BLExID: Blockchain-based Novel Collaborative LLM for Explainable Adaptive Intrusion Detection

Md Tayeb Adnan, Paul Angelo Oroceo, Jae-Min Lee, Dong-Seong Kim
Networked Systems Laboratory, Department of IT Convergence Engineering,
Kumoh National Institute of Technology, Gumi, South Korea.
(mdtayebadnan, oroceopaul, ljmpaul, dskim)@kumoh.ac.kr

***Abstract*—Next-generation networks pose significant security challenges due to their dynamic and decentralized nature. Traditional intrusion detection systems often fail to adapt to evolving threats while ensuring transparency and collaboration. This paper proposes a blockchain-enabled collaborative framework employing a lightweight Large Language Model (LLM) for context-aware intrusion detection and Gaussian Mixture Models (GMM) for clustering unknown attacks. SHAP-based explainability provides transparent decision insights, while a permissioned blockchain secures decentralized coordination and model updates. Experimental results demonstrate high accuracy, adaptability to novel threats, and robust performance, validating the framework's applicability to real-world network environments.**

***Index Terms*—Blockchain, PureChain, Intrusion Detection, LLM, IoT**

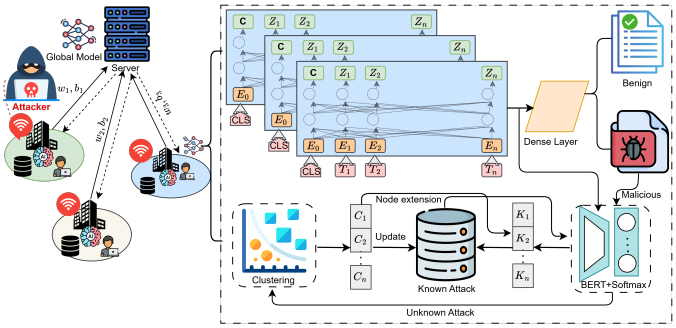

Fig. 1: Overview of the proposed BLExID architecture for Explainable Adaptive Intrusion Detection using Lightweight Large Language Models (LLM) and Blockchain

I. Introduction

The rapid evolution of next-generation networks—including 6G, IoT ecosystems, edge computing infrastructures, and cyber-physical systems—has significantly expanded the attack surface for cyber threats [1], [2]. Characterized by high device heterogeneity [3], [4], massive data flows, and stringent real-time requirements, these networks present complex challenges for securing digital assets. Traditional Intrusion Detection Systems (IDSs), often based on static rules, predefined signatures, or isolated machine learning models, increasingly fail to detect emerging sophisticated threats in such dynamic and decentralized environments.

Recent advances in deep learning and transformer-based models [5]–[8] have improved intrusion detection performance. Nonetheless, these models are typically trained offline on static datasets, limiting adaptability to novel attacks without full retraining. Additionally, most existing approaches lack explainability and collaborative intelligence, which are crucial for trust and auditability in distributed, mission-critical systems.

To address these challenges, we propose a blockchain-based collaborative intrusion detection framework integrating a lightweight Large Language Model (LLM). **Our key contributions are summarized are:** (i) A modular, adaptive IDS framework leveraging a compressed BERT-based LLM to detect and classify known attack patterns, (ii) A GMM-based clustering mechanism for identifying previously unseen attack types, enabling continuous learning through dynamic classifier expansion, (iii) Integration of SHAP-based explainability for interpretable intrusion decision justification, and (iv) A blockchain-powered collaboration mechanism providing verifiable, tamper-resistant, decentralized model updates and decision audits across IDS nodes.

II. Proposed Framework

As depicted in Figure 1, the proposed architecture is composed of four key modules: (i) data acquisition and preprocessing, (ii) decentralized binary detection, (iii) explainable attack classification, and (iv) adaptive model updating through blockchain coordination.

A. Lightweight LLM-Based Detection

The framework's first stage uses a compressed BERT model for binary classification, where each node locally distinguishes malicious from benign flows. Decisions are shared via a permissioned blockchain; smart contracts ensure integrity and consensus-based aggregation, providing provenance and resistance to poisoning. Formally, the model $f_\theta : \mathbb{R}^d \to [0, 1]$ with parameters θ predicts label $\hat{y}_i = 1$ if $\sigma(f_\theta(x_i)) > \tau$, and 0 otherwise, where σ is the sigmoid function and $\tau \in (0, 1)$ is a threshold. Malicious samples ($\hat{y}_i = 1$) proceed to classification.

Malicious flows are classified into known attack classes $\mathcal{K} = \{K_1, \ldots, K_n\}$ by encoding x_i into $z_i \in \mathbb{R}^h$ via encoder ϕ, followed by softmax classification. SHAP values approximate predictions as $f_\theta(x_i) \approx \phi_0 + \sum_j \psi_j(x_i)$, where ψ_j measures feature contributions for explainability.

