Explainable AI for Intrusion Detection Systems: Challenges and Opportunities

Jisung Brayden Rhee
Department of Computer Science
Georgia State University
Atlanta, GA, USA
jrhee5@student.gsu.edu

***Abstract*—This paper addresses the explainability challenges inherent in modern AI-driven Intrusion Detection Systems (AI-IDS). While these systems have demonstrated strong performance in detecting novel threats, the rationale behind their alerts often remains opaque. To enhance interpretability, we examine Explainable Artificial Intelligence (XAI) techniques, specifically, Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) and evaluate their effectiveness in clarifying model outputs. By comparing their functionality, computational efficiency, and integration within real-world intrusion detection architectures, we review the respective strengths and limitations of each method. Findings from prior studies suggest that combining local and global explanation techniques can significantly improve the transparency, auditability, and operational trustworthiness of AI-IDS frameworks.**

***Index Terms*—Cloud Security, Cybersecurity, Explainable AI, IDS, LIME, Model Interpretability, Network Security, SHAP.**

I. Introduction

As cloud systems grow more complex, attackers continue to exploit new vulnerabilities, outpacing traditional defenses. Static and rule-based systems like signature-based intrusion detection (SIDS) struggle to identify novel attacks. To fill this gap, AI-driven intrusion detection systems (AI-IDS) are gaining traction for their ability to adapt and respond to unfamiliar threats in real-time. However, despite their effectiveness, AI-IDS face a major challenge, explainability. These systems often generate alerts without offering clear insight into how or why a threat was flagged. This opacity complicates the work of security analysts, who must interpret and act on alerts with confidence and speed. This paper analyzes the strengths and limitations of AI-IDS with a particular focus on explainability. It also examines Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), the core tools of explainable artificial intelligence (XAI), and explores the potential impact of XAI-IDS on cybersecurity outcomes [1].

II. Limitations of Traditional AI-IDS

A. Benefits of AI-IDS

AI-IDS use machine learning to monitor network traffic and flag abnormal activity. Unlike SIDS, which rely on predefined attack patterns, AI-IDS adapt by learning from historical data. In their study presented at the Proceedings of the 16th International Conference on Security of Information and Networks (SIN), U. Upadhyay et al. highlight that "machine learning and AI-driven IDS solutions are gaining traction, leveraging algorithms that can adapt and learn from historical data to identify new attack patterns" [2]. This adaptability gives AI-IDS a clear advantage in detecting zero-day and polymorphic threats. The Center for Internet Security, a nonprofit organization founded in 2000 and focused on cyber threats, expressed a neutral view of SIDS, stating the cons that "signature-based detection relies on a preprogrammed list of known indicators of compromise (IOCs)…. However, signature-based security systems will not detect zero-day exploits" [3]. This limitation exists because SIDS rely on static rule sets.

B. Limitations and Explainabilities

In contrast, AI-IDS recognize emerging behaviors and flag threats based on real-time deviations from normal patterns. However, explainability remains a core limitation. Many AI-IDS function like black boxes—systems that produce alerts without revealing how or why those decisions were made. Upadhyay et al. emphasize that "lack of interpretability can lead to skepticism and hinder decision-making efficacy" [2]. This lack of transparency delays response, complicates auditing, and undermines trust in automated systems.

To address this issue, researchers have begun embedding explainability into AI-IDS. Writing in Applied Sciences, O. Arreche et al. argue that "It is essential to provide explanations of such models and accompanied features (traffic log parameters) and labels (attack types)" [1]. This shift signals a new standard in AI-IDS, where interpretability is no longer treated as something to sacrifice, but as a strength that contributes to effective security.

III. Explainable AI Tools for IDS

A. XAI-IDS, SHAP and LIME

XAI plays a pivotal role in addressing the opacity of AI-IDS. Unlike traditional models that merely output binary alerts, XAI provides transparency by generating interpretable insights into why specific traffic is flagged. At the core of this transformation are SHAP and LIME, two key tools within the XAI-IDS framework.

Rather than relying solely on theoretical claims, Arreche et al. demonstrate that SHAP enables analysts to visualize how input features contribute to alerts by ranking their importance and explaining their role in both global and local contexts [1]. SHAP computes these contributions using Shapley values from cooperative game theory, helping analysts pinpoint which features were decisive in triggering specific alerts, even in encrypted traffic. Fig. 1 visualizes the conceptual difference between SHAP's global attribution and LIME's instance-specific modeling.

