International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 2, 2026

Peer Reviewed, Referred & Indexed Journal

Improving Transparency In Intrusion Detection Systems Through LIME And SHAP-Based Explanation Of MLP Models

[1]Lokesh Devathati,[2]V. Vaishnavi,[3]S. Akshith,[4]V. Vengamma

[1]Assistant Professor, Department of Computer Science & Engineering, Eluru College of Engineering and Technology

[2,3,4]B. Tech Student, Department of Computer Science & Engineering, Eluru College of Engineering and Technology

ABSTRACT

Intrusion Detection Systems (IDS) are critical for safeguarding network security by identifying malicious activities in real-time. While Multi-Layer Perceptron (MLP) neural networks have demonstrated high accuracy in detecting intrusions, their complex decision-making processes remain largely opaque, limiting their practical adoption. This study explores the application of Explainable Artificial Intelligence (XAI) techniques—specifically LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations)—to enhance the interpretability of MLP-based IDS. By providing transparent and interpretable explanations of individual intrusion predictions, these methods help security analysts understand, trust, and effectively respond to alerts generated by the system. The comparative analysis highlights the strengths and limitations of LIME and SHAP in terms of explanation quality, computational efficiency, and applicability in real-world intrusion detection scenarios. The integration of XAI with MLP models promises to bridge the gap between high-performance detection and explainability, advancing the development of trustworthy cybersecurity solutions.

Keywords: Intrusion Detection System (IDS), Explainable Artificial Intelligence (XAI), Multilayer Perceptron (MLP), LIME, SHAP, Cybersecurity, Network Attack Detection, Machine Learning Interpretability, Feature Importance Analysis, Transparent AI Models.

INTRODUCTION

As cyber threats become increasingly sophisticated, organizations depend heavily on IDS to safeguard their networks. Traditional IDS approaches rely on manually crafted rules or shallow machine learning models, which often fail to adapt to novel attack patterns. Deep learning models like MLPs, trained on large-scale network traffic datasets, have emerged as powerful alternatives, capable of detecting both known and zero-day attacks with high precision.

Despite their predictive power, deep learning models pose a major challenge: their decision-making process is opaque. In security-sensitive domains, stakeholders demand transparency to understand why an alert is triggered—whether to validate its authenticity, improve system defenses, or meet regulatory requirements. Explainable AI bridges this gap by making complex models interpretable without significantly compromising performance.

LIME generates local surrogate models to explain individual predictions, while SHAP assigns Shapley values to quantify each feature's contribution to a prediction. Applying these methods to MLP-based IDS can enhance analyst trust, improve attack forensics, and guide model improvement.

This research focuses on assessing LIME and SHAP applicability in explaining MLP-driven IDS, analyzing their trade-offs in accuracy, speed, and interpretability.

I. LITERATURE SURVEY

1. Ribeiro et al. (2016) – "Why Should I Trust You?": Explaining the Predictions of Any Classifier

Ribeiro et al. introduced LIME, a model-agnostic method for interpreting black-box predictions by approximating them with simple interpretable models locally. Their study demonstrated LIME's flexibility across domains, including text and image classification, but its application to high-dimensional network intrusion data was not specifically addressed.

2. Lundberg & Lee (2017) – A Unified Approach to Interpreting Model Predictions

Lundberg and Lee developed SHAP, a game-theoretic approach based on Shapley values, offering consistent and additive feature importance scores. SHAP has proven effective in providing global and local interpretability, but computational cost remains a challenge for large-scale IDS datasets.

3. Shapoorifard et al. (2021) – Explainable Artificial Intelligence for Intrusion Detection

Shapoorifard et al. explored the integration of XAI methods, including LIME and SHAP, into IDS frameworks. Their experiments on NSL-KDD and CICIDS datasets showed that interpretability improved analyst trust and reduced false alarm rates, though runtime performance was a limiting factor for real-time deployment.

4. Zhang et al. (2020) – Interpreting Deep Learning Models for Cybersecurity

Zhang et al. applied SHAP to deep learning-based intrusion detection and found that certain features, such as packet length and connection duration, consistently influenced model decisions. Their work highlighted that XAI can help security experts prioritize relevant features for improved IDS design.

5. Khan et al. (2022) – Local and Global Explanations in Cybersecurity Models

Khan et al. compared LIME and SHAP in explaining MLP and Random Forest models for network anomaly detection. They concluded that SHAP provided more stable and globally consistent explanations, whereas LIME was faster and more suitable for instance-specific analysis, suggesting a hybrid use in IDS environments.

II. EXISTING SYSTEM

Intrusion Detection Systems have traditionally relied on signature-based or rule-based techniques that detect attacks by matching known patterns or anomalies. However, these methods often fail to detect novel or evolving threats due to their static nature.

