SDN-Xplain: LLM-Driven Anomaly Detection and Human-Readable Explanation

Khan Hafiz Muhammad Ahmad Raza*, Alex Olazabal Dominguez†, Chanseo Park*,
Shakuova Leila*, Mose Gu*, Jaehoon (Paul) Jeong*, and Tae (Tom) Oh‡
*Sungkyunkwan University, Suwon, Republic of Korea
†Universidad de Deusto, Bilbao, Spain
‡School of Information, Rochester Institute of Technology, Rochester, NY, USA
Email: {030626, cstim, rna0415, pauljeong}@g.skku.edu, alex.o@opendeusto.es,
leyla.shakuova14@gmail.com, thoics@rit.edu

***Abstract*—Software-Defined Networking (SDN) offers programmability and centralized control but faces persistent challenges from anomalies such as intrusion attempts, high-rate flows, and protocol misuse. Traditional ML-based detectors can classify anomalous traffic patterns but provide no interpretability [1]–[4].**

This paper presents SDN-Xplain, a lightweight framework that detects anomalies using machine learning models trained on publicly available SDN traffic datasets and generates natural-language explanations with a Large Language Model (LLM).

Although our implementation does not include a live SDN controller or real-time topology, SDN-Xplain demonstrates a clear proof of concept: combining ML prediction with LLM-based reasoning significantly improves the interpretability of anomaly alerts and supports human-centric network analysis.

***Index Terms*—Software-Defined Networking (SDN), machine learning, Large Language Models (LLMs), network security, SHAP**

I. Introduction

Software-Defined Networking (SDN) decouples the control and data planes, offering global visibility and centralized management. This architectural shift, however, makes SDN traffic analysis increasingly complex. Machine learning models are widely used to detect anomalous or malicious flows [4]–[6], yet they behave as black-box classifiers: they output a label but provide no justification [7], [8]. Network operators must interpret these alerts manually, which is inefficient and error-prone.

Large Language Models (LLMs) provide an opportunity to improve interpretability [9], [10]. By generating context-aware explanations for detected anomalies, LLMs can bridge the cognitive gap between automated detection and human understanding. This paper introduces SDN-Xplain, a simple yet effective framework that focuses on two tasks: 1) anomaly detection using ML models trained on SDN traffic datasets, and 2) natural-language explanation of anomalies using an LLM. Our implementation is dataset-driven rather than deployed on a live SDN topology, but it demonstrates that LLM-based reasoning can meaningfully enhance anomaly interpretability even in offline or simulated analysis pipelines.

II. Related Work

Existing approaches to network anomaly detection rely heavily on machine learning classifiers [2], [4], [6] such as Random Forest, Support Vector Machines, and gradient boosting methods [1]. While these models achieve high accuracy, their decision-making process remains opaque. Recent work has explored explainable AI (XAI) techniques, particularly SHAP (SHapley Additive exPlanations) values [7], [11], to quantify feature contributions [11]. However, raw SHAP values require domain expertise [12] to interpret, limiting their practical utility for network operators.

The integration of LLMs for network security analysis is an emerging area. Previous studies have used LLMs for log analysis and threat intelligence [13], but few have combined ML-based detection with LLM-generated explanations in a unified framework [6], [9], [10], [13]. SDN-Xplain addresses this gap by providing an end-to-end pipeline that transforms technical ML outputs into actionable, human-readable insights.

III. Proposed Framework Design

SDN-Xplain is designed to be modular and adaptable. It consists of three primary components: Preprocessing, Classification via XGBoost, and Explanation Generation.

A. System Architecture

As illustrated in Fig. 1, the pipeline begins with the ingestion of network flow records. Each record, containing 41 distinct features, undergoes numerical encoding before being processed by the XGBoost classifier. Upon detection of an anomaly, the SHAP module calculates the contribution of each feature to the prediction. These contributions, along with the predicted class and confidence score, are formatted into a structured prompt for the LLM.

B. Anomaly Detection and XGBoost

We utilize XGBoost (Extreme Gradient Boosting), which demonstrates superior performance in handling imbalanced network security datasets [14]. The model is trained on the KDD Cup 1999 dataset [15], encompassing 33 classes including Denial of Service (DoS), Probing, Remote-to-Local (R2L), and User-to-Root (U2R). XGBoost's ability to model non-linear interactions allows it to outperform traditional models in both macro-F1 score (0.75) and inference speed (0.01s).

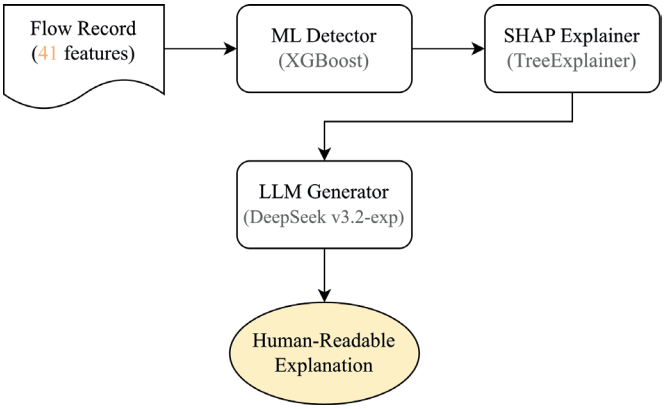

Fig. 1. Overall SDN-Xplain pipeline: Integrating ML-based detection, SHAP feature attribution, and LLM-driven explanation generation.