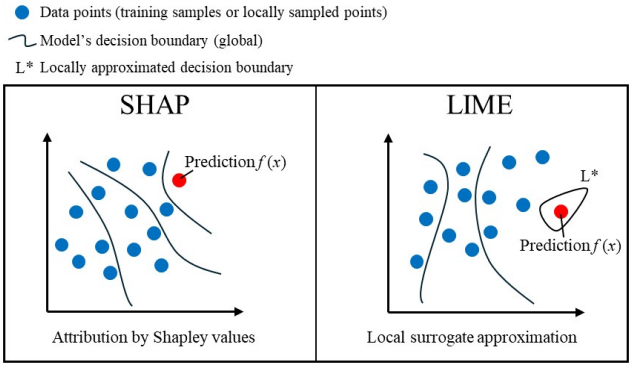

Fig. 1. Comparison of SHAP and LIME explanation methods. SHAP attributes the prediction $f(x)$ to each feature using Shapley values based on the global decision boundary. In contrast, LIME constructs a locally interpretable surrogate model L^* around the instance to approximate the prediction.

LIME complements SHAP by creating a simple, interpretable model that mimics the behavior of the original complex model around a specific prediction. In other words, when a model is mathematically too complex to explain directly, LIME generates a simplified version that behaves similarly and presents it to analysts, helping them better understand security events. Arreche et al. reported that "Overall, SHAP is more time-efficient for global explanations, whereas LIME is better in local explanation scenarios across all tested AI models" [1]. Together, these tools bridge the gap between algorithmic output and human reasoning, offering actionable insights that improve trust, usability, and auditability in operational environments. Upadhyay et al. also describe LIME as suited for instance-level interpretation and SHAP for feature-level attribution, implying their complementary strengths in building XAI-IDS [2]. Table I provides a comparative summary of the key differences between SHAP and LIME in terms of scope, approach, cost, and usability.

TABLE I. COMPARATIVE SUMMARY OF SHAP AND LIME

	SHAP	**LIME**
Explanation Scope	Global, Local	Local
Approach	Game theory-based[a]	Linear approximation
Intuitiveness	Moderate	High
Computational Cost	Medium to High	Low
Strengths	Quantitative, precise contributions	Fast, interpretable
Weaknesses	Slow	Inconsistent on repeated runs

[a] Shapley values

B. *Integrated XAI-IDS Framework*

Building on this foundation, Arreche et al. proposed an integrated XAI-IDS framework that embeds SHAP and LIME into a multi-model architecture utilizing classifiers such as deep neural networks, and XGBoost. The system was validated using widely recognized intrusion detection benchmarks like NSL-KDD, and RoEduNet-SIMARGL2021. Each contains diverse attack types ranging from DoS and brute force to infiltration and botnet activity.

This study found that the explanations generated by XAI-IDS helped analysts disambiguate overlapping anomalies and assign threat severity levels more effectively. K. Fatema et al. effectively support the use of SHAP in IDS by stating that "employing SHAP in Intrusion Detection Systems aids researchers and practitioners in identifying significant attack patterns, reducing false positives, and enhancing model robustness" [4]. The framework also maintained low computational overhead, making it suitable for real-time deployment in operational cloud environments. Arreche et al. state, "Notably, the study shows that the use of XAI methods results in negligible additional time overhead" [1]. These findings affirm that XAI not only enhances interpretability but does so efficiently and without compromising performance.

IV. CONCLUSION

Incorporating XAI into IDS represents a broader paradigm shift in cloud security architecture. Detection alone is no longer sufficient, systems must also be able to explain why something is flagged. Transparent alerts help build trust in AI-made decisions, streamline incident response, and support legal and regulatory compliance. This emphasis on interpretability aligns with the evolving standards of responsible AI, especially in sensitive and high-stakes fields such as cloud security. Fatema et al. support this by stating, "We may move closer to a cybersecurity environment that is more safe, considerate of privacy, and interpretable by further improving and developing federated explainable AI-based IDS solutions" [4]. XAI-IDS represents a meaningful step toward building intelligent and transparent cybersecurity systems that match the growing complexity of today's digital infrastructure.

REFERENCES

[1] O. Arreche, T. Guntur, and M. Abdallah, "XAI-IDS: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems," *Applied Sciences*, vol. 14, no. 10, p. 4170, 2024, doi:10.3390/app14104170.

[2] U. Upadhyay, A. Kumar, S. Roy, U. Rawat, and S. Chaurasia, "Defending the cloud: Understanding the role of explainable AI in intrusion detection systems," in *Proc. 16th Int. Conf. Security of Information and Networks (SIN), Jaipur, India, Nov*. 2023, pp. 1–9, doi:10.1109/SIN60469.2023.10475080.

[3] Center for Internet Security, "Election security spotlight – Signature-based vs anomaly-based detection," 2020. [Online]. Available: https://www.cisecurity.org/insights/spotlight/cybersecurity-spotlight-signature-based-vs-anomaly-based-detection (accessed: May 27, 2025).

[4] K. Fatema, S. K. Dey, M. Anannya, R. T. Khan, M. Rashid, S. Chunhua, and R. Mazumder, "Federated XAI IDS: An explainable and safeguarding privacy approach to detect intrusion combining federated learning and SHAP," *Preprints*, 2025, doi: 10.20944/preprints202503.1902.v1.