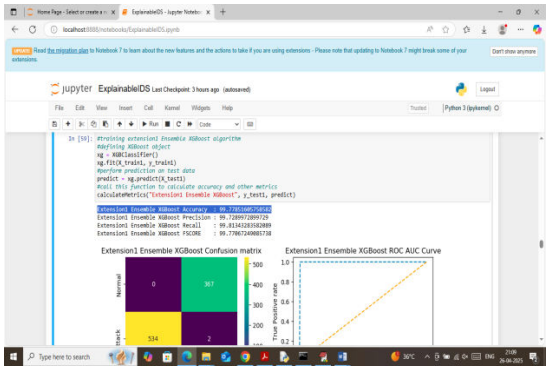
In above screen training TCN algorithm and below is the output



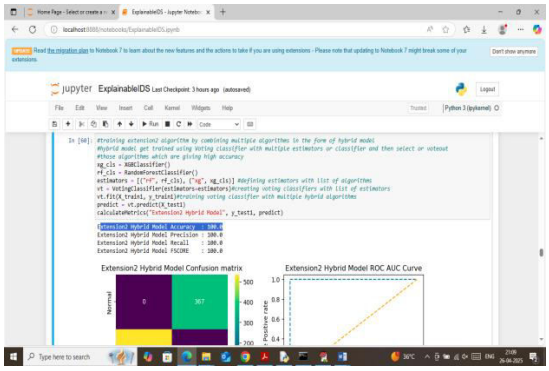
In above screen TCN got 99.33% accuracy and can see other metrics also



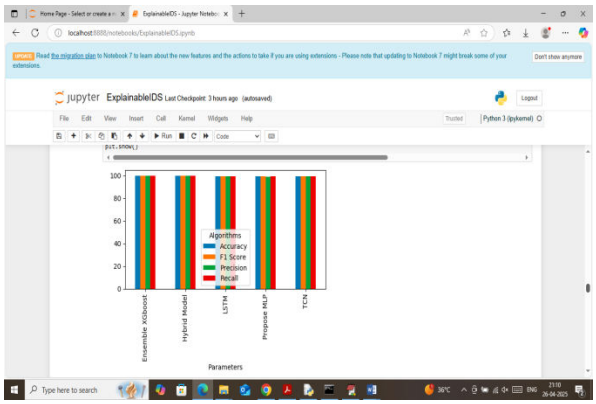
In above screen propose MLP algorithm got 99.22% accuracy and can see other metrics also



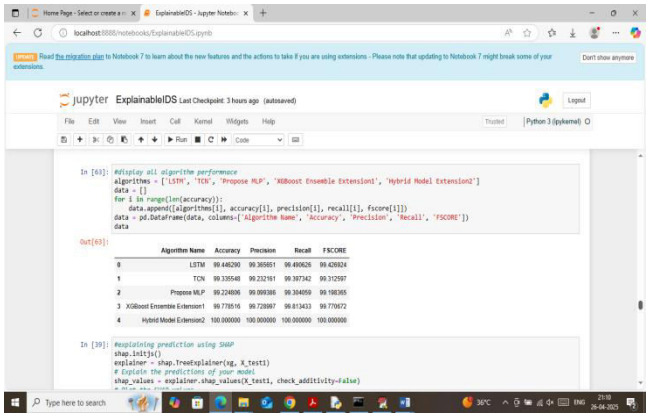
In above screen training extension1 XGBOOST algorithm which got 99.77% accuracy



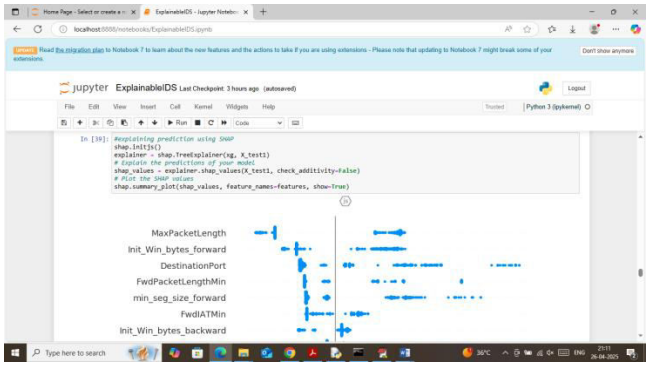
In above screen training extension 2 Voting classifier algorithm which got 100% accuracy