To improve detection accuracy, machine learning (ML) and deep learning (DL) approaches—especially Multi-Layer Perceptrons (MLPs)—have been widely adopted in IDS. MLPs learn complex patterns from network traffic data and can classify intrusions with high accuracy. Despite their effectiveness, these models act as "black boxes," providing little insight into why a particular network event is flagged as malicious.

Several existing IDS implementations utilize MLPs or other deep learning models but lack explainability, limiting trust and interpretability for cybersecurity analysts. Some works have explored model-agnostic interpretability methods like LIME and SHAP in other domains, but their application in IDS, particularly on MLP-based models, remains underexplored.

Recent research has started integrating explainability techniques with IDS, but challenges remain around balancing interpretability, computational efficiency, and real-time applicability. Furthermore,

many existing approaches provide either local or global explanations but not both, which can limit their usefulness in operational security environments.

Overall, while the performance of MLP-based IDS has improved, the absence of transparent decision-making hampers the practical deployment of these systems. This motivates the need for applying and evaluating explainability tools like LIME and SHAP tailored for IDS environments.

III. PROPOSED SYSTEM

The proposed system focuses on improving the transparency and interpretability of Intrusion Detection Systems (IDS) by integrating explainable artificial intelligence techniques with a Multilayer Perceptron (MLP) model. In this approach, network traffic data is first collected and preprocessed through steps such as data cleaning, normalization, and feature selection to ensure high-quality input for the learning model. The processed data is then used to train an MLP classifier capable of detecting different types of network intrusions and malicious activities. To enhance the transparency of the prediction process, two popular explainability techniques, Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), are applied to the trained model. These techniques analyze the model's decisions and highlight the most influential features contributing to each prediction. LIME provides local explanations for individual predictions, while SHAP offers both local and global feature importance insights based on cooperative game theory. By combining accurate intrusion detection with interpretable explanations, the proposed system allows security analysts to understand why specific network activities are classified as normal or malicious. This improves trust, facilitates faster incident analysis, and supports better decision-making in cybersecurity environments.

IV. SYSTEM ARCHITECTURE

Explainable AI for Intrusion Detection System (XAI-IDS) operates as a comprehensive three-tier pipeline designed for transparent network security. The architecture begins with the Data Preprocessing Module, which is responsible for ingesting raw network traffic flows—such as those found in datasets like CIC-IDS2017—and preparing them for analysis. This preparation involves critical steps like feature selection to isolate relevant metrics (e.g., flow duration and packet lengths), as well as normalization and scaling to ensure numerical consistency, which is vital for stabilizing the downstream machine learning process.

The cleaned data is then fed into the Detection and Classification Module, the core of the system, which utilizes a high-performance Multi-Layer Perceptron (MLP) model. This MLP, comprising dense layers and a Softmax output, is trained to classify traffic flows into various categories, such as 'Normal' or specific intrusion types like 'DoS' or 'PortScan', and simultaneously provides a probability-based confidence score for its decision. This module is optimized for fast, real-time assessment, flagging potential threats as they occur on the network.

Crucially, the architecture is enhanced by the Explainability Generation Module (XAI), which addresses the "why" behind the MLP's predictions. When an intrusion is flagged, both LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) techniques are applied post-hoc to the detected flow. LIME provides local fidelity by highlighting the few most influential features that pushed a single data point toward a malicious classification, while SHAP offers a more

rigorous, game-theoretic perspective by calculating the fair contribution (SHAP value) of every input feature to the final prediction, often used for both local and aggregated global importance insights.

Finally, the entire process culminates in the Presentation and Visualization Layer, which takes the MLP's Classification Label and the corresponding XAI feature attributions and renders them on an analyst dashboard. LIME provides local fidelity by highlighting the few most influential features that pushed a single data point toward a malicious classification This integrated output moves the system beyond simple alerts, providing security analysts with clear, auditable evidence—visualized often through charts like SHAP plots or LIME weight displays—which enhances trust in the autonomous detection process, reduces false positive investigation time, and allows security teams to efficiently triage and respond to network threats.

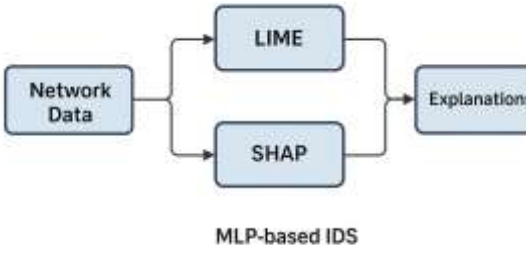

Fig 5.1: Structure of the Proposed System

V. IMPLEMENTATION

Fig 6.1: Admin Dashboard Page

Fig 6.2: Model Training Page

Fig 6.3: Dataset Page

Fig 6.4: Prediction inputs Page

Fig 6.5: Result Page

VI. CONCLUSION

The integration of Explainable AI into MLP-based Intrusion Detection Systems

addresses one of the most critical limitations of deep learning in cybersecurity: the lack of transparency. LIME and SHAP offer complementary strengths—LIME excels in generating quick, instance-specific insights, while SHAP provides consistent and globally interpretable feature importance measures.

