Stochastic Modelling and Computational Sciences

EXPLAINABLE AI FOR INTRUSION DETECTION SYSTEMS: LIME AND SHAP APPLICABILITY ON MULTI-LAYER PERCEPTRON

[1]Dr. G. Jagan Naik, [2]Dr. A. Vijendar and [3]Mrs. S. Sreshta Joshi
[1]Associate Professor, Computer Science & Engineering (Data Science), CMR Institute of Technology, Hyderabad, Telangana, India
[2]Associate Professor, Computer Science & Engineering (AI&ML), CMR Engineering College, Hyderabad, Telangana, India
[3]Assistant Professor, Computer Science & Engineering (Data Science), CMR Institute of Technology, Hyderabad, Telangana, India
[1]jagannaikg@cmritonline.ac.in., [2]vijendar.amgothu@cmrec.ac.in and [3]sreshtajoshi @gmail.com

ABSTRACT
Intrusion Detection Systems (IDS) form an inseparable part of modern cybersecurity systems, where they are required to monitor network traffic and identify malicious activity. The fast development of advanced cyberattacks and zero-day attacks has made traditional signature-based detection systems inadequate. Multi-layer perceiving machines (MLPs) and other machine-learning models have proven to be effective in learning complex non-linear patterns in high-dimensional network traffic data sets. Nevertheless, regardless of their good classification performance, these deep-learning models are black-box systems, which reduces transparency, accountability, and trust of analysts in security-sensitive settings. This paper touches upon the interpretability issue by incorporating into an MLP-based IDS model Explainable Artificial Intelligence (XAI) approaches, i.e. Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). Benchmark datasets, such as NSL-KDD and CICIDS2017 were used to evaluate the proposed approach. The MLP model achieved detection accuracy of over 95% and the explanations of instances by local surrogate modeling have been provided by the LIME and the explanations of global and local features by the SHAP based on the cooperative game theory. Experimental evidence indicates that explainability can be made without affecting predictive performance. With high detection rates and multilevel disclosure, the proposed framework increases the reliability of operations, promotes forensic examination, and raises the level of trust of AI -based cybersecurity systems among the analysts. The results prove that predictive effectiveness and interpretability are not the mutually exclusive goals, and, thus, the application of deep-learning-based IDS in the real-world should be trusted.

Keywords: *Explainable AI, Intrusion Detection Systems, LIME, SHAP, Multi-Layer Perceptron, Cybersecurity.*

I. INTRODUCTION
The exponential growth of digital communication infrastructure has resulted in a significant increase in cyber threats, including distributed denial-of-service attacks, malware intrusions, phishing campaigns, insider threats, and unauthorized access attempts. Intrusion Detection Systems (IDS) are designed to monitor network traffic and system behavior in order to identify anomalies and potential security breaches. Traditional IDS solutions rely heavily on predefined signatures and rule-based mechanisms, which struggle to adapt to evolving attack patterns and previously unseen threats.

To address these limitations, machine learning techniques have been increasingly adopted in IDS frameworks. Among them, Multi-Layer Perceptrons (MLP) demonstrate strong classification performance owing to their ability to model complex nonlinear relationships between network features and attack categories. Despite achieving high detection accuracy, MLP-based systems lack transparency, making it difficult for security professionals to understand why specific traffic instances are classified as malicious or benign. In safety-critical environments, such opacity limits trust and hinders effective incident response.
Explainable Artificial Intelligence (XAI) seeks to bridge this gap by providing human-understandable explanations for model decisions. This study investigates the applicability of LIME and SHAP in explaining MLP-based IDS predictions, with the objective of combining predictive performance with interpretability.

By integrating explanation mechanisms directly into the IDS pipeline, the proposed approach supports informed decision-making, improves analyst confidence, and enhances the practical deployment of AI-driven cybersecurity systems in real-world environment.

II. RELATED WORK

IDS has greatly evolved over the last few decades. The first systems mainly used misuse detection models that were based on rule based and signature-based models that were largely successful at defending against known threats but failed to detect the zero-day attacks or attacks that had never been seen before. Statistical anomaly detection algorithms were trying to describe the normal network behaviour using a probabilistic and statistical approach; however, it often had high false-positive rates and could not adapt to changing attacks [6], [7].The introduction of benchmark datasets like KDD Cup 1999 made it easier to systematically assess the IDS methods, and spurred the development of machine-learning-based intrusion detection [11]. Machine-learning algorithms such as Support Vector Machines (SVM), Decision Trees, k-Nearest Neighbours (k-NN), Naive Bayes, and Random Forest showed better generalisation in comparison to entirely statistical models [6], [7]. But these procedures require a lot of feature engineering and cannot easily address high dimensional and skewed intrusion data [6].

