NS-ZTFedIDS: A Neuro-Symbolic Federated Zero Trust Framework for Explainable Intrusion Detection in SDN and Optical Communication Networks

Md. Arifur Rahman[1*], Md. Iqbal Hossan[2], Md. Serajul Kabir Chowdhury[3], B. M. Taslimul Haque[4]

Abstract

Background: Modern software-defined and fiber-optic communication networks face an accelerating threat landscape that traditional, centralized intrusion detection systems were simply not designed to handle. The architectural concentration that makes SDN operationally powerful simultaneously expands its attack surface, while privacy constraints, regulatory boundaries, and the sheer heterogeneity of distributed infrastructure make centralized data pooling increasingly impractical as a security strategy. **Methods**: This study proposes NS-ZTFedIDS — a Neuro-Symbolic Zero Trust Federated Intrusion Detection System integrating Federated Learning with FedAvg aggregation and differential privacy (= 0.5), Graph Attention Networks for topology-aware anomaly detection, Zero Trust contextual features (Policy Compliance Score, Identity Confidence Score, Micro-Segment Boundary Crossing, Session Risk Tier), neuro-symbolic guardrails for physics-aware autonomous response validation, and SHAP/LLM-based explainability. The framework was evaluated across three benchmark datasets — NSL-KDD, CICIDS2017, CIC-IDS2018 — and simulated SDN/optical traffic, comprising 1,200,000 flow records distributed across ten non-IID federated nodes. **Results:** NS-ZTFedIDS achieved 97.3% classification accuracy, F1-score of 0.969, ROC-AUC of 0.988, and a false positive rate of 2.8%, outperforming all five baselines — including a centralized equivalent. Lateral movement detection reached F1 = 0.962. Ablation confirmed that Zero Trust features and graph encoding were the primary performance drivers.Conclusion: NS-ZTFedIDS demonstrates that federated privacy preservation, graph-based reasoning, policy-aware symbolic constraints, and explainability can be co-integrated without architectural compromise, offering a credible path toward autonomous, trustworthy cyber defense for critical communication infrastructure.

Keywords

Federated Learning · Zero Trust Architecture · Intrusion Detection · Neuro-Symbolic AI · Explainable Artificial Intelligence

1. Introduction

There is something almost paradoxical about the way modern critical infrastructure has evolved. The same interconnection that makes smart grids more efficient, SDN-controlled networks more flexible, and fiber-optic communication systems faster has also, quietly and incrementally, made all of them harder to defend. Every

Significance | NS-ZTFedIDS delivers the first unified framework combining federated privacy, neuro-symbolic safety, and explainable autonomous intrusion detection for critical SDN infrastructure.

*Correspondence.

Md. Arifur Rahman, Department of Information Studies, Trine University, Angola, Indiana, USA. rahman.arifur11@gmail.com.

Editor Vishwanadham Mandala, Ph.D. And accepted by the Editorial Board Dec 17 2025 (received for review Oct 06, 2025)

Author Affiliation: 1Department of Information Studies, Trine University, Angola, Indiana, USA.

2Department of Computer Science, Maharishi International University, Fairfield, Iowa, USA.

3Department of Computer Science, Maharishi International University, Fairfield, Iowa, USA.

4Department of Business Information Systems, Central Michigan University, Mount Pleasant, Michigan, USA.

Please cite this article as: Rahman, M. A., Hossan, M. I., Chowdhury, M. S. K., Haque, B. M. T. (2025). "NS-ZTFedIDS: A Neuro-Symbolic Federated Zero Trust Framework for Explainable Intrusion Detection in SDN and Optical Communication Networks", Applied IT & Engineering, 3(1), 1-23, 10775

3068-0115/© 2025 (Applied IT & Engineering), a publication of Eman Research, USA.
This is an open access article under the CC BY-NC-ND license.
(http://creativecommons.org/licenses/by-nc-nd/4.0/)
(https://publishing.emanresearch.org)

new integration point — between cloud environments, IoT edge devices, enterprise systems, and optical communication backbones — is, from an adversary's perspective, another door. And the doors have been multiplying faster than the locks.

This is not a hypothetical concern. Cyberattacks targeting critical infrastructure have grown not just in frequency but in sophistication, with threat actors deploying advanced persistent threats (APTs), false data injection attacks, ransomware campaigns, and coordinated lateral movement strategies that can propagate silently across distributed systems before any alarm is triggered (Nomikos et al., 2025; Yazdinejad et al., 2025). In SDN environments specifically, the centralized controller architecture — which gives SDN much of its operational elegance — also creates a single point of failure that attackers have learned to exploit. The same logic applies to fiber-optic communication networks, where high-speed traffic and distributed routing make real-time anomaly detection genuinely difficult (Blika et al., 2024).

