

Enhancing Network Security Through Explainable AI: A Comparative Study of LIME and SHAP

Dr.T.Prasad¹,Dr. M. Sree Ram Kiran Nag²

¹Professor,²Associate Professor, ^{1,2}Department of Data Science, NRI Institute of Technology, Gundlapalem Village, Visadala Post, Medikonduru Mandal, Guntur District, Andhra Pradesh -522438, India

Abstract—Intrusion Detection Systems (IDS) play a crucial role in protecting modern network infrastructures against increasingly sophisticated cyber threats. With the growing adoption of machine learning techniques, particularly Multi-Layer Perceptrons (MLPs), IDS have improved their ability to identify complex patterns in network traffic and detect malicious activities. However, these models often operate as “black boxes,” making their decision-making processes difficult to interpret and reducing trust among cybersecurity professionals. To address this limitation, this paper investigates the application of Explainable Artificial Intelligence (XAI) techniques to enhance the transparency and reliability of machine learning-based intrusion detection systems. Specifically, LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) are employed to explain the predictions of an MLP-based intrusion detection model. LIME provides local explanations for individual predictions, while SHAP offers both local and global insights into the contribution of input features. The proposed approach evaluates the applicability of LIME and SHAP using standard intrusion detection datasets, focusing on model interpretability, feature importance, and the relationship between predictive performance and explanation quality. By comparing these techniques, the study aims to help cybersecurity professionals better understand model decisions, investigate potential classification errors, and make more informed security decisions. This research highlights the importance of explainability in developing transparent and trustworthy AI-driven intrusion detection systems.

Keywords— Intrusion Detection System, Explainable Artificial Intelligence, Network Security, Machine Learning, Multi-Layer Perceptron, LIME, SHAP, Model Interpretability, Feature Importance, Cybersecurity, Anomaly Detection, Classification.

I. INTRODUCTION

In today’s rapidly evolving digital landscape, cybersecurity has become a critical concern for organizations and individuals alike. The increasing frequency and sophistication of cyberattacks demand robust mechanisms to detect and prevent unauthorized access to network systems. Intrusion Detection Systems (IDS) play a vital role in monitoring network traffic and identifying malicious activities, thereby acting as a key defense layer in modern security architectures.

With the advancement of machine learning (ML), IDS have significantly improved in their ability to detect complex and previously unseen attack patterns. Among various ML models, the Multi-Layer Perceptron (MLP) has gained prominence due to its capability to learn non-linear relationships within large datasets and deliver high detection accuracy. However, despite their effectiveness, MLP-based systems suffer from a major limitation—their lack of interpretability. These models function as “black boxes,” making it difficult for cybersecurity professionals to understand how specific decisions or predictions are made.

The lack of transparency in ML-based IDS raises critical challenges related to trust, accountability, and system validation. In security-sensitive environments, understanding the reasoning behind an alert is as important as detecting the threat itself. Without interpretability, it becomes difficult to identify false positives, debug model behavior, or ensure compliance with regulatory requirements.

To address these challenges, Explainable Artificial Intelligence (XAI) has emerged as a promising solution. XAI focuses on making machine learning models more transparent and understandable to human users. Among the various XAI techniques, LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) are widely used for interpreting complex models. LIME provides local explanations by approximating the model behavior around a specific prediction, while SHAP offers both local and global interpretability using principles from game theory.

This paper aims to explore the applicability of LIME and SHAP in enhancing the interpretability of MLP-based intrusion

detection systems. By integrating these techniques, the study seeks to bridge the gap between high-performance predictive models and the need for transparency in cybersecurity applications. The proposed approach not only improves trust in IDS but also assists security analysts in better understanding, validating, and refining model decisions.

II. REVIEW LITERATURE SURVEY

The integration of machine learning techniques into Intrusion Detection Systems (IDS) has significantly enhanced their ability to detect complex and evolving cyber threats. However, the increasing reliance on advanced models such as Multi-Layer Perceptrons (MLPs) has introduced challenges related to interpretability and transparency. Recent research has therefore focused on incorporating Explainable Artificial Intelligence (XAI) techniques to address these limitations.