To overcome these constraints, automated extraction of the features and more powerful detection were proposed using deep-learning. This was followed by the proposal of NSL-KDD dataset to eliminate redundancy and evaluation problems of KDD Cup 1999 [11]. Deep neural networks such as Multi-layer Perceptrons (MLP) performed better by modeling a nonlinear relationship between network-traffic data [5], [6] in the networks. Convolutional Neural Networks (CNN) were used to map spatial traffic patterns, and Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks were used to model successfully temporal patterns on sequential flows [5]. Unsupervised learning of features in high-dimensional intrusion data were also proposed as autoencoder-based architectures [13]. The recent surveys [5], [6] provide comprehensive analyses of deep-learning-based models of IDS.

Recent IDS studies rely on realistic contemporary data sets like the CICIDS2017 which are more reflective of current attack vectors and traffic behaviour [10]. Such datasets have been shown to be robust and flexible when trained using deep-learning and ensemble techniques [5], [6].

Although they have better detection performance, deep-learning-based models of the IDS are black-box systems, thus creating a transparency, accountability and trust concern in cybersecurity decision-making [3]. Security analysts need obtainable explanations of automated alerts in the high stakes operational environments in order to respond effectively to any incident.

Explainable Artificial Intelligence (XAI) methods are developed to cope with this issue. LIME (Local Interpretable Model-Agnostic Explanations) was proposed to produce local surrogate explanations of black-box classifiers [8], whereas SHAP (SHapley Additive explanations) is a theory-driven feature attribution based on a cooperative game theory [9]. In recent reviews, XAI methods in the context of cybersecurity are discussed comparatively [3].

In IDS research, SHAP has been used on classical machine-learned models, e.g., Random Forest to obtain rankings of global feature-importances and explain major indicators of an attack [2]. Similarly, the use of LIME transferred to the IDS frameworks has been used to produce instance-level explanations of network-traffic classification [4]. However, investigations that examine the strength and riskiness of LIME explanations suggest that high dimensional spaces can be unstable [12].

According to comparative research, SHAP offers consistency and additive feature-attribution guarantee due to its game-theoretic nature [9], which makes it apt to global interpretability and auditing. On the other hand, LIME is computationally efficient and supports local explanations that are flexible, and is useful in real-time Security Operations Center (SOC) environments [8].

Nevertheless, the majority of available literature uses either LIME or SHAP on its own and concentrates more on classical machine-learning models than on deep-neural-network structures [2], [4]. There are limited

studies that apply both of these methods of explanation in a single deep-learning-based IDS systematically. Also, scanty studies have specifically assessed the synergies of LIME and SHAP in analyzing Multi-layer Perceptron (MLP)-based IDS models in cybersecurity applications.

Table1: Literature Review Table:

Title	Problem Statement	Solution	Methodologies	Limitations
Pinto et al., 2023 – Survey on ML-based IDS	Lack of structured comparison of ML & DL IDS	Comprehensive IDS technique survey	Comparative evaluation of ML, DL models	No dual XAI integration
Patil et al., 2022 – XAI-based IDS	Lack of interpretability in IDS	SHAP-based explanation for Random Forest IDS	Random Forest + SHAP	Limited to classical ML
Charmet et al., 2022 – XAI in Cybersecurity Review	Limited interpretability in AI-driven cybersecurity tools	Systematic review of XAI techniques in security	Comparative study of SHAP, LIME, IG, attention models	No implementation-level validation
Mane et al., 2021 – Explainable IDS	IDS alerts lack human interpretability	LIME integration for IDS	ML + LIME	No global explanation
Ferrag et al., 2020 – Deep Learning for Cyber Security Intrusion Detection	Increasing sophistication of cyber-attacks requires advanced detection beyond traditional ML	Comprehensive evaluation of deep learning models for IDS	Comparative study of CNN, RNN, LSTM on benchmark IDS datasets	Does not integrate explainability mechanisms
Ahmad et al., 2020 – Machine Learning and Deep Learning for Network Intrusion Detection	High false positive rates and poor generalization in IDS models	Systematic evaluation of ML and DL techniques for intrusion detection	Performance comparison using accuracy, precision, recall, F1-score	Lacks interpretability and transparency analysis
Choraś et al., 2020 – Machine Learning in Cyber Attack Detection	Difficulty in adapting IDS to evolving attack patterns	Adoption of ML-based detection frameworks	Supervised ML and deep learning-based classifiers	Limited focus on deep neural model explainability
Zhang et al., 2019 –	Instability in local	Quantitative evaluation of	Statistical fidelity and	Not specific to IDS domain

Understanding Uncertainty in LIME Explanations	explanation methods like LIME	LIME robustness	perturbation analysis	
Shone et al., 2018 – Autoencoder IDS	High-dimensional intrusion data	Unsupervised feature learning	Autoencoder + RF	Black-box nature

III. METHODOLOGY

The proposed framework comprises four primary stages: data preprocessing, MLP model training, integration of explainability, and performance evaluation. Benchmark IDS datasets such as NSL-KDD and CICIDS2017 were utilized to ensure representative attack coverage. Preprocessing steps include handling missing values, encoding categorical attributes, normalization, and stratified dataset splitting to maintain class balance.