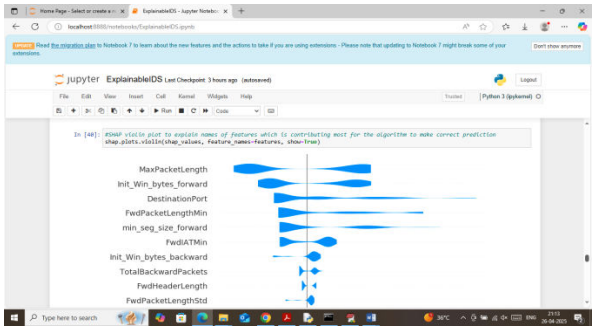
In above graph x-axis represents algorithm names and y-axis represents metric values like accuracy, precision, recall in different bar colour and in all algorithms Hybrid extension2 got high accuracy



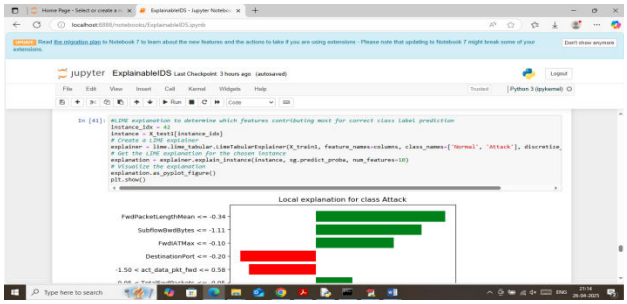
In above screen displaying all algorithms performance in tabular format



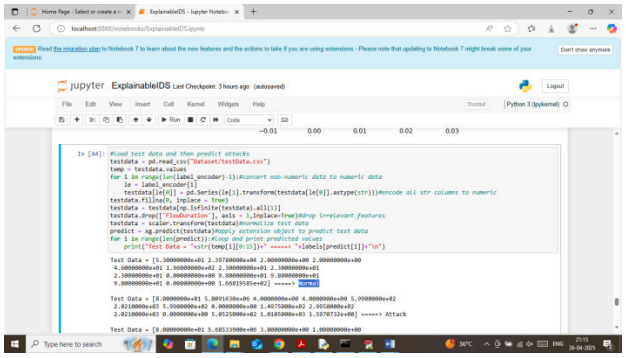
In above screen visualizing SHAP summary graph which is explaining about features which contribute model for predicting class label. In above graph first half represents ATTACK label and second half represents Normal label and then the half which contains more blue dots is the predicted class label. Features with more dots contribute most for prediction and the feature name can see in the beginning of graph



In above screen displaying SHAP Violin plot which is explaining same thing about model and features



In above screen visualizing LIME explanation which is explaining same about model and contributing features



In above screen loading test data and then performing data processing and then applying extension XGBOOST algorithm to predict test data class label. In above output in square bracket can see test data values and after =➔ arrow symbol can see predicted label as Normal or Attack.

5. Conclusions

The proposed Explainable AI-based Intrusion Detection System (IDS) using Multi-Layer Perceptron (MLP) integrated with SHAP and LIME successfully addresses the limitations of traditional and existing machine learning-based security systems. While conventional IDS models provide good detection accuracy, they lack transparency and interpretability, making them less reliable for real-world security applications.

In this work, the MLP model effectively classifies network traffic into normal and malicious categories by learning complex patterns from the dataset. To overcome the black-box nature of deep learning models, SHAP and LIME are incorporated to provide clear and meaningful explanations for model predictions. SHAP offers global interpretability by identifying the most influential features, while LIME provides local explanations for individual predictions.

The integration of explainable AI techniques improves trust, transparency, and usability of the intrusion detection system. It enables security analysts to better understand model decisions, validate outputs, and respond effectively to cyber threats.

Overall, the proposed system achieves a balance between high detection performance and interpretability, making it suitable for deployment in modern cybersecurity environments such as cloud computing, IoT networks, and enterprise systems.

REFERENCES

1. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). *Communication-Efficient Learning of Deep Networks from Decentralized Data*. Proceedings of AISTATS.
2. Lundberg, S. M., & Lee, S. I. (2017). *A Unified Approach to Interpreting Model Predictions (SHAP)*. Advances in Neural Information Processing Systems (NeurIPS).
3. Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). *"Why Should I Trust You?" Explaining the Predictions of Any Classifier (LIME)*. Proceedings of KDD.
4. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). *A Detailed Analysis of the KDD CUP 99 Data Set*. IEEE Symposium on Computational Intelligence for Security and Defense Applications.

5. Moustafa, N., & Slay, J. (2015). *UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems*. Military Communications and Information Systems Conference.
6. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). *Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization (CICIDS)*.
7. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
8. Zhang, X., et al. (2020). *Intrusion Detection Using Machine Learning Techniques in Cybersecurity: A Survey*. IEEE Access.