International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

Explainable Artificial Intelligence (AI) for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron

Malyala Mahesh
(M.Tech Artificial Intelligence)
Aurora's Scientific and Technological Institute,Telangana,India
Email : maheshmalyalasssssss318@gmail.com
Dr.M. Sridhar
Head Of The Department Computer Science and Engineering
Aurora's Scientific and Technological Institute,Telangana,India
Email: msridhar.msr@gmail.com
Manipaul Panem
Aurora's Scientific and Technological Institute,Telangana,India
Email: manipal.panem@gmail.com

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), Multi-Layer Perceptron (MLP), LIME, SHAP, Cybersecurity, Network Security, Model Interpretability, Feature Importance Analysis, Attack Detection.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

I. 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.

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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

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 aims to enhance the transparency and trustworthiness of Multi-Layer Perceptron (MLP)-based Intrusion Detection Systems (IDS) by integrating Explainable Artificial Intelligence (XAI) techniques—specifically LIME and SHAP.

IV. SYSTEM ARCHITECTURE

The 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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

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. 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.

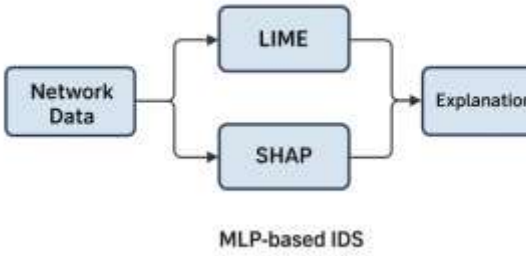

Fig 5.1: System Architecture

V. IMPLEMENTATION

Fig 6.1: Home Page

Fig 6.2: DatasetPreview

Fig 6.3: Model Training

Fig 6.4: Prediction Inputs

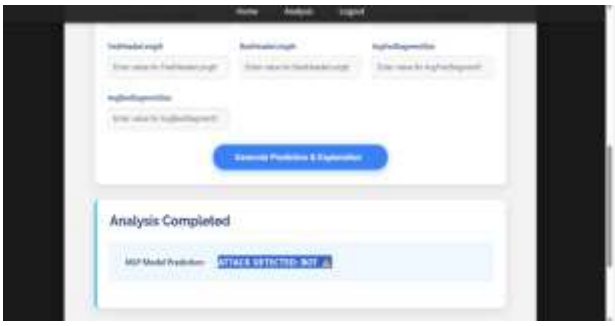

Fig 6.5: Result Page

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. M. A. Kadhim, A. Al-Hamami, and A. A. Hussein, “Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron,” IEEE Access, vol. 12, pp. 30164–30175, 2024. DOI: 10.1109/ACCESS.2024.3368377.

2. S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774, 2017. DOI: 10.48550/arXiv.1705.07874.

3. 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, pp. 1135–1144, 2016. DOI: 10.1145/2939672.2939778.

4. V. Z. Mohale and I. C. Obagbuwa, “A Systematic Review on the Integration of Explainable Artificial Intelligence in Intrusion Detection Systems to Enhance Transparency and Interpretability in Cybersecurity,” Frontiers in Artificial Intelligence, vol. 8, 2025. DOI: 10.3389/frai.2025.1526221.

    A. S. Khan et al., “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.

5. V. Nair et al., “Enhanced Intrusion Detection in Cybersecurity Through Dimensionality Reduction and

Explainable Artificial Intelligence," Scientific Reports, vol. 15, 2025. DOI: 10.1038/s41598-025-06761-9.

6. S. M. Lundberg et al., "From Local Explanations to Global Understanding with Explainable AI for Trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020. DOI: 10.1038/s42256-019-0138-9.

7. M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-Agnostic Interpretability of Machine Learning," arXiv preprint, 2016. DOI: 10.48550/arXiv.1606.05386.

8. F. Chollet, Deep Learning with Python, 2nd ed. Manning Publications, 2021. ISBN: 9781617296864.

9. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016. DOI: 10.5555/3086952.

10. Todupunuri, A. (2024). Explore How AI Can Be Used To Create Dynamic And Adaptive Fraud & Rules That Improve The Detection And Prevention Of Fraudulent & Activities In Digital Banking. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5014699

11. Babburi, S. Privacy-Preserving Collaborative Framework with Auditable Federated Learning.

12. Gaddam, S. (2024). Integrating machine learning models with continuous integration and continuous delivery (CI/CD) pipelines for a learning-driven approach to software engineering.

13. Immadi, S. K. (2025). Optimizing ERP for Human Capital Management. Applied Research for Growth, Innovation and Sustainable Impact, 377–384. https://doi.org/10.1201/9781003684657-63

14. Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.

15. Poojari, R. INTELLIGENT SYSTEMS+B108 AND APPLICATIONS IN ENGINEERING.

16. Vasagam, M. (2024, August 30). Ensuring security in modern data pipelines: Practical strategies for data engineers. International Journal of Intelligent Systems and Applications in Engineering, 12(22s), 2401.

17. Santthosh Saai Reddy Purmani. (2026). Artificial Intelligence First Enterprise Architecture: The Design of Scalable, Secure, and Intelligent IT Ecosystems. American Journal of AI Cyber Computing Management, 6(1(2)), 1–8. https://doi.org/10.64751/ajaccm.2026.v6.n1(2).pp1-8

18. Purmani, S. S. R. (2024). Aligning IT investment decisions with overall business strategy from an enterprise program management perspective, focusing on the integration of IT leadership in strategic decision-making processes. International Journal of Communication Networks and Information Security, 16(5), 1213–1219

19. Kumara, S. (2026, February). A Lightweight Deep Learning Based Classification Models for Non-Human Identity Threat Detection. In 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC) (pp. 1-6). IEEE.