The MLP architecture consists of input, hidden, and output layers with ReLU activation and cross-entropy loss. Hyperparameter tuning was performed to optimize learning rate, batch size, and network depth. Once trained, the model predicts traffic labels as normal or malicious.

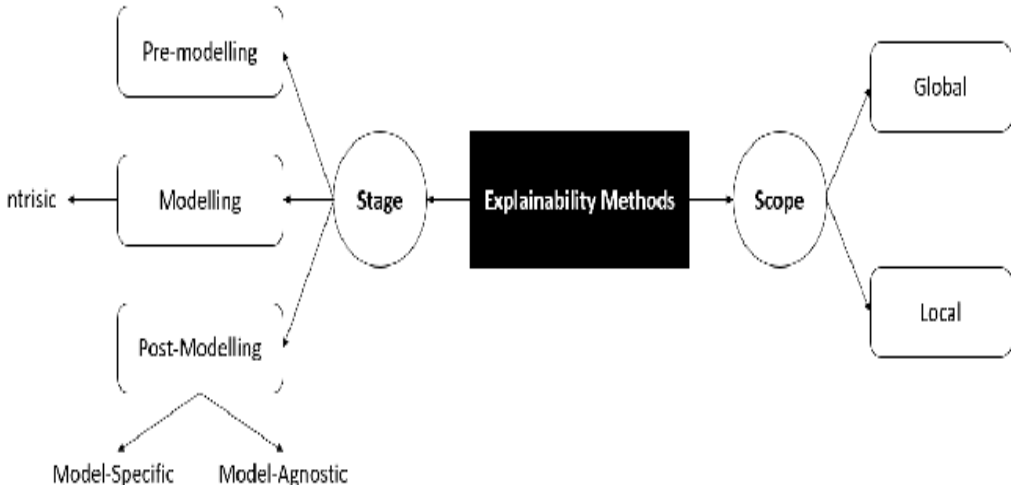

Fig.1. System Architecture

LIME is incorporated to generate localized explanations by perturbing feature values around individual instances and fitting interpretable surrogate models. SHAP is applied to compute Shapley-based feature attributions, providing both global importance rankings and local contribution values. Visualization techniques, such as summary and force plots assist analysts in interpreting the results. The evaluation metrics included accuracy, precision, recall, F1-score, and explanation consistency. The integration of LIME and SHAP ensures multilevel transparency while preserving detection effectiveness.

The process begins with data acquisition from benchmark datasets such as NSL-KDD and CICIDS2017, followed by preprocessing steps including data cleaning, categorical encoding, and normalization to ensure data consistency and quality. Feature engineering is then performed to extract meaningful traffic attributes that enhance detection capability.

The processed data is used to train a Multi-Layer Perceptron (MLP) model, which classifies network traffic as either benign or malicious. Model performance is evaluated using standard metrics such as accuracy, precision, recall, and F1-score to ensure reliable detection performance.

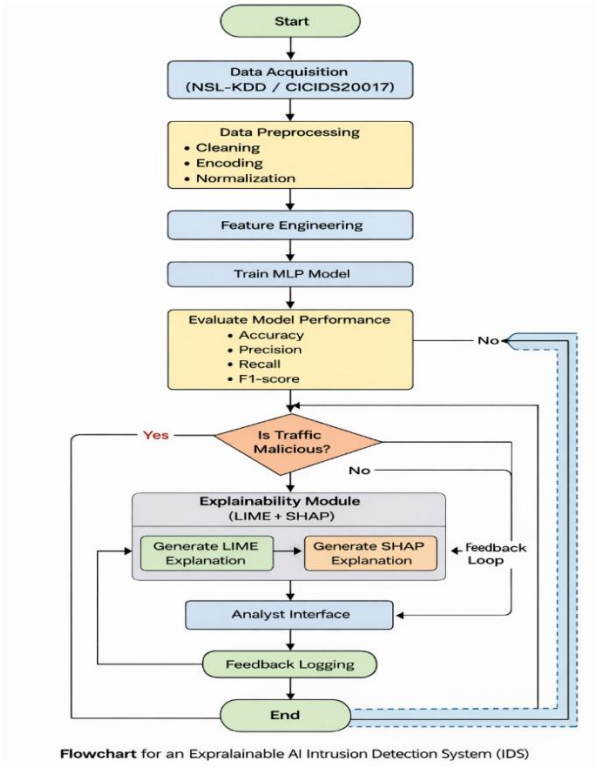

Fig.2. Flow Chart

When traffic is identified as malicious, the explainability module is activated. This module integrates LIME and SHAP to generate both local (instance-level) and global (feature-level) explanations. The generated explanations are presented through an analyst interface, enabling human validation and interpretation of model decisions. Analyst feedback is logged and incorporated into a feedback loop to support continuous model improvement and operational transparency.

