Explainable Transformer-Based Zero-Day Attack Detection in Next-Generation Networks: A Comprehensive Literature Review

Sanjith S[1*], Dr. S. K. Manju Bargavi[2],

[1]Postdoctoral Fellow, Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia.

[2]Professor, School of Computer Science and IT, Jain Deemed-to-be University, Bangalore 560069, Karnataka, India.

Abstract

Network Intrusion Detection Systems (IDSs) are a critical part of cybersecurity. However, there are still great challenges for IDSs to detect zero-day attacks and provide transparent and reliable decision support. Deep learning architectures based on transformers have achieved detection accuracies higher than 99% on several benchmark intrusion detection datasets. But the black-box nature of these models limits their practical use and raises concerns about interpretability, regulatory compliance, and user trust. We explore three main research directions: (1) performance of detection on benchmark datasets such as NSL-KDD, CICIDS2017, UNSW-NB15, CICIoT2023, and ToN-IoT; (2) explainability techniques, including SHAP, LIME, attention visualization, DeepLIFT, and Integrated Gradients; and (3) robustness against adversarial evasion attacks. Results show that hybrid supervised/unsupervised Transformer models outperform single model approaches consistently with some studies reporting improvements up to 18.6% in zero-day attack detection. Furthermore, the adversarial training can achieve the detection accuracies above 94% in the adversarial attack cases. Attention-based mechanisms offer inherent model interpretability, eliminating the necessity of post hoc explanation techniques. However, it remains challenging to generalize across datasets, and performance drops by roughly 10% to 40% when models are evaluated in network settings that are different from training. This review presents a practical deployment architecture based on the synthesized evidence, including Transformer inference, dual-method explainability (SHAP and LIME), and decision support at the Security Operations Center (SOC). It also discusses some open research challenges like standardized evaluation metrics for explainability, realistic benchmarks for adversarial robustness, and privacy-preserving federated learning frameworks for collaborative threat intelligence.

Keywords: Intrusion Detection System; Transformer; Explainable Artificial Intelligence; Zero-Day Attacks; Network Security; SHAP; LIME; Adversarial Robustness; Self-Attention Mechanism; Deep Learning; Cybersecurity; Edge-Cloud Security; Federated Learning

1. Introduction and Motivation

The explosion of digital connectivity by the Internet of Things (IoT) devices, fifth and sixth-generation (5G/6G) networks, and cloud-edge computing paradigms has radically changed the cybersecurity landscape [1]. These technologies hold immense promise for innovation and efficiency, but they also introduce sophisticated attack vectors that traditional security mechanisms are not equipped to handle effectively. Intrusion Detection Systems (IDS) are the backbone of network defense architectures and are facing increasing challenges in detecting zero-day attacks, which are exploits that target previously unknown vulnerabilities with no signatures available [2]. On average, modern cyber threats go undetected for 312 days, indicating a crucial urgency of advanced detection methodologies [3].

2. Cybersecurity Challenges in 5G/6G, IoT, SDN, Edge, and Cloud Environments

2.1 Fifth-Generation and Sixth-Generation Network Vulnerabilities

Fifth and sixth generation networks are architecturally complex and the attack surface grows dramatically. The 5G and 6G paradigm involves Network Function Virtualization (NFV), network slicing and software-defined networking (SDN) which improve network agility but generate new security risks [4]. Network slicing allows for logical isolation of network resources, but, incorrect implementation might allow for cross-slice attacks. The centralized control architecture of SDN provides programmability advantages, but also centralizes the security vulnerabilities around the controller plane, which is a high value target for adversarial actions [5].

3. Zero-Day Attack Detection Methods

3.1 Machine Learning and Statistical Approaches

Statistical outliers for zero-day attacks are found using unsupervised learning techniques such as Isolation Forest and One-Class SVM by analyzing the divergence from the distribution of normal behavior [7]. Isolation Forest is computationally efficient, as it builds a random tree ensemble, reducing the time complexity to O(n log n) compared to typical clustering algorithms. These techniques have achieved detection accuracy of 96-99% in the zero-day attack detection jobs, but the false positive rates are still a challenge at 1-3% [8].

3.2 Deep Learning and Transformer Approaches

We show that transformer encoder architectures trained solely on benign traffic are able to recover the attributes of normal flow through the use of multi-head self-attention approaches. Deviations from thresholds of reconstruction errors indicate possible irregularities or zero-day attacks. In the context of zero-day detection, the unsupervised Transformer encoder has proven its utility, reaching 95.09% accuracy and 97.31% precision on the UNSW-NB15 dataset. In IoT networks, detection of zero-day attacks via dual-stage Transformer architectures on raw packet payloads achieves a 44% F1-score gain over baseline approaches [2].