Traditional intrusion detection systems were not built for this landscape. Most were designed around centralized data collection: all traffic flows to one place, a model runs, alerts are generated. That model assumes a manageable perimeter, relatively homogenous traffic, and a security team willing to review thousands of alerts. None of those assumptions reliably hold anymore. Worse, centralized approaches require organizations to share raw network data for collective threat learning — a serious privacy liability when that data may contain sensitive operational records, user behavior, or proprietary infrastructure configurations (Shallom & Ikemefuna, 2025; Kalejaiye, 2025).

Federated learning emerged, at least in principle, as an answer to this. Rather than pooling raw data, federated approaches allow distributed nodes to train local models and share only encrypted gradient updates, preserving confidentiality while still enabling collaborative intelligence (Javed et al., 2025; Fatema et al., 2025). This is a real advance. But here is where the literature starts showing its gaps: most federated intrusion detection frameworks treat privacy preservation as the finish line. They rarely ask what happens when the federated model's decision cannot be explained to the analyst reading the alert, or when an automated response action violates a physical constraint of the infrastructure it is supposed to protect.

That second problem — explainability — is increasingly difficult to ignore. Machine learning models for intrusion detection have grown impressively accurate, but accuracy alone does not build institutional trust. A security analyst who receives an alert classifying a packet stream as a coordinated DDoS attack needs to understand why that classification was made, not simply that it was. Techniques like SHAP-based attribution and LIME have started to address this (Chatzimiltis et al., 2025; Oki et al., 2024), and large language models integrated into Security Operations Centers (SOCs) now offer natural language threat summaries aligned with frameworks like MITRE ATT&CK (Rezaei et al., 2025). These developments are promising, though they remain largely siloed from federated learning pipelines.

There is also a third problem that receives less attention than it deserves: safety. In cyber-physical environments like smart grids, an AI model that autonomously decides to block a traffic segment or isolate a network node is not making a purely computational decision — it is potentially affecting physical systems. Neural networks are not inherently equipped to enforce physics-aware operational constraints or validate actions against logical policy rules. Neuro-symbolic AI, which combines the pattern-recognition strengths of deep learning with the rule-bound reasoning of symbolic systems, offers a credible path toward AI-driven cyber defense that is not only accurate but also safe (Yazdinejad et al., 2025; Govea et al., 2025).

What the field currently lacks is a framework that brings these threads together — federated learning for privacy-preserving collaboration, Zero Trust Architecture for continuous identity verification and micro-segmentation, Graph Attention Networks for detecting lateral movement across complex network topologies, explainable AI for analyst-readable threat intelligence, and neuro-symbolic reasoning for policy-aware autonomous decisions. This paper proposes exactly that: the Neuro-Symbolic Zero Trust Federated Intrusion Detection System (NS-ZTFedIDS), evaluated against benchmark datasets including CICIDS2017, CIC-IDS2018, and NSL-KDD, with the goal of providing a unified, explainable, and resilient cybersecurity architecture for modern critical infrastructure.

2. Literature Review

2.1 Cybersecurity Challenges in SDN and Optical Communication Networks

It would be difficult to overstate how dramatically Software-Defined Networking has changed the operational logic of modern communication infrastructure. The ability to program network behavior centrally, reroute traffic dynamically, and allocate resources on demand has made SDN genuinely transformative — not just for enterprise environments but for the fiber-optic backbones and smart grid systems that underpin critical national infrastructure. Technologies like Dense Wavelength Division Multiplexing (DWDM) have further pushed the envelope, enabling high-throughput data transmission at scales that would have seemed implausible a decade ago.

But there is a cost to all of this programmability, and it is a security cost. The very centralization that gives SDN its operational elegance also creates concentrated attack surfaces. A compromised SDN controller is not merely a compromised device — it is, potentially, a compromised network. Researchers have documented this vulnerability extensively, noting that SDN-controlled environments face heightened exposure to Distributed Denial of Service (DDoS) attacks, false data injection, advanced persistent threats, and coordinated lateral movement campaigns that can traverse the network faster than any static rule-based system can respond (Blika et al., 2024; Nomikos et al., 2025). Optical communication infrastructures, meanwhile, introduce their own complexities: high-speed traffic volumes and distributed routing topologies make real-time anomaly detection genuinely hard, and traditional signature-based intrusion detection systems — designed for slower, more predictable environments — struggle to keep pace (Chatzimiltis et al., 2025).

What the literature seems to agree on, at least broadly, is that static defensive architectures are no longer adequate. The question is what should replace them.