This structured pipeline ensures that predictive performance and interpretability are tightly integrated within the IDS framework.

PREPROCESSING METHODS

Data Collection and Preprocessing

The proposed Explainable Intrusion Detection System (X-IDS) begins with collecting high-quality labelled network traffic data that includes both normal and various attack activities such as DoS, brute force, port scanning, and botnet attacks. Accurate labeling is essential because supervised learning models rely heavily on reliable ground-truth data. The collected dataset may consist of flow records, packet captures, or host logs that represent realistic network behavior.

After collection, the data underwent thorough preprocessing to ensure quality and consistency. This includes removing duplicates, handling missing values, standardizing numeric features, and encoding categorical attributes using techniques like one-hot encoding. To prevent data leakage, the preprocessing steps were fitted only on the training data and then applied to validation and test sets. The dataset was split using stratified sampling to maintain the class distribution, which is crucial because of the common class imbalance in intrusion detection tasks. A curated subset of representative samples may also be prepared for the explanation analysis.

Feature engineering was performed to enhance detection performance by creating meaningful cybersecurity-related the feature such as packet ratios, duration metrics, and protocol statistics. Throughout this process,

interpretability is maintained by clearly documenting feature definitions and units, ensuring that the generated explanations remain understandable and useful for security analysts.

Intrusion Detection Model: Multi-Layer Perceptron (MLP)

The core detection mechanism of the proposed framework is a multi layer perceptron (MLP), a feedforward neural network capable of modeling nonlinear relationships in structured tabular data. The input layer of the MLP corresponds to the number of engineered features, followed by multiple hidden layers designed to capture increasingly abstract representations of network behavior. Activation functions such as ReLU are employed to introduce nonlinearity, whereas dropout layers are incorporated to reduce overfitting and improve generalization. The output layer uses either sigmoid activation for binary classification or softmax activation for multi-class intrusion detection tasks.

Model training is conducted using cross-entropy loss, which is well-suited for classification problems. The Adam optimizer is typically employed to achieve efficient convergence. Hyperparameters such as learning rate, batch size, number of hidden layers, and number of neurons are tuned using validation data to ensure optimal performance. In cases where the dataset exhibits class imbalance, strategies such as class weighting or resampling are applied to prevent bias toward majority classes. After training, the model weights, preprocessing pipeline, and associated configuration parameters are securely stored to ensure reproducibility and proper version control.

Explainable AI Integration

Although MLP models offer strong predictive performance, they operate as black-box systems, making it difficult to interpret individual decisions. To address this limitation, the proposed system integrates two complementary explainability techniques: LIME and SHAP. These techniques provide transparency at both local and global levels, thereby improving trust and accountability in automated intrusion detection.

LIME is incorporated to provide local explanations for individual predictions. When a network flow is flagged as suspicious, LIME generates perturbed samples around that specific instance and fits an interpretable surrogate model, such as a linear classifier, to approximate the MLP's decision boundary in the local neighborhood. The resulting explanation highlights the most influential features contributing positively or negatively to the prediction. This produces a concise, human-readable summary that security analysts can quickly understand during incident investigation. Because LIME focuses on sparse explanations, it is particularly useful in real-time operational settings where analysts require immediate clarity regarding why an alert was generated.

In parallel, SHAP is integrated to provide consistent and theoretically grounded feature attributions based on Shapley values from cooperative game theory. SHAP calculates the contribution of each feature relative to a baseline dataset, ensuring additive consistency across predictions. Unlike LIME, SHAP supports both local explanations for individual instances and global analysis across the entire dataset. Through visualization tools such as force plots, summary plots, and feature importance bar charts, SHAP reveals which features most strongly influence model behavior overall. This dual capability enables both operational insight and strategic model auditing.

Visualization Module

To operationalize explainability, a visualization dashboard is designed for security analysts. When an alert is triggered, the dashboard presents the model's prediction along with its confidence score. LIME outputs are displayed as a ranked list of top contributing features with directional influence, providing a simplified explanation. SHAP visualizations supplement this information by showing comprehensive feature attributions and comparative importance metrics. Historical similar alerts and their resolution outcomes may also be presented to support faster decision-making.

The dashboard allows analysts to accept or reject alerts, add contextual notes, flag suspicious features, and provide feedback that can later be used for retraining. Logging explanation usage and analyst decisions supports compliance auditing and continuous improvement of the detection pipeline.