3.3 Advanced Techniques: Quantum Machine Learning and LLM Agents

Quantum machine learning (QML) for zero-day detection is an emerging research topic that uses the quantum advantage in the exploration of high-dimension feature space. Geometric Quantum Machine Learning (GQML) detects 96.8% of known threats from normal traffic with 89.3% detection rate for zero-day variants and 3.2% false positives [3].

4. Transformer-Based Intrusion Detection Architectures

4.1 Fundamental Transformer Architecture and Self-Attention Mechanisms

Transformer architectures were originally proposed for sequence-to-sequence modeling, and they leverage multi-head self-attention to learn complicated feature interactions without recurrence or convolution [4]. For IDS tasks, the categorical features (protocol type, flags, etc.) and numerical data (packet count, duration, etc.) are embedded into a single latent space with positional encodings and fed into stacked multi-head attention blocks. The self-attention technique provides query-key-value projections which enable the model to give varying weights to different network flow characteristics [9].

4.2 Hybrid Transformer Architectures

Transformer-LSTM hybrid models combine transformer global context modeling with LSTM temporal dependency capture, achieving 99.89% accuracy on CICIDS2017 with Optuna-based hyperparameter optimization [10]. Transformer-KAN (Kolmogorov-Arnold Networks) models enhance pattern

approximation through non-linear transformations, achieving near-perfect accuracy on both binary and multi-class classification tasks in maritime network environments [12].

5. Explainable Artificial Intelligence Techniques for IDS

5.1 SHAP (SHapley Additive exPlanations)

SHAP presents a theoretically sound global and local explanation of feature importance using Shapley values from cooperative game theory [15]. For IDS applications, SHAP also produces force plots for visualizing the contribution of features for a certain prediction, which may allow security analysts to identify the cause of the traffic flows classified as intrusions. Global SHAP summary graphics rank characteristics by their mean relevance over all predictions. The method explains the decisions of transformer-based IDS with great fidelity and finds out the protocol type, packet count and connection length as the main attack signs [4]. The computational complexity of SHAP is O(n2), which is suitable for batch analysis but not for real-time deployment. Kernel explainer employs approximation SHAP variations with sampling to bring down the calculation to O(n3) enabling quasi real-time explanations on edge devices [35].

5.2 LIME (Local Interpretable Model-Agnostic Explanations)

LIME explains individual predictions with local linear approximations by perturbing the input characteristics and watching the changes in the output [15]. In the IDS use case, LIME provides packet level explanations, showing the most relevant information leading to an intrusion categorization conclusion. Its model-agnostic property allows it to be used to any IDS design without the need to access the model, offering explainability for proprietary systems. Hybrid techniques, combining LIME and SHAP, provide complimentary explanations, where LIME provides instance-level transparency and SHAP provides the global model behavior knowledge [16]. Hybrid SHAP-LIME methods boost trust ratings by 25% for cybersecurity specialists, but at the cost of 40% more computational overhead than single methods [17].

5.3 Attention Visualization and Gradient-Based Techniques

The attention weights explicitly encode the input attributes that the transformer considers most important for classification decisions [15]. Attention visualization heatmaps of traffic sequences show attention patterns that indicate risky areas that produce intrusion alarms. The process of self-attention in transformers is intrinsically interpretable through the attention weights, making post-hoc explanation approaches redundant. Gradient-weighted Class Activation Mapping (Grad-CAM) and Integrated Gradients provide gradient-based explanations complementing attention visualization [14]. These algorithms compute the gradients of the input with respect to the outputs of the model and locate the features that have the most impact on the predictions. Gradient-based methods are more faithful in characterizing complicated deep learning IDS models compared to perturbation-based methods [18].

6. Research Gaps, Challenges, and Future Directions

6.1 Critical Research Gaps

There has been a lot of development yet there are key gaps in the research:

1. Generalization across datasets is currently limited. Models trained on NSL-KDD suffer from 15-25% accuracy deterioration on CIC-IDS2017 due to distributional shifts and change in feature representation [19]. It is vital to research domain adaption methods for IDS.

2. There is no examination of adversarial robustness, where MLP-based IDS models have a 66.81% drop in accuracy under FGSM attacks on CIC-IDS2017 and full failure (0% detection) under transfer-based PGD attacks [20]. Adversarial training recovers some robustness (96%+ detection on CIC-IDS2017) but standardized robustness benchmarks are still missing.

3. There are no agreed upon measures for standardizing explainability. SHAP, LIME and attention visualization are offering distinct viewpoints on interpretability, but there is no consistent framework to evaluate the quality of explanations, fidelity or user understanding across methodologies [21].

4. Limited number of real-world deployment case studies restricts the knowledge of performance of the research prototypes in production network contexts with idea drift, traffic evolution and adversary adaptation [22].