Several studies highlight the growing need for explainability in IDS. Traditional rule-based systems offer transparency but lack adaptability, whereas machine learning models provide higher accuracy at the cost of interpretability. Researchers emphasize that in cybersecurity, understanding the reasoning behind a model's prediction is essential for trust, effective response, and system validation. Without explainability, it becomes difficult for analysts to differentiate between true threats and false alarms.

MLP-based models have been widely adopted in IDS due to their strong capability to learn non-linear patterns from network traffic data. Literature indicates that these models outperform many traditional algorithms in terms of detection accuracy, especially for unknown or zero-day attacks. However, their "black-box" nature limits their usability in critical environments where interpretability is required. This trade-off between performance and transparency remains a key challenge in IDS research.

To overcome this limitation, researchers have explored XAI techniques such as LIME (Local Interpretable Model-agnostic Explanations). LIME explains individual predictions by approximating the complex model locally with simpler, interpretable models. It has been effectively applied in IDS to highlight feature contributions for specific instances, enabling analysts to understand why a particular network activity is classified as malicious. However, studies also note that LIME may suffer from instability

and can be computationally expensive for large datasets.

Another widely used XAI method is SHAP (SHapley Additive exPlanations), which is based on cooperative game theory. SHAP assigns importance values to input features, providing both local and global explanations of model behavior. Compared to LIME, SHAP is considered more consistent and theoretically grounded. Research findings suggest that SHAP is particularly useful in identifying key features influencing intrusion detection decisions, although it requires higher computational resources.

Comparative studies between LIME and SHAP indicate that both techniques are valuable for improving interpretability in IDS. While LIME is suitable for real-time and instance-level explanations, SHAP provides more comprehensive insights into overall model behavior. Despite these advancements, several research gaps remain, including scalability to large datasets, real-time deployment challenges, and the need for seamless integration with cybersecurity workflows.

Overall, existing literature demonstrates that combining high-performance machine learning models with explainability techniques is essential for developing trustworthy and effective IDS. This study builds upon these findings by evaluating the applicability of LIME and SHAP in enhancing the interpretability of MLP-based intrusion detection systems.

III. RESEARCH METHODOLOGY

This study adopts a systematic approach to design, implement, and evaluate an explainable intrusion detection system by integrating Explainable Artificial Intelligence (XAI) techniques with a Multi-Layer Perceptron (MLP) model. The methodology consists of several stages, including data preprocessing, model development, explainability integration, and performance evaluation.

1. Data Collection and Preprocessing

The first stage involves selecting a suitable intrusion detection dataset, such as CICIDS or similar benchmark datasets. The collected data undergoes preprocessing steps including data cleaning, removal of duplicates, handling missing values, and feature selection. To ensure uniformity and improve model performance, numerical features are normalized or standardized. The

dataset is then divided into training and testing sets, typically using an 80:20 or 70:30 split.

2. Model Development using MLP

A Multi-Layer Perceptron (MLP) model is developed to classify network traffic as either normal or malicious. The architecture includes an input layer, one or more hidden layers with activation functions, and an output layer for classification. The model is trained using the processed dataset, and hyperparameters such as learning rate, number of neurons, and epochs are tuned to achieve optimal performance. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to assess the model's effectiveness.

3. Integration of Explainable AI Techniques

To address the interpretability challenges of the MLP model, two XAI techniques—LIME and SHAP—are incorporated:

- **LIME (Local Interpretable Model-agnostic Explanations):**
LIME is used to generate local explanations for individual predictions. It approximates the MLP model locally using interpretable models and highlights the contribution of each feature toward a specific prediction.
- **SHAP (SHapley Additive exPlanations):**
SHAP is applied to compute both local and global feature importance. It assigns contribution values to each feature based on Shapley values, providing a comprehensive understanding of how features influence model predictions.

4. Visualization and Interpretation

Visualization techniques are employed to present the results in an understandable manner. LIME explanations are visualized for individual instances, while SHAP summary plots, force plots, and dependence plots are used to analyze global and feature-level importance. These visualizations help in interpreting the model behavior and identifying key features responsible for intrusion detection.