System Evaluation and Comparison

The performance of the proposed X-IDS framework is evaluated using standard classification metrics including accuracy, precision, recall, F1-score, and area under the ROC curve. A per-class confusion matrix is analyzed to identify patterns in false positives and false negatives, which are critical in cybersecurity applications. Beyond predictive performance, explanation quality is also evaluated. Local fidelity measures how closely LIME approximates the MLP's predictions, while stability testing examines whether minor input perturbations lead to drastically different explanations. Analyst feedback surveys may be conducted to assess the practical usefulness of explanations in reducing triage time and improving trust.

Comparative analysis between LIME and SHAP highlights trade-offs in computational cost, consistency, and interpretability. While SHAP provides theoretically consistent attributions and global feature insights, it can be computationally intensive. LIME, in contrast, offers faster and more intuitive local explanations but may lack global consistency. Together, they provide a comprehensive explainability framework.

Deployment

For deployment in real-world environments, performance and scalability must be carefully managed. Because explanation generation can introduce latency, LIME and SHAP computations may be selectively triggered for high-priority alerts or performed offline for global analysis. Batch SHAP processing can be scheduled periodically to update feature importance trends, while real-time LIME explanations support urgent investigations.

Security considerations are also paramount, as explanations may reveal internal model details. Access control mechanisms are implemented to restrict visibility to authorized analysts. Model versioning and explainer configuration documentation ensure reproducibility across deployments. Continuous learning mechanisms collect analyst feedback to refine both the MLP model and explanation parameters over time.

Documentation and Reporting

Comprehensive documentation accompanies the system to ensure transparency and academic rigor. This includes detailed descriptions of dataset characteristics, preprocessing steps, feature definitions, model architecture, hyperparameters, evaluation metrics, and explainability configurations. Reports present both detection performance statistics and interpretability insights, including global feature importance trends and case studies of explained alerts. Proper documentation not only supports academic publication but also enhances operational maintainability and future research extension.

IV. RESULT ANALYSIS

Experimental evaluation was conducted using benchmark intrusion detection datasets to assess both classification performance and explainability effectiveness. The Multi-Layer Perceptron model achieved consistent accuracy levels exceeding 95% across multiple training and testing iterations, demonstrating strong capability in learning complex traffic patterns. Precision and recall values indicated balanced detection of both benign and malicious instances, reducing false positives while maintaining sensitivity to attack activities.

LIME was applied to generate instance-level explanations for randomly selected predictions. These explanations highlighted the most influential network attributes contributing to individual classifications, such as packet length variance, source-to-destination byte ratio, and connection duration. By visualizing positive and negative feature contributions, analysts were able to validate model decisions and quickly identify abnormal traffic behavior. This local interpretability proved particularly useful during incident investigation, allowing security professionals to understand why specific alerts were triggered.

SHAP analysis provided global and local feature attribution using Shapley values. Global summary plots revealed that protocol type, flow duration, packet size statistics, and flag count consistently ranked among the most significant predictors. Local SHAP force plots further illustrated how combinations of features influenced specific outcomes, offering deeper insight into model reasoning. Compared to LIME, SHAP produced more stable explanations across multiple runs, though at the cost of increased computational overhead.

Quantitative comparison between the baseline MLP model and the XAI-integrated framework showed no significant degradation in detection accuracy. While SHAP introduced additional processing time, particularly for large datasets, this overhead remained acceptable for offline analysis and semi-real-time security environments. The combined use of LIME and SHAP enabled comprehensive transparency, providing both rapid instance explanations and reliable global feature understanding.

Overall, the results demonstrate that explainability can be incorporated into IDS pipelines without sacrificing predictive performance. The proposed framework enhances operational trust by transforming opaque model outputs into interpretable insights. Compared to traditional black-box IDS approaches, the XAI-enabled system supports faster alert validation, improved feature engineering, and more informed decision-making, making it suitable for practical cybersecurity deployments.

The results from this project have demonstrated that machine learning techniques can effectively predict employee attrition and layoffs strategies, when it was trained with historical hr data sets.
After preprocessing with historical hr data sets and doing exploratory data analysis and those are evaluated using standard performance metrics like precision and accuracy.

This analysis shows that ensemble-based model such as Random Forest algorithm generally outperform over simpler models like Logistic Regression and Decision Tree.

This improvement is mainly due to the ability to handle complex features, interactions and reduces overfitting which directly improves the efficiency of the project.

This analysis highlights factors like age, job satisfaction, salary level, overtime and promotion history etc, which have a impact on employee attrition.

These findings were consistent with existing researches in HR analytics, and by confirming that employee behavior is strongly connected to both organizational and personal factors.

These metrics indicate that recall is the critical measure in this context, as it is failing to identify employees at risk that may result in unexpected workforce loss.