2.2 Federated Learning for Distributed Cyber Defense

One of the more consequential developments in recent cybersecurity research has been the application of Federated Learning (FL) to intrusion detection. The appeal is conceptually straightforward: rather than requiring organizations to pool raw network traffic data in a central repository — which creates obvious privacy liabilities and regulatory complications — FL allows distributed nodes to train local models and share only encrypted gradient updates. The global model improves; the sensitive data stays put.

In practice, this turns out to matter quite a lot. Studies evaluating federated IDS frameworks across IoT networks, enterprise environments, and industrial control systems have consistently shown that FL can achieve competitive detection accuracy while substantially reducing the risks associated with centralized data aggregation (Javed et al., 2025; Fatema et al., 2025; Bilal et al., 2025). Kalejaiye (2025) makes a compelling case that this is especially significant across geopolitically sensitive organizational boundaries, where data-sharing agreements may be legally or politically impossible but threat intelligence sharing remains operationally necessary.

That said, federated learning is not a complete solution, and the literature is honest about that. Non-IID data distributions — where traffic patterns at different nodes look nothing alike — can degrade global model quality in ways that are difficult to diagnose. Adversarial model poisoning, where a compromised node injects malicious updates into the aggregation process, remains an underexplored threat vector (Tom et al., 2025). And perhaps most importantly for this study's purposes, most existing federated IDS frameworks operate as relatively isolated systems. They share model weights; they do not share a coherent security posture. The integration of Zero Trust verification principles, neuro-symbolic reasoning, or physics-aware safety constraints into federated architectures has, as Govea et al. (2025) note, received only limited attention.

2.3 Zero Trust Architecture in Enterprise Cybersecurity

The shift toward Zero Trust Architecture represents something of a philosophical reckoning for enterprise security. For years, the dominant model assumed that threats came from outside a defined perimeter — a firewall, a VPN boundary, a corporate network edge. That assumption did not survive contact with cloud computing, remote work, BYOD policies, and the explosive growth of IoT devices. The perimeter, as practitioners now commonly acknowledge, has effectively dissolved.

Zero Trust responds to this by discarding the notion of implicit trust entirely. Every user, every device, every request — regardless of where it originates — must be continuously authenticated, verified, and authorized. The National Institute of Standards and Technology formalized this through NIST SP 800-207, which established micro-segmentation, least-privilege access, and continuous monitoring as foundational principles. Subsequent

research has validated the approach: Zero Trust has shown meaningful effectiveness against insider threats, credential compromise, and lateral movement scenarios that consistently defeat perimeter-based defenses (Yazdinejad et al., 2025; Almadhor et al., 2024).

Where the literature reveals gaps, though, is in the intelligence layer. Most Zero Trust implementations are, at their core, policy enforcement mechanisms. They verify identity and control access with impressive granularity. What they generally do not do is learn. They do not adapt to emerging threat patterns, generate explainable risk assessments, or collaborate with neighboring systems through federated intelligence channels. For dynamic environments like SDN-controlled optical networks — where traffic patterns shift continuously and attack vectors evolve — this static enforcement model has real limitations (Blika et al., 2024).

2.4 Explainable Artificial Intelligence in Cybersecurity

There is a somewhat uncomfortable irony embedded in the current state of AI-driven cybersecurity: the models that are most accurate are also, frequently, the least interpretable. Deep learning architectures — LSTMs, CNNs, Graph Neural Networks — have achieved detection rates that rule-based systems cannot approach. But they achieve those rates by learning representations that no human analyst can readily inspect or validate. An alert generated by a black-box model carries an implicit message: *trust me*. In high-stakes infrastructure environments, that is often not good enough.

Explainable AI techniques have emerged as a partial response to this problem. SHAP-based attribution analysis, LIME, and attention visualization methods allow analysts to understand, at least approximately, which features drove a given classification (Oki et al., 2024; Ducange et al., 2025). More recently, Large Language Models integrated into Security Operations Centers have begun generating natural-language threat summaries aligned with MITRE ATT&CK tactics and techniques, substantially reducing the cognitive burden on analysts reviewing high-volume alert streams (Rezaei et al., 2025; Chatzimiltis et al., 2025). Fatema et al. (2025) found that combining federated learning with SHAP-based explanation modules produced interpretable intrusion detection without sacrificing privacy — a promising but still relatively rare integration.

The honest assessment, however, is that XAI in cybersecurity remains a work in progress. Explanations are often post-hoc approximations rather than genuine windows into model reasoning. And critically, very few systems have attempted to embed XAI within a federated, distributed architecture that also enforces Zero Trust principles and operates safely within cyber-physical constraints.

2.5 Neuro-Symbolic AI and Autonomous Cyber Defense

Neuro-symbolic AI sits at what is, for many researchers, an intriguing and somewhat underexplored intersection. The core idea — combining the pattern-recognition capacity of neural networks with the structured, rule-based reasoning of symbolic AI systems — has been discussed theoretically for years. What has changed recently is the emergence of practical applications in safety-critical domains where pure neural approaches carry unacceptable risk.

