

Integrating Post-Hoc Explainable Artificial Intelligence (XAI) Frameworks for Transparent, Accountable, and Auditable Deep Learning Network Intrusion Controls

Aravind Chagantipati

Department of Computer Science and Engineering, Kennedy University

ABSTRACT

While deep neural networks achieve excellent classification rates in automated threat parsing, their multi-layered hidden structures operate as opaque black boxes. This paper presents an implementation framework that integrates post-hoc Explainable AI (XAI) algorithms—specifically SHAP and LIME—onto optimized deep network intrusion classifiers.

1. INTRODUCTION

The deployment of high-capacity Machine Learning and multi-layered Deep Neural Networks has dramatically improved classification metrics inside automated Network Intrusion Detection Systems (NIDS)...

2. MATHEMATICAL FORMULATIONS OF POST-HOC XAI

2.1 SHAP Framework (Shapley Additive Explanations)

The SHAP framework relies on cooperative game theory to compute feature contributions.

$$g(z') = \phi_0 + \sum \phi_i z'_i$$

2.2 LIME Framework (Local Interpretable Model-Agnostic Explanations)

LIME works by approximating the complex global model locally around a specific target data point.

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

3. EXPERIMENTAL IMPLEMENTATION AND RESULTS

We integrated SHAP and LIME explainer layers into our deep network pipelines.

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

Operational Criteria | SHAP | LIME

Explanation Scope | Global + Local | Local

Mathematical Consistency | High | Moderate

Human Interpretability | High | High

Computational Cost | High | Moderate

Primary Use Case | Validation | Real-time Diagnostics

Attribution Stability | Stable | Slightly Stochastic

4. FORENSIC ANALYSIS AND SECURITY ANALYST INSIGHTS

SHAP provides strong global consistency, while LIME focuses on fast local explanations.

5. CONCLUSION

Combining SHAP and LIME post-hoc layers effectively opens the black-box of deep network intrusion models.

REFERENCES

- [1] Lundberg and Lee, 2017.
- [2] Ribeiro, Singh, and Guestrin, 2016.
- [3] Gilpin et al., 2018.
- [4] Samek et al., 2017.