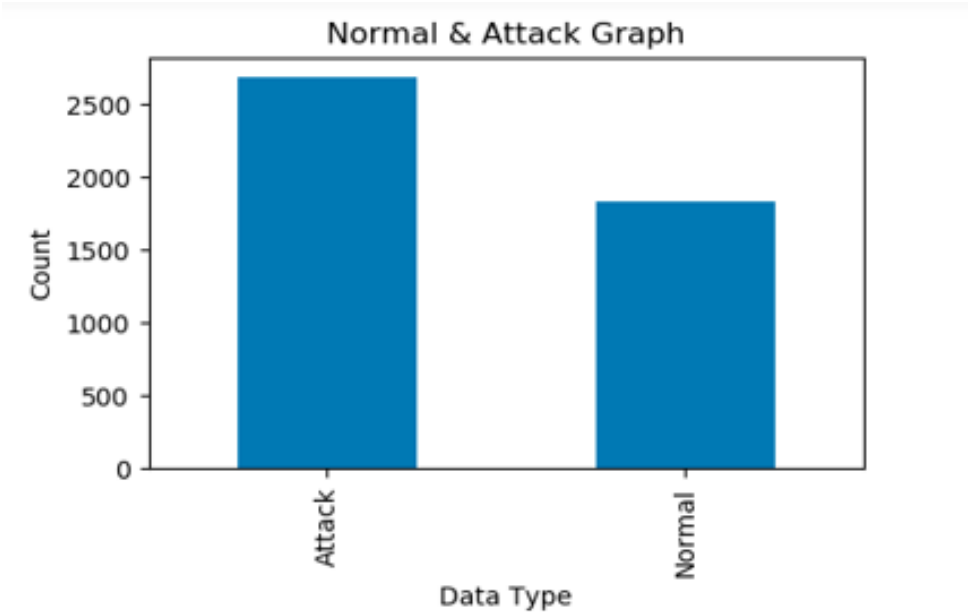

Fig.3. Normal & Attack Graph

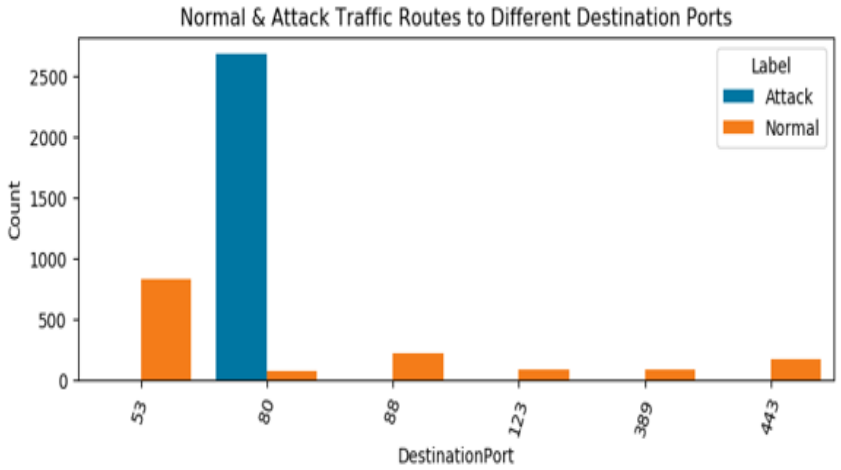

Fig.4. Normal & Attack Traffic Routes to Different Destination ports

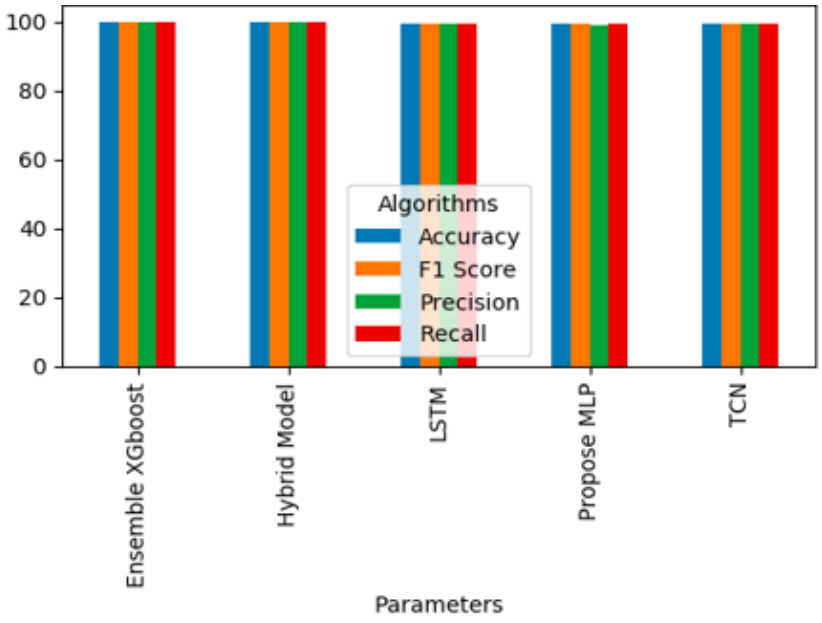

Fig.5. Confusion Matrix Parameters

V. DISCUSSIONS

The experimental findings demonstrate that integrating explainability mechanisms within deep learning–based Intrusion Detection Systems significantly enhances transparency without degrading classification performance. The MLP model successfully captured nonlinear traffic behavior and achieved high accuracy across benchmark datasets. However, the incorporation of LIME and SHAP provided deeper insight into the internal reasoning of the model, transforming opaque predictions into interpretable decision outputs.

