Aravind Chagantipati, 2024, 12:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Architecting Post-Hoc Explainable Artificial Intelligence Frameworks for Verifiable, Transparent, and Auditable Deep Learning Network Intrusion Controls

Aravind Chagantipati
Department of Computer Science and Engineering, Kennedy University, Saint Lucia, France

Abstract- Although multi-layered neural architectures deliver exceptional precision in autonomous cyber threat identification, their complex internal mechanisms remain functionally opaque. This hidden operational layer compromises institutional trust, severely impacts the speed of post-incident forensic validation, and introduces regulatory compliance vulnerabilities under modern international data governance mandates. This study introduces an operational deployment strategy that embeds post-hoc Explainable AI (XAI) methodology—specifically leveraging Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME)—directly onto high-performance network anomaly classifiers. Comprehensive experimental assessments utilizing the NSL-KDD, CICIDS, and UNSW-NB15 reference datasets reveal that while the SHAP framework establishes rigorous global structural stability, the LIME tool yields rapid, localized diagnostic clarity for isolated alerts.

Keywords: Explainable AI (XAI), SHAP Framework, LIME Tool, Deep Black-Box Models, Security Threat Auditing, Transparency.

I. INTRODUCTION

The operational scale of modern security systems has expanded significantly through the deployment of complex Machine Learning models and deep neural architectures within Network Intrusion Detection Systems (NIDS). These mathematical models excel at detecting sophisticated modern cyber attacks and subtle telemetry variations without relying on manual rule generation. Nonetheless, because modern hidden layers—such as recurrent nodes, convolutional layers, or specialized long short-term memory networks—map intricate, non-linear relationships across highly expansive high-dimensional spaces, they act as functional black-boxes. Consequently, security teams routinely receive raw classification verdicts without an interpretable breakdown of the underlying feature weight dynamics that generated the alert.

In highly regulated enterprise ecosystems, this lack of transparency presents a critical barrier to production adoption. Forensic analysts must cross-examine automated alerts against strict infrastructure parameters before executing defensive isolation protocols. Opaque diagnostic outputs slow incident response times, complicate mandatory auditing tasks under current global compliance frameworks (such as GDPR), and expose underlying models to hidden adversarial exploitation. To bridge this structural vulnerability, this paper provides an implementation blueprint that couples SHAP and LIME post-hoc evaluation frameworks directly into operational deep learning network pipelines, successfully translating complex mathematical outputs into verifiable, human-readable forensic insights.

II. MATHEMATICAL FORMULATIONS OF POST-HOC XAI

SHAP Framework (Shapley Additive Explanations)

The SHAP diagnostic model utilizes axiomatic cooperative game theory to quantify individual feature contributions. Within this formulation, each metric vector from the network dataset is modeled as a participant in a cooperative game, where the total game payoff is represented by the continuous prediction value of the network classifier. Feature

importance is determined by calculating exact Shapley values, which ensure balanced attribution by taking the weighted average of marginal impacts across every possible feature subset configuration. The basic additive explanation model is formulated as:

$$g(z') = \varphi_0 + \sum_{i=1}^{M} q_i z_i$$

where g represents the localized proxy explanation model, z' ∈ {0,1}M denotes a binary array representing the specific feature coalition pattern, and φi signifies the exact calculated Shapley attribution coefficient assigned to feature element i.

LIME Framework (Local Interpretable Model-Agnostic Explanations)

The LIME methodology functions by constructing a simplified local approximation of the complex global neural model directly around a specific target data vector. This approach perturbs the localized target sample inputs, synthesizes a localized evaluation dataset of variations, and documents the corresponding output predictions from the black-box model. Subsequently, an inherently interpretable model (such as a regularized sparse linear regression) is trained on this localized, proximity-weighted dataset. The primary objective optimization equation is formalized as:

$$\xi(x) = arg\ min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

where L(f, g, πx) represents the local loss metric measuring the deviation of the simple explanation model g from the black-box classifier f under a distance proximity metric πx, and Ω(g) enforces a penalty constraint on the model complexity to ensure high human interpretability.

III. EXPERIMENTAL IMPLEMENTATION AND RESULTS

The experimental configuration integrated the SHAP and LIME explainer layers directly into the automated deep classification pipelines. Both explanation modules evaluated validation samples drawn from the standard NSL-KDD, CICIDS, and UNSW-NB15 datasets, extracting comprehensive global feature profiles alongside fast instance-specific decision pathways.

Table 2: Comparative Metric Breakdown between SHAP and LIME Explainer Layers

Operational Criteria	Shap Explainer Implementation	Lime Explainer Implementation
Explanation Scope Profile	Global System + Local Capability	Local Instance Focus Exclusive
Mathematical Consistency	High (Guaranteed Via Shapley Axioms)	Moderate (Dependent On Local Sampling)
Human Interpretability Index	High (Visual Force And Summary Graphs)	High (Local Rule Threshold Bars)
Computational Cost Footprint	High (Requires Full Power Combinations)	Moderate (Processes Local Perturbations Fast)
Primary Best Use Case	Global Feature Selection And Model Validation	Real-Time Instance Diagnostics And Alerts
Attribution Stability Profile	Stable Across Repeated Processing Runs	Slight Stochastic Variations Per Run

IV. FORENSIC ANALYSIS AND SECURITY ANALYST INSIGHTS

The empirical comparative assessments uncover distinct operational trade-offs inherent to each explainability technique. The SHAP framework yields exceptional global mathematical consistency, proving highly effective for off-line architectural validation, feature engineering, and model optimization. It empowers security infrastructure engineers to confidently prune low-impact or redundant telemetry features across the entire historical dataset. However, its combinatoric mathematical requirements generate a significant computational load, which can introduce processing latencies within high-throughput production environments.

In contrast, the LIME framework focuses strictly on localized data points, mapping individual data parameters swiftly by skipping complex global interactions. This property makes it well-suited for interactive, real-time operations centers, equipping threat analysts with immediate, contextual feature weights for isolated alerts (e.g., highlighting a specific 42% elevation in attack likelihood driven by a sharp variance in packet_size coupled with an irregular protocol_type identifier).

V. CONCLUSION

This study demonstrates that the deployment of unified SHAP and LIME post-hoc analytical frameworks successfully deconstructs the opaque nature of deep neural intrusion classifiers. This layered interpretability maps abstract mathematical vectors into logical, highly auditable patterns that operators can verify. Future research paths will focus on the acceleration of real-time local explanation modules to seamlessly sustain line-rate detection across high-speed multi-gigabit routing environments.

REFERENCES

1. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.
2. 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, 2016, pp. 1135–1144.
3. L. H. Gilpin, D. Bau, B. Z. Yuan, J. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 2018, pp. 80–89.
4. W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and justifying deep learning models," IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 39–48, 2017.