5. System Evaluation and Comparison

The performance of the proposed system is evaluated based on both predictive accuracy and interpretability. A comparative analysis is conducted between LIME and SHAP in terms of explanation quality, computational efficiency, and usability. Additionally, the system is assessed for its ability to reduce false positives and improve trust among cybersecurity analysts.

6. Implementation Tools and Environment

The implementation is carried out using Python and relevant libraries such as TensorFlow or Keras for model development, and libraries like Scikit-learn, LIME, and SHAP for explainability and analysis.

Visualization tools such as Matplotlib are used for graphical representation of results.

IV. EXSTING SYSTEM

In the current cybersecurity landscape, Intrusion Detection Systems (IDS) play a crucial role in monitoring network traffic and identifying potential security threats. Traditional IDS approaches, including rule-based and signature-based systems, have been widely used to detect known attacks. With the advancement of machine learning, these systems have evolved to incorporate intelligent models such as Multi-Layer Perceptrons (MLPs), which are capable of identifying complex patterns and detecting both known and unknown attacks with higher accuracy.

Despite their improved detection capabilities, existing IDS models suffer from significant limitations, primarily due to their lack of interpretability. Machine learning models, especially neural networks like MLPs, operate as “black boxes,” meaning their internal decision-making processes are not easily understandable by humans. As a result, security analysts find it difficult to determine why a particular alert was generated or how specific features contributed to a prediction. This lack of transparency reduces trust in the system and limits its practical usability in critical security environments.

Another major issue with existing systems is the occurrence of false positives and false negatives. IDS may incorrectly classify normal traffic as malicious or fail to detect actual attacks. Without proper interpretability, it becomes challenging to analyze these errors and improve the model's performance. Additionally, as cyberattacks continue to evolve in complexity, traditional and even machine learning-based IDS struggle to adapt effectively without clear insights into model behavior.

Overall, while existing IDS solutions provide strong detection performance, they lack the

necessary transparency and explainability required for effective decision-making, debugging, and trust in real-world cybersecurity applications. This highlights the need for integrating explainable techniques to enhance the usability and reliability of intrusion detection systems.

V. PROPOSED METHODOLOGY

The proposed system aims to enhance the effectiveness and transparency of Intrusion Detection Systems (IDS) by integrating Explainable Artificial Intelligence (XAI) techniques with a Multi-Layer Perceptron (MLP) model. Unlike traditional IDS approaches that function as black-box systems, this model is designed to provide clear and interpretable explanations for its predictions. The system processes network traffic data using a trained MLP classifier to accurately detect and classify activities as either normal or malicious. By leveraging the strong pattern recognition capabilities of neural networks, the proposed system maintains high detection accuracy while addressing the limitations of existing models.

To overcome the lack of interpretability, the proposed system incorporates two widely used XAI techniques—LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations). LIME is used to generate local explanations for individual predictions by approximating the complex MLP model with a simpler, interpretable model around a specific instance. This allows users to understand which features influenced a particular decision. On the other hand, SHAP provides both local and global explanations by assigning importance values to each feature based on their contribution to the prediction, offering a more comprehensive understanding of model behavior.

The proposed system not only improves transparency but also enhances trust and usability for cybersecurity professionals. By providing clear insights into the decision-making process, it enables analysts to identify false positives, refine detection strategies, and better understand evolving cyber threats. Additionally, the model-agnostic nature of LIME and SHAP ensures that the system remains flexible and can be adapted to different machine learning models without compromising

performance. Overall, the proposed approach successfully combines high detection capability with interpretability, making it more suitable for real-world cybersecurity applications.