A comparative evaluation reveals complementary strengths between LIME and SHAP. LIME is computationally efficient and suitable for rapid instance-level investigation, making it practical for operational Security Operations Center (SOC) workflows. Its local surrogate modeling approach enables analysts to quickly identify key contributing features influencing specific alerts. However, due to its perturbation-based approximation, explanation stability may vary in high-dimensional feature spaces.

SHAP, grounded in cooperative game theory, provides additive and theoretically consistent feature attributions. Unlike LIME, SHAP supports both global model interpretability and local instance analysis, making it valuable for long-term auditing, policy refinement, and feature optimization. Although SHAP

introduces higher computational overhead, especially for large-scale datasets, its stability and consistency make it suitable for batch or offline analysis environments.

The dual integration of LIME and SHAP creates a comprehensive explainability framework that supports both real-time incident investigation and strategic model auditing. Moreover, explanation outputs enable improved understanding of false positive and false negative patterns, facilitating feature refinement and threshold calibration. From a deployment perspective, the study confirms that explainability enhances analyst trust and strengthens human–AI collaboration in cybersecurity operations.

Overall, the results indicate that explainability should not be treated as an optional visualization add-on but as a fundamental component of AI-driven IDS design. Embedding transparent reasoning mechanisms directly into detection pipelines improves accountability, regulatory compliance readiness, and operational reliability in high-stakes network security environments.

VI. CONCLUSION

In this study, we explored the applicability of Explainable AI (XAI) techniques—specifically LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations)—in enhancing the interpretability of Multi-Layer Perceptron (MLP) models for Intrusion Detection Systems (IDS). Our experimental results demonstrated that both LIME and SHAP significantly aid in understanding the internal decision-making mechanisms of the MLP classifier, offering transparency and trustworthiness for end-users and security analysts. LIME provided intuitive local explanations for specific instances, while SHAP offered consistent global and local attributions, highlighting the contribution of each feature to the model's predictions.

Beyond interpretability, the integration of LIME and SHAP enabled deeper insight into feature behavior, allowing us to identify which network attributes most strongly influenced attack detection. This not only improved confidence in the model’s predictions but also supported more informed security investigations. By analyzing explanation outputs, patterns behind false positives and false negatives were better understood, creating opportunities for feature refinement and model improvement. Thus, explainability did not merely act as a visualization tool but functioned as a diagnostic mechanism to enhance overall IDS performance.

The comparative use of LIME and SHAP revealed complementary strengths. LIME proved effective for rapid, instance-level explanations suitable for operational scenarios where analysts need immediate clarity. In contrast, SHAP provided theoretically grounded, additive feature attributions that supported both local interpretation and global feature importance analysis across the dataset. Together, these methods formed a comprehensive explainability framework capable of addressing both short-term incident analysis and long-term model auditing requirements.

Importantly, this work demonstrates that high detection accuracy and interpretability are not mutually exclusive objectives. By embedding explainability into the IDS pipeline, the proposed framework bridges the gap between advanced deep learning models and practical cybersecurity deployment. The resulting system promotes transparency, accountability, and analyst trust—key requirements in high-stakes security environments.

Overall, this research contributes to the development of trustworthy AI-driven intrusion detection systems by systematically integrating MLP-based classification with dual explainability mechanisms. The findings affirm that combining predictive performance with meaningful explanations enhances not only model transparency but also operational effectiveness, paving the way for more responsible and reliable AI adoption in cybersecurity.

VII. FUTURE WORK

While this work concentrates on improving the interpretability of MLP-based Intrusion Detection Systems using LIME and SHAP, numerous opportunities exist for further advancement and research.One important direction is extending the explainability framework to more sophisticated deep learning architectures such as hybrid CNN–LSTM models, Graph Neural Networks (GNNs) for network topology-aware detection, and

Transformer-based models designed for sequential traffic analysis. As network traffic increasingly exhibits temporal and relational dependencies, evaluating how explainability techniques adapt to these architectures would significantly enhance practical applicability.