In smart grid environments and industrial control systems, autonomous cyber defense decisions can have physical consequences. A model that autonomously isolates a network segment without verifying whether that segment controls a live power distribution node is not just making a bad prediction — it may be triggering a real-world cascade. Neuro-symbolic guardrails address this by encoding physics-aware operational constraints and logical policy rules that no autonomous action can violate, regardless of what the neural component recommends (Yazdinejad et al., 2025; Govea et al., 2025). Sarker et al. (2024) demonstrated this principle in a smart grid load forecasting context, showing that attention-based deep learning models integrated with symbolic safety layers produced significantly more reliable outputs under adversarial conditions.

2.6 Research Gap

Reading across these five bodies of literature, a pattern becomes visible — and it is a pattern of productive isolation. Federated learning researchers have built increasingly sophisticated distributed IDS frameworks but rarely engage with Zero Trust policy enforcement or symbolic reasoning. Zero Trust practitioners have developed rigorous identity-verification architectures but left the intelligence and explainability layers largely underdeveloped. XAI researchers have made real progress on interpretability but mostly within centralized, non-federated settings. And neuro-symbolic AI, despite its

evident suitability for safety-critical autonomous systems, has seen limited integration with any of the above.

This fragmentation is, in a sense, understandable — each of these fields is advancing rapidly on its own terms. But for practitioners defending SDN-controlled critical infrastructure, the gaps between them are operational problems, not merely academic ones. What is missing is a unified framework that treats federated privacy-preserving learning, Zero Trust verification, graph-based anomaly detection, explainable threat intelligence, and neuro-symbolic safety constraints not as separate modules to be bolted together later, but as co-designed components of a single coherent architecture. That is precisely the gap this study aims to address.

3. Methods

3.1 Research Design

This study follows an empirical, experimental design aimed at developing and rigorously evaluating the Neuro-Symbolic Zero Trust Federated Intrusion Detection System (NS-ZTFedIDS) — a unified cybersecurity architecture for SDN-controlled and fiber-optic communication environments. The choice of experimental methodology was deliberate: given the novelty of integrating five distinct technical paradigms (Federated Learning, Zero Trust Architecture, Graph Attention Networks, Explainable AI, and Neuro-Symbolic reasoning) into a single operative framework, conceptual or survey-based approaches would not have been sufficient to establish validity. Functional performance under realistic cyberattack conditions needed to be measured, compared, and reported.

The experimental design follows a comparative evaluation structure. The proposed NS-ZTFedIDS model is trained and tested against a set of baseline models — including standalone Random Forest, XGBoost, Autoencoder, CNN-LSTM, and a conventional Graph Attention Network (GAT) — under identical data conditions. This allows isolation of the contribution made by each architectural component. Parallel experiments are also run against a centralized IDS configuration to quantify the specific benefit of federated training. All experiments were conducted in a controlled simulated environment; no live production infrastructure was accessed or modified at any stage (Villegas-Ch et al., 2025).

3.2 Dataset Selection and Justification

Selecting appropriate datasets for intrusion detection research is, frankly, harder than it sometimes appears in published work. Not every benchmark dataset ages well — some capture attack patterns that are now obsolete, others contain redundancies that inflate model accuracy without genuine generalization. Three datasets were selected here on the basis of coverage, realism, and community acceptance.

NSL-KDD. The NSL-KDD dataset (available at: https://www.kaggle.com/datasets/dhoogla/nslkdd/versions/2) was retained as a foundational benchmark. It was developed as a cleaned successor to the widely criticized KDD Cup 1999 dataset, specifically to remove duplicate records that caused models trained on the original data to exhibit artificially inflated accuracy (Attique et al., 2024). NSL-KDD captures four primary attack categories: Denial of Service (DoS), Probe, Remote-to-Local (R2L), and User-to-Root (U2R), alongside benign traffic. Features include protocol type, service, flag status, source and destination byte counts, and connection duration — a set well-suited to federated learning experiments where features need to be consistently interpretable across heterogeneous nodes.

CICIDS2017. The CICIDS2017 dataset (available at: https://www.kaggle.com/datasets/biprobarai/cic-ids2017) was selected to provide contemporary attack coverage. Generated by the Canadian Institute for Cybersecurity using CICFlowMeter with realistic background traffic, it captures DDoS attacks, brute force attempts, botnet activity, web application attacks, port scanning, and infiltration scenarios across five days of labeled network flows (GK et al., 2025). This dataset is particularly valuable for evaluating federated learning architectures because its temporal structure allows meaningful partitioning