Unknown embeddings $\mathcal{U} = \{z_k\}$ are modeled by a Gaussian mixture with m clusters representing novel attacks $\mathcal{C} = \{C_1, \ldots, C_m\}$. The number of clusters is determined by the

TABLE I: Performance Comparison of existing methods on CICIDS2017 and NSL-KDD Datasets

Method	CICIDS2017					NSL-KDD				
	Accuracy	Precision	Recall	F1-Score	AUC	Accuracy	Precision	Recall	F1-Score	AUC
BiLSTM	0.970	0.972	0.978	0.975	0.977	0.978	0.979	0.981	0.976	0.978
CNN-LSTM	0.975	0.973	0.970	0.971	0.973	0.972	0.975	0.974	0.974	0.975
LSTM-DNN	0.975	0.976	0.979	0.977	0.979	0.978	0.981	0.983	0.980	0.981
Multi-MLP	0.974	0.973	0.976	0.974	0.975	0.975	0.977	0.978	0.976	0.977
Proposed	**0.979**	**0.982**	**0.985**	**0.983**	**0.986**	**0.984**	**0.986**	**0.988**	**0.985**	**0.987**

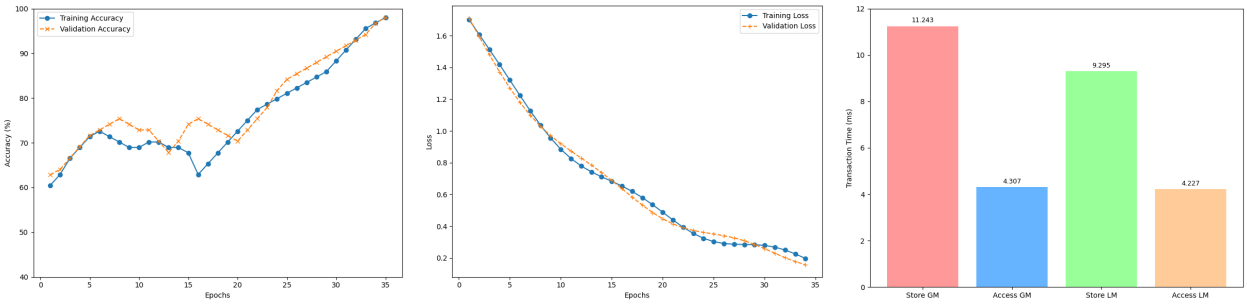

Fig. 2: Experimental results on CICIDS2017. Training and validation accuracy over epochs (left), Training and validation loss curves over epochs (middle), Transaction time for storing and accessing Global Model (GM) and Local Model (LM) (right).

silhouette coefficient $S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, balancing intra- and inter-cluster distances $a(i), b(i)$. These clusters expand the classifier output from n to $n+m$, followed by incremental retraining. Model updates are synchronized via blockchain, ensuring secure and tamper-proof coordination.

III. Experimental Results and Discussion

The proposed framework was tested on the CICIDS2017 and NSL-KDD datasets.

Figure 2 shows steady improvements in training and validation accuracy over 35 epochs on CICIDS2017, with aligned loss curves indicating strong generalization and optimization stability. Transaction times for storing and retrieving Local Models (LM) and Global Models (GM) via blockchain remain within acceptable limits, despite higher latency for GM due to consensus overhead.

Table II compares our model against state-of-the-art IDS architectures. On CICIDS2017, it achieves accuracy of 0.979, precision 0.982, recall 0.985, F1-score 0.983, and AUC 0.986, outperforming baselines. On NSL-KDD, it attains similarly superior metrics, demonstrating robust generalizability across diverse datasets and network environments.

IV. Conclusion

This study proposes a collaborative, explainable intrusion detection framework using a lightweight language model, Gaussian Mixture Models, and blockchain for secure coordination. SHAP-based explainability enhances trust. Experiments show high accuracy and adaptability for real-world decentralized environments.

Acknowledgment

This work was partly supported by Innovative Human Resource Development for Local Intellectualization program through the IITP grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201612, 25%) and by Priority Research Centers Program through the NRF funded by the MEST (2018R1A6A1A03024003, 25%) and by the MSIT, Korea, under the ITRC support program (IITP-2025-RS-2024-00438430, 25%), by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ICAN (ICT Challenge and Advanced Network of HRD) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2025-RS-2022-00156394, 25%).

References

[1] S. K. Reddy Mallidi and R. R. Ramisetty, "Optimizing intrusion detection for iot: A systematic review of machine learning and deep learning approaches with feature selection and data balancing," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 15, no. 2, p. e70008, 2025.

[2] A. Tayeb, M. Alam, A. Khatun, G. Mohtasin, M. F. Rahaman, M. J. A. Shanto, and D.-S. Kim, "Pureparking: A decentralized, secure framework for parking space sharing using blockchain," 06 2024.

[3] M. Golam, A. M. Tayeb, M. A. Khatun, M. F. Rahaman, A. Aouto, O. P. Angelo, D.-S. Kim, J.-M. Lee, and J.-H. Kim, "Blackicenet: Explainable ai-enhanced multimodal for black ice detection to prevent accident in intelligent vehicles," *IEEE Internet of Things Journal*, pp. 1–1, 2025.

[4] A. Md Tayeb and T.-H. Kim, "Unestformer: Enhancing decoders and skip connections with nested transformers for medical image segmentation," *IEEE Access*, vol. 12, pp. 190996–191009, 2024.

[5] Y. Li, Z. Xiang, N. D. Bastian, D. Song, and B. Li, "IDS-agent: An LLM agent for explainable intrusion detection in iot networks," in *NeurIPS 2024 Workshop on Open-World Agents*, 2024.

[6] A. M. Tayeb, J.-M. Lee, and D.-S. Kim, "Hiclassgen: High-resolution image augmentation with class and shape controllable diffusion models," in *2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 0814–0819, 2025.

[7] H. Dong and I. Kotenko, "Cybersecurity in the ai era: analyzing the impact of machine learning on intrusion detection," *Knowledge and Information Systems*, vol. 67, pp. 3915–3966, 2025.

[8] A. M. Tayeb, H. L. Nakayiza, H. Shin, S. Lee, J.-M. Lee, and D.-S. Kim, "Defectdiffusion: A generative diffusion model for robust data augmentation in industrial defect detection," in *2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 0066–0071, 2025.