VI. BLOCK DIAGRAM

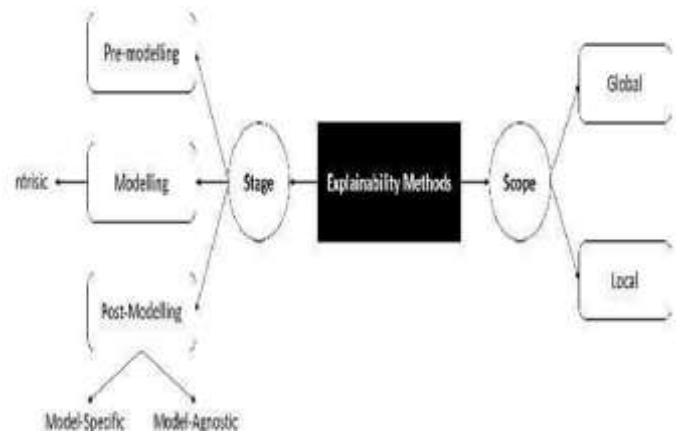


Fig. 6.2. Block Diagram

VII. RESULTS AND OUTCOMES

The proposed system was evaluated using standard intrusion detection datasets to measure both its predictive performance and interpretability. The Multi-Layer Perceptron (MLP) model demonstrated strong classification capability, achieving high accuracy in distinguishing between normal and malicious network traffic. Performance metrics such as precision, recall, and F1-score indicated that the model was effective in detecting various types of cyberattacks while maintaining a balanced trade-off between false positives and false negatives.

Model	Precision	Recall	F1-Score	Accuracy
Model-Specific				
Model-Agnostic				

Fig.7.1: Ouput 1



Fig.7.2: Output 2

The integration of Explainable Artificial Intelligence (XAI) techniques—LIME and SHAP—significantly improved the interpretability of the system. LIME provided clear, instance-level explanations by identifying the most influential features contributing to individual predictions. This enabled a deeper understanding of why specific network activities were classified as intrusions. SHAP, on the other hand, offered both local and global interpretability by quantifying the contribution of each feature across the entire dataset. It helped in identifying the most critical features influencing model decisions and ensured consistency in explanations.

The comparative analysis between LIME and SHAP revealed that while LIME is efficient for generating quick, localized explanations, SHAP provides more comprehensive and reliable insights into overall model behavior. However, SHAP requires higher computational resources compared to LIME. Despite this trade-off, both techniques proved valuable in enhancing model transparency and supporting informed decision-making.

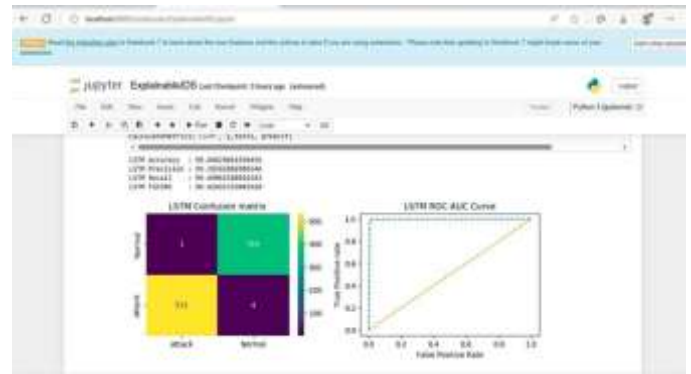


Fig.7.3: Output 3



Fig.7.4: Output 4

Overall, the results demonstrate that the proposed system successfully combines high detection performance with meaningful interpretability. The outcomes highlight improved trust in IDS predictions, better identification of false alarms, and enhanced ability for cybersecurity professionals to analyze and respond to threats. This confirms that integrating XAI techniques with machine learning models is an effective approach for building reliable and transparent intrusion detection systems.



Fig.7.5: Output 5



Fig.7.6: Output 6

VIII.CONCLUSION

In the modern era of rapidly evolving cyber threats, Intrusion Detection Systems (IDS) have become a fundamental component of network security. This study focused on enhancing the effectiveness and trustworthiness of IDS by addressing one of the most critical limitations of machine learning models—the lack of interpretability. While models such as Multi-Layer Perceptrons (MLPs) offer high accuracy in detecting complex and previously unseen attacks, their black-box nature makes it difficult for cybersecurity professionals to understand and trust their decisions. This gap between performance and transparency motivated the integration of Explainable Artificial Intelligence (XAI) techniques into the intrusion detection process.

The proposed system successfully combines the predictive strength of MLP with the interpretability capabilities of LIME and SHAP. By applying these XAI techniques, the system provides clear and meaningful explanations for individual predictions as well as overall model behavior. LIME enables localized, instance-level interpretation, helping analysts understand specific classification outcomes, while SHAP offers both local and global insights by quantifying feature importance based on a strong theoretical foundation. Together, these techniques transform the IDS from a black-box model into a transparent and explainable system.