Another promising area involves real-time and streaming intrusion detection environments. Future work could focus on developing lightweight or approximate explainers that reduce computational overhead, enabling low-latency explanations suitable for production-scale Security Operations Centers (SOCs). Optimization strategies such as incremental SHAP computation, surrogate caching, or hardware acceleration (GPU/TPU-based explainability) could be investigated.

Robustness analysis under adversarial conditions also presents a critical research direction. Attackers may attempt to manipulate input features not only to evade detection but also to produce misleading explanations. Studying explanation stability under adversarial perturbations and developing adversarially robust explanation techniques would strengthen trust in deployed systems. Defensive strategies such as explanation consistency checks or ensemble-based explanation validation could also be explored.

Future research could further incorporate human-in-the-loop learning mechanisms. Analyst feedback on explanations could be systematically collected and integrated into retraining pipelines, enabling adaptive IDS models that evolve with emerging threats. Measuring how explainability impacts analyst decision speed, false positive reduction, and overall operational efficiency would provide valuable real-world validation.

Another direction involves combining explainability with automated response systems. High-confidence SHAP patterns could be translated into rule-based detection heuristics or policy updates, bridging the gap between data-driven learning and traditional signature-based approaches. This hybrid defense strategy may improve both accuracy and response time.

Cross-dataset generalization is another area worth exploring. Evaluating the proposed framework across multiple modern intrusion datasets and real enterprise traffic would test the scalability and robustness of explanations. Transfer learning approaches could also be examined to determine whether explanation patterns remain consistent across domains.

Furthermore, fairness and bias analysis within IDS models can be investigated. Certain network segments, protocols, or user behaviors might be disproportionately flagged as malicious. Using SHAP-based global analysis to detect biased feature dependencies could lead to more equitable and reliable detection systems.

Finally, future work may explore standardized metrics for evaluating explanation quality in cybersecurity contexts. While fidelity and stability are useful indicators, domain-specific metrics that measure operational usefulness, interpretability for non-technical analysts, and compliance readiness would provide a more holistic evaluation framework.

By addressing scalability, robustness, human interaction, fairness, and real-time deployment challenges, future research can further advance the development of trustworthy, transparent, and resilient explainable intrusion detection systems suitable for modern cybersecurity infrastructures.

REFERENCES

[1] A. Pinto, L.-C. Herrera, Y. Donoso, and J. A. Gutierrez, “Survey on Intrusion Detection Systems Based on Machine Learning Techniques,” *Sensors*, vol. 23, no. 5, 2023.

[2] S. Patil, R. S. Deshpande, and P. Kulkarni, “Explainable Artificial Intelligence for Intrusion Detection Systems,” *Electronics*, vol. 11, no. 19, 2022.

[3] A. Charmet, J. Kwon, and M. Choo, “Explainable Artificial Intelligence in Cybersecurity: A Systematic Review,” *Journal of Cybersecurity and Privacy*, 2022.

[4] S. Mane and P. K. Sinha, “Explaining Network Intrusion Detection System Using Explainable Artificial Intelligence (XAI),” *arXiv preprint*, 2021.

[5] M. A. Ferrag, L. Maglaras, A. Ahmim, and H. Janicke, "Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study," *Journal of Information Security and Applications*, vol. 50, 2020.

[6] Z. Ahmad, A. Shahid, M. A. Khan, and S. U. Khan, "Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches," *IEEE Access*, vol. 8, 2020.

[7] G.Jagan Naik, "effective distributor based decision making approach using ETL, data warehousing based on smart business intelligent technology" ISSN: 2229-7359, international journal of environmental sciences

[8] M. Choraś, R. Kozik, W. Holubowicz, and R. Renk, "Machine Learning – The New Cyber Attack Detection Paradigm," *Computers & Security*, vol. 90, 2020.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1135–1144.

[10] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[11] G.Jagan Naik, "explainable AI and Blockchain for cyber resilient online retail :A framework for enhanced security and trust"ISSN: 2229-7359, international journal of environmental sciences

[12] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization (CICIDS2017)," in *ICISSP*, 2018.

[13] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A Detailed Analysis of the KDD Cup 99 Dataset," in *IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009.

[12] J. Zhang, H. Chen, and Y. Li, "Understanding Uncertainty in LIME Explanations," *arXiv preprint*, 2019.

[13] A. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A Deep Learning Approach to Network Intrusion Detection Using Autoencoders," *IEEE Access*, vol. 6, 2018.

[14] G. Jagan Naik, Effective Distributed Based Decision-Making Approach Using ETL Data Warehousing Based on Smart Business Intelligence Technology International Journal of Environmental Sciences ISSN: 2229-7359Vol. 11 No. 21s,2025 https://theaspd.com/index.php.

[15] A Vijendar, P Madhavi Advancing Healthcare with Deep Learning: Innovations in Medical Image Analysis International Journal of Environmental Sciences ISSN: 2229-7359Vol. 11 No.18s,2025 https://theaspd.com/index.php.