C. Explainability via SHAP and LLM

To demystify the XGBoost model, we employ the SHAP TreeExplainer [11]. SHAP values assign each feature an importance value by calculating its marginal contribution. SDN-Xplain identifies the "Top-K" influential features and feeds them into the LLM module.

The LLM module, powered by DeepSeek v3.2-exp, consumes this data through a prompt containing technical semantic meanings (e.g., mapping "count" to "number of connections to the same host"). This enables the LLM to generate a narrative explaining *how* specific features led to the detection of an attack, such as a "portsweep".

IV. Performance Evaluation

A. Classification Results

SDN-Xplain was evaluated on 4,509 test samples from the KDD Cup 1999 dataset, achieving an overall accuracy of 98.8% and a weighted F1-score of 0.99. Table I details the classification performance for selected attack categories.

TABLE I
Classification Performance for Selected Classes (Table II)

Attack Type	Precision	Recall	F1-Score	Support
normal	0.99	1.00	0.99	2,249
neptune	1.00	1.00	1.00	1,331
portsweep	0.96	0.99	0.97	74
satan	0.96	0.97	0.97	140
smurf	1.00	1.00	1.00	108
ipsweep	0.95	1.00	0.97	96

The model shows exceptional robustness in identifying high-volume attacks like *neptune*. For minority classes like *portsweep*, the model maintained high recall (0.99), ensuring critical threats are not missed.

B. Qualitative Analysis of Explanations

For a *portsweep* anomaly, the framework identified *dst_host_srv_count* and *same_src_port_rate* as primary contributors. The LLM synthesized this: "The system detected a portsweep attack with 99.8% confidence. This is driven by an abnormal increase in services accessed on the destination host from a single source port, characteristic of systematic scanning." This allows operators to prioritize firewall updates immediately.

V. Conclusion

SDN-Xplain successfully integrates XGBoost, SHAP, and LLMs to create an explainable anomaly detection system. By achieving 98.8% accuracy and providing human-centric insights, it addresses the "black-box" challenge in network security.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No. RS-2024-00398199) and the National Research Foundation of Korea (NRF) grant (No. 2023R1A2C2002990) funded by the Korea government (MSIT). Jaehoon (Paul) Jeong is the corresponding author.

References

[1] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, pp. 1–22, 2019.
[2] J. Lansky, S. Ali, M. Mohammadi, M. K. Majeed, S. H. T. Karim, S. Rashidi, M. Hosseinzadeh, and A. M. Rahmani, "Deep learning-based intrusion detection systems: a systematic review," *IEEE Access*, vol. 9, pp. 101 574–101 599, 2021.
[3] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, "A survey of deep learning-based network anomaly detection," *Cluster Computing*, vol. 22, no. Suppl 1, pp. 949–961, 2019.
[4] S. Wang, J. F. Balarezo, S. Kandeepan, A. Al-Hourani, K. G. Chavez, and B. Rubinstein, "Machine learning in network anomaly detection: A survey," *IEEe Access*, vol. 9, pp. 152 379–152 396, 2021.
[5] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *Journal of network and computer applications*, vol. 36, no. 1, pp. 16–24, 2013.
[6] M. Zolanvari, M. A. Teixeira, L. Gupta, K. M. Khan, and R. Jain, "Machine learning-based network vulnerability analysis of industrial internet of things," *IEEE internet of things journal*, vol. 6, no. 4, pp. 6822–6834, 2019.
[7] F. Yan, S. Wen, S. Nepal, C. Paris, and Y. Xiang, "Explainable machine learning in cybersecurity: A survey," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 12 305–12 334, 2022.
[8] M. Wang, K. Zheng, Y. Yang, and X. Wang, "An explainable machine learning framework for intrusion detection systems," *IEEE Access*, vol. 8, pp. 73 127–73 141, 2020.
[9] H. F. Atlam, "Llms in cyber security: Bridging practice and education," *Big Data and Cognitive Computing*, vol. 9, no. 7, p. 184, 2025.
[10] H. Xu, S. Wang, N. Li, K. Wang, Y. Zhao, K. Chen, T. Yu, Y. Liu, and H. Wang, "Large language models for cyber security: A systematic literature review," *ACM Transactions on Software Engineering and Methodology*, 2024.
[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
[12] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.
[13] V. Jüttner, M. Grimmer, and E. Buchmann, "Chatids: Explainable cybersecurity using generative ai," *arXiv preprint arXiv:2306.14504*, 2023.
[14] M. Nalluri, M. Pentela, and N. R. Eluri, "A scalable tree boosting system: Xg boost," *Int. J. Res. Stud. Sci. Eng. Technol*, vol. 7, no. 12, pp. 36–51, 2020.
[15] KDD, "The kdd cup 1999 dataset," 1999, available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.