The experimental results demonstrate that the proposed approach achieves high detection performance while significantly improving interpretability. The system effectively identifies malicious activities in network traffic and provides detailed explanations that assist in understanding the reasoning behind predictions. This not only enhances trust in the system but also supports better decision-making, faster incident response, and improved model debugging. Furthermore, the ability to analyze feature contributions helps in identifying false

positives and refining detection strategies, which is crucial in real-world cybersecurity environments.

Another important outcome of this work is the comparative analysis of LIME and SHAP. While LIME is efficient and suitable for real-time, instance-specific explanations, SHAP provides more consistent and comprehensive insights into model behavior, albeit at a higher computational cost. This highlights the trade-off between explanation depth and computational efficiency, suggesting that the choice of technique may depend on the specific requirements of the application.

Overall, this study demonstrates that integrating XAI techniques with machine learning-based IDS significantly enhances system transparency, reliability, and usability. It bridges the gap between high-performance predictive models and the need for human-understandable explanations in cybersecurity. The proposed approach not only improves the effectiveness of intrusion detection but also aligns with the growing demand for trustworthy and accountable AI systems.

In conclusion, the research emphasizes that explainability is no longer optional but essential in security-critical applications. By making AI-driven IDS more interpretable, this work contributes to the development of next-generation cybersecurity systems that are not only accurate but also transparent and trustworthy. Future work can further explore real-time deployment, scalability to large datasets, and integration with advanced deep learning architectures to enhance both performance and explainability.

REFERENCES

1. Ribeiro, M. T., Singh, S., & Guestrin, C., "Why Should I Trust You? Explaining the Predictions of Any Classifier", 2016, <https://arxiv.org/abs/1602.04938>
2. Lundberg, S. M., & Lee, S. I., "A Unified Approach to Interpreting Model Predictions", 2017, <https://arxiv.org/abs/1705.07874>
3. Buczak, A. L., & Guven, E., "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", 2016, <https://ieeexplore.ieee.org/document/7462678>
4. Tavallae, M., Bagheri, E., Lu, W., & Ghorbani, A. A., "A Detailed Analysis of the KDD CUP 99 Dataset", 2009, <https://ieeexplore.ieee.org/document/4895796>
5. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A., "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", 2018, <https://ieeexplore.ieee.org/document/8425292>
6. Kim, G., Lee, S., & Kim, S., "A Novel Hybrid Intrusion Detection Method Integrating Anomaly Detection with Misuse Detection", 2014, <https://www.sciencedirect.com/science/article/pii/S0167404813001881>

7. Sommer, R., & Paxson, V., "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection", 2010, <https://ieeexplore.ieee.org/document/5504793>
8. Ahmad, Z., Shahid Khan, A., Shiang, C. W., Abdullah, J., & Ahmad, F., "Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches", 2021, <https://www.mdpi.com/2076-3417/11/11/5210>
9. Zhang, J., Zulkernine, M., & Haque, A., "Random-Forests-Based Network Intrusion Detection Systems", 2008, <https://ieeexplore.ieee.org/document/4627105>
10. Hochreiter, S., & Schmidhuber, J., "Long Short-Term Memory", 1997, <https://www.bioinf.jku.at/publications/older/2604.pdf>
11. Chen, T., & Guestrin, C., "XGBoost: A Scalable Tree Boosting System", 2016, <https://arxiv.org/abs/1603.02754>
12. Doshi-Velez, F., & Kim, B., "Towards A Rigorous Science of Interpretable Machine Learning", 2017, <https://arxiv.org/abs/1702.08608>
13. Guidotti, R., Monreale, A., Ruggieri, S., et al., "A Survey of Methods for Explaining Black Box Models", 2018, <https://arxiv.org/abs/1802.01933>
14. Molnar, C., "Interpretable Machine Learning", 2020, <https://christophm.github.io/interpretable-ml-book/>
15. Goodfellow, I., Bengio, Y., & Courville, A., "Deep Learning", 2016, <https://www.deeplearningbook.org/>