20. Kotte, G. (2025). Overcoming Challenges and Driving Innovations in

API Design for High-Performance AI Applications. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5283649

21. Kotte, G. (2025). Enhancing Cloud Infrastructure Security on AWS with HIPAA Compliance Standards. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5283660

22. Ranjbareslamloo, S., Dzukeya, G. A., Muhit, M. M. I., & Qattawi, A. (2025). Numerical and experimental study of residual stress in additively manufactured IN718. Manufacturing Letters, 44, 915–927. https://doi.org/10.1016/j.mfglet.2025.915927

23. Viswanathan, V. (2023). AI-Augmented Decision Intelligence for Enterprise Systems: Integrating Cognitive Analytics for Resource and Talent Optimization.

24. Viswanathan, V. Generative AI for Smarter Workforce Planning and Enterprise Resource Decisions.

25. Mudusu, S. (2025). Health Insurance Fraud Detection: The Role Of Advanced It Systems In Preventing And Identifying Fraud. International Journal, 16(1), 3769-3777

26. Mudusu, S. K. (2026, April 15). The secure intelligence framework: Architecting AI systems for a data-driven world. CIO (Foundry Expert Contributor Network).

27. Agrawal, A. M., Gajula, S., Shinde, R. P., Shah, H., & Ghosh, H. (2025, July). Machine Translation for Long Sequences with Enhanced Attention Mechanisms. In 2025 5th International Conference on Electrical, Computer and Energy Technologies (ICECET) (pp. 1-6). IEEE.

28. Gajula, S. (2026, March). Two Pillars of Banking Intelligence: A Comparative Analysis of AI Techniques for Fraud Prevention and Churn Mitigation. In 2026 14th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-6). IEEE.

29. Maturi, S. Y. (2021). Blockbond hardening: Securing pooled-hash protocols against traffic tampering, MITM hash-rate hijacking, and template coercion. International Journal of Communication Networks and Information Security, 13(3), 718–728.

30. Maturi, S. Y. (2023). Crowdsourced frontier: Unveiling autonomous adversarial cybercapabilities via open AI competition. International Journal of Intelligent Systems and Applications in Engineering, 11(1s), 275–284.

31. Sikder, M. Z., Shakil, M. A. I., Ahad, A., Karim, M. F., Intakhab, B., & Islam, D. A. (2025, June). Microwave-Based Detection of Early-Stage Renal Cell Carcinoma Using UHF Range Antenna. In 2025 International Conference on Computer Systems and Technologies (CompSysTech) (pp. 1-6). IEEE.

32. Manoharan, D. (2024). Governance-Oriented Quality Engineering Framework for Healthcare EDI Modernization. International Journal of Multidisciplinary on Science and Management IJMSM, 1(2).

33. Manoharan, D. (2026). Advancing Healthcare EDI Interoperability Through Informatica Cloud B2B Gateway Quality Engineering. Available at SSRN 6385719.

34. Ravishankara, M. (2026, February). PlotChain: Deterministic Checkpointed Evaluation of Multimodal LLMs on Engineering Plot Reading. In SoutheastCon 2026 (pp. 1-8). IEEE.

35. Doragacharla, V. R. (2026). Building

Real-Time Pricing Systems for Modern Retail. Available at SSRN 6451760.

36. Adabala, P. K. (2024). Utilizing predictive analytics to improve efficiency and decision-making in ERP-connected supply chains. International Journal of Intelligent Systems and Applications in Engineering, 12(22s), 2465

37. Venkata Ramana, P. (2024). AI-driven predictive analytics in ERP systems for proactive supply chain optimization. International Journal of Research in Information Technology and Computing, 8(4).

38. Kavuri, S. (2026). An Explainable Machine Learning Framework for Predicting Software Defects in Large-Scale Software Systems. 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC), 1–6. https://doi.org/10.1109/icaic67076.2026.11395777

39. Srikanth Kavuri. (2025). AI-DRIVEN TEST AUTOMATION FRAMEWORKS: ENHANCING EFFICIENCY AND ACCURACY IN SOFTWARE QUALITY ASSURANCE. International Journal of Applied Mathematics, 38(10s), 699–710. https://doi.org/10.12732/ijam.v38i10s.990

40. Venkata Pavan Kumar Gummadi. (2023). MuleSoft Batch Processing: High-Volume Streaming Architecture. Computer Fraud and Security, 50–57. https://doi.org/10.52710/cfs.886

41. Venkata Pavan Kumar Gummadi. (2026). Infrastructure Optimization Techniques for Enterprise Integration Platforms: A Comprehensive Analysis. Computer Fraud and Security, 37–44. https://doi.org/10.52710/cfs.875

42. Shashank, A. (2025). Self-Healing Data Pipelines for Enhanced Reliability: A Paradigm Shift in Enterprise Data Management. Journal of Computer Science and Technology Studies, 7(8), 1097-1104.

43. Harshitha, G. K., Nandigama, C., & Thiripalu, P. (2026). An exploration into identification of opportunities and challenges of establishing and running an enterprise in the area of biofuels. Minnesota Journal of Business Law and Entrepreneurship, 2026(1), 1159–1168.

44. Ghali Krishna Harshitha & P. Thiripalu. (2025). Assessing the influence of age and gender on soft skills among emerging Gen Z HR professionals. Advances in Consumer Research, 2(2), 991–999.