By applying these methods to IDS, security analysts can better understand model behavior, verify the validity of alerts, and adapt detection strategies to emerging threats. However, the computational cost of SHAP and the potential instability of LIME in high-dimensional spaces must be carefully considered. Future work should explore optimization techniques for XAI methods, as well as real-time deployment strategies that balance interpretability with detection speed.

VII. FUTURE SCOPE

While this work focuses on the interpretability of MLP-based IDS using LIME and SHAP, several avenues remain for future exploration. One potential direction is the extension of this framework to more complex and real-time models, such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), or hybrid deep learning architectures. Additionally, exploring the performance of XAI tools under adversarial attack conditions can provide insights into the robustness of explanations.

Further research could also assess user-centric evaluation of explainability, where domain experts judge the usefulness of LIME and SHAP explanations in actual decision-making scenarios. Moreover, integrating explainable components into online learning or federated learning IDS architectures can address issues related to data privacy and continuous adaptation. Lastly, establishing standardized benchmarks for evaluating the fidelity and usability of XAI methods in cybersecurity contexts would enhance the comparability and applicability of future work in this field.

VIII. REFERENCES

[1] D. Gaspar, P. Silva, and C. Silva, “Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron,” *IEEE Access*, 2024. DOI: 10.1109/ACCESS.2024.3368377.

[2] B. Sharma, L. Sharma, C. Lal, and S. Roy, “Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach,” *Expert Systems with Applications*, vol. 232, 2023. DOI: 10.1016/j.eswa.2023.121751.

[3] P. Hermosilla, S. Berríos, and H. Allende-Cid, “Explainable AI for forensic analysis: A comparative study of SHAP and LIME in intrusion detection models,” *Applied Sciences*, vol. 15, no. 13, 2025. DOI: 10.3390/app15137329.

[4] V. Z. Mohale and I. C. Obagbuwa, “A systematic review on the integration of explainable artificial intelligence in intrusion detection systems,” *Frontiers in Artificial Intelligence*, 2025. DOI: 10.3389/frai.2025.1526221.

[5] L. Wang, Y. Zhang, and H. Li, “Explainable AI-based innovative hybrid ensemble model for intrusion detection systems,” *Journal of Cloud Computing*, vol. 13, 2024. DOI: 10.1186/s13677-024-00712-x.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016. DOI: 10.1145/2939672.2939778.

[7] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, 2017. DOI: 10.48550/arXiv.1705.07874.

[8] K. K. R. Choo, A. Dehghantanha, and R. M. Parizi, “Machine learning and intrusion detection systems: A survey,” *IEEE Access*, vol. 8, pp. 73127–73141, 2020.

DOI: 10.1109/ACCESS.2020.2988359.
[9] G. Yedukondalu, G. Bindu, and J. Pavan, "Intrusion detection system framework using machine learning," in *Proc. Int. Conf. Inventive Research in Computing Applications (ICIRCA)*, 2021.
DOI: 10.1109/ICIRCA51532.2021.9544717.
[10] S. Patil, V. Varadarajan, S. Mazhar, et al., "Explainable artificial intelligence for intrusion detection system," *Electronics*, vol. 11, no. 19, 2022.
DOI: 10.3390/electronics11193079.
[11] F. Ebrahimi et al., "Intrusion detection in the Internet of Things using explainable deep learning," *Cybersecurity*, 2025.
DOI: 10.1186/s42400-025-00369-2.
[12] R. Kalakoti, R. Vaarandi, H. Bahsi, and S. Nõmm, "Evaluating explainable AI for deep learning-based network intrusion detection system alert classification," *Proc. Int. Conf. Information Systems Security and Privacy*, 2025.
DOI: 10.5220/0012488300003687.
[13] M. Masum et al., "Bayesian hyperparameter optimization for deep neural network-based network intrusion detection," *Journal of Big Data*, 2022.
DOI: 10.1186/s40537-022-00697-5.
[14] V. Vimbi et al., "Interpreting artificial intelligence models: A systematic review of LIME and SHAP frameworks," *Artificial Intelligence Review*, 2024.
DOI: 10.1007/s10462-024-10725-4.
[15] P. Hermosilla et al., "Explainable AI techniques for cybersecurity and digital forensics," *Applied Sciences*, 2025.
DOI: 10.3390/app15137329.