6.2 Explainability and Interpretability Challenges

The opaque nature of deep learning models is at odds with security needs for transparent, auditable decision-making. XAI solutions address this challenge but some of the concerns are still open:

• Explanation fidelity and consistency: SHAP explanations can occasionally contradict domain knowledge and LIME local linearity assumptions fail on substantially non-linear decision boundaries [18].

Computational overhead: The calculation of SHAP is cubic to the feature dimension, making real-time explanation infeasible for high-dimensional traffic representations (greater than 100 features) [23].

• User misinterpretation: Despite the enhanced openness of the model, security analysts can misinterpret SHAP force plots or attention visualizations, which can result in an incorrect threat assessment [14].

6.3 Future Research Directions

Proposed Zero-day IDS Explainable Transformer Based Framework: Future research should be focused on an integrated architecture integrated with: 1. Multi-level Transformers with hierarchical attention, collecting packet level, flow level and session level anomalies simultaneously, addressing temporal patterns from second to minute [2]. 2. Federated Learning Integration allows collaborative model training by a number of dispersed companies in a privacy-preserving way, without centralizing sensitive network traffic [24]. 3. Quantum-Resistant Explainability [3] prepared for post-quantum encryption age by combining quantum machine learning with SHAP explanation creation. 4. Continual Learning techniques that change the model parameters incrementally when new attack patterns are encountered, without catastrophic forgetting, while keeping past information [6]. 5. Multi-Modal Threat Intelligence Fusion. Holistic zero-day detection using attention-weighted fusion techniques combining network telemetry, endpoint signals, threat feeds and behavioral analytics. 6. Explainable Decision Fusion that combines diverse IDS components (signature, anomaly, behavior) using interpretable ensemble methods, giving transparency in multi-source decision synthesis [25].

Conclusion

Transformer-based intrusion detection systems (IDSs) mark a paradigm change from manual feature engineering to automatic discovery of harmful patterns in network data. Their integration with explainable AI approaches fulfills major transparency requirements for implementation in security-critical infrastructure. A lot of progress has been made, e.g., state-of-the-art systems attain 99%+ accuracy on benchmark datasets. However, cross-dataset generalization, adversarial robustness, and scalability still remain key obstacles for practical application. Synthesis of the Literature study The synthesis of the literature study indicates that transformer topologies show promising potential in zero-day attack detection with an increase of 44% in the F1-score over baseline techniques [2]. But bridging the chasm between research excellence and operational implementation requires standardized evaluation frameworks, adversarial robustness assurances and human-centric explainability metrics. Future work combining quantum machine learning, federated learning, and continual adaptation will enable next-generation intrusion detection systems that can protect against unprecedented threats in 5G/6G, IoT, and edge computing settings [5].

References

[1] B. Chouhan, A. Sharma, H. Pathak, R. Raj, A. Gopi, and A. Joshi, “A comprehensive review of cybersecurity in 5G and 6G networks: SDN, NFV, slicing, and massive IoT security,” *International Conference on Computing for Sustainable Global Development*, Apr. 2026, doi: 10.23919/INDIACom70271.2026.11526205.

[2] M. D. Hssayeni and I. Mahgoub, “A transformer-based intrusion detection system for zero-day attack detection in IoT networks,” *Future Internet*, May 2026, doi: 10.3390/fi18060282.

[3] Mr. J. Krishna[1], P. Manasa[2], P. D. Kumar[3], M. Prathyusha[4], and M. G. [1]Associate Professor, “Zero-day attack detection using quantum machine learning: A hybrid framework with GQML-enhanced tkinter application for network monitoring and rule-based prediction,” *International Journal of Data Science and IoT Management System*, Apr. 2026, doi: 10.64751/ijdim.2026.v5.n2.pp114-124.

[5] Mrs. Darakshan, J. A. Dr, and A. Shaikh, “ENHANCING 6G NETWORK SECURITY WITH ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING INTEGRATION.” *International Journal of Applied Mathematics*, Nov. 2025, doi: 10.12732/ijam.v38i9s.870.

[6] S. Maric, R. Baidar, R. Abbas, and S. Reisenfeld, “System security framework for 5G advanced /6G IoT integrated terrestrial network-non-terrestrial network (TN-NTN) with AI-enabled cloud security,” *arXiv.org*, Aug. 2025, doi: 10.48550/arXiv.2508.05707.

[17] T. T and M. Subalakshmi, “Development of an adaptive machine learning framework for real-time anomaly detection in cybersecurity,” *THE SCIENTIFIC TEMPER*, Aug. 2025, doi: 10.58414/scientifictemper.2025.16.8.07.

[21] S. Anand and S. Zehra, “A comparative study of isolation forest and autoencoder models for proactive zero-day attack detection,” *IEEE International Conference on Consumer Electronics*, Feb. 2026, doi: 10.1109/ICCE67443.2026.11449661.

[22] R. R, S. M, N. S. T, and P. Archana, “Dual-phase learning approach for zero-day intrusion detection using NSL-KDD,” *International Research Journal on Advanced Engineering Hub (IRJAEH)*, Feb. 2026, doi: 10.47392/irjaeh.2026.0068.

[26] Y. Saheed and J. Chukwuere, “Autonomous LLM agent: A memory-augmented, edge-optimized SHAP explanations with zero-day attack resilience in IoT/industrial IoT networks,” *IEEE Internet of Things Journal*, Apr. 2026, doi: 10.1109/JIOT.2025.3648649.

[27] K. Harshdeep, K. Sumalatha, and R. Mathur, “DeepTransIDS: Transformer-based deep learning model for detecting DDoS attacks on 5G NIDD,” *Results in Engineering*, Apr. 2025, doi: https://doi.org/10.1016/j.rineng.2025.104826.

[28] K. Benjamin and T. Kacem, “UAV DoS attack detection using a hybrid transformer-LSTM approach,” *International Conference on Wireless Communications and Mobile Computing*, May 2025, doi: 10.1109/IWCMC65282.2025.11059673.

[29] Z. Lu, H. Zeng, Z. Zheng, C. Li, and X. Dong, “Research on deep learning-based network intrusion detection methods for smart ships,” *Brodogradnja*, 2026, doi: 10.21278/brod77205.

[30] A. Akgündoğdu and Ş. Çelikbaş, “Explainable deep learning framework for brain tumor detection: Integrating LIME, grad-CAM, and SHAP for enhanced accuracy.” *Medical Engineering and Physics*, July 2025, doi: 10.1016/j.medengphy.2025.104405.

[34] T. Vengatesh, “Transparent decision-making with explainable ai (xai): Advances in interpretable deep learning.” *Journal of Information Systems Engineering & Management*, May 2025, doi: 10.52783/jisem.v10i4.10584.

[35] N. Almolhis, “Intrusion detection using hybrid random forest and attention models and explainable AI visualization,” *Journal of Internet Services and Information Security*, Feb. 2025, doi: 10.58346/jisis.2025.i1.024.

[36] Z. Tayari and M. Zaied, “Beyond the black box: A hybrid SHAP-LIME approach for transparent and explainable deep neural networks,” *ACS/IEEE International Conference on Computer Systems and Applications*, Oct. 2025, doi: 10.1109/AICCSA66935.2025.11315251.

[37] M. S. Islam *et al.*, “Explainable AI in healthcare: Leveraging machine learning and knowledge representation for personalized treatment recommendations,” *Journal of Posthumanism*, Jan. 2025, doi: 10.63332/joph.v5i1.1996.

[38] F. Basheer, “Interpretability in deep learning: A comparative study of grad-CAM, LIME, and SHAP across imaging domains,” *2025 IEEE 4th International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET)*, Oct. 2025, doi: 10.1109/TEMSMET65536.2025.11467494.

[41] C. Xin and K. Xu, “Cross-dataset transformer-IDS with calibration and AUC optimization (evaluated on NSL-KDD, UNSW-NB15, CIC-IDS2017),” *Journal of Cyber Security*, 2025, doi: 10.32604/jcs.2025.071627.

[46] S. Tarek, M. Muthmainnah, and A. J. Obaid, “A cross-dataset empirical evaluation of adversarial evasion attacks and defenses in machine learning-based intrusion detection systems,” *Computational Discovery and Intelligent Systems (CDIS)*, Apr. 2026, doi: 10.66279/3k5mqs50.

[47] V. Z. Mohale and I. C. Obagbuwa, “Evaluating machine learning-based intrusion detection systems with explainable AI: Enhancing transparency and interpretability,” *Frontiers in Computer Science*, May 2025, doi: https://doi.org/10.3389/fcomp.2025.1520741.

[48] V. Sharma, V. Singh, V. Tiwari, S. Narayanasamy, S. R, and G. Rambabu, “Enhanced machine learning algorithms for real-time anomaly detection in network security: Addressing scalability, adaptability and privacy challenges in the face of emerging cyber threats,” *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies*, 2025, doi: 10.5220/0013944000004919.

[49] S. Kumar, N. Dhanda, and K. Gupta, “Exploring explainable AI methods for decoding machine learning model decisions,” *International Conference Intelligent Data Communication Technologies and Internet Things*, June 2025, doi: 10.1109/ICICI65870.2025.11069883.

[51] M. A. Bilal, I. U. Islam, N. Iltaf, M. J. Khan, and M. J. Khan, “Federated learning with explainable AI for malicious traffic detection in IoT networks,” *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3613459.

[52] V. Nitin, Aarti, and V. N. Thatha, “Deep learning for web application security: A comprehensive survey on intrusion detection systems,” *IEEE International Symposium on Compound Semiconductors*, Nov. 2025, doi: 10.1109/ISCS69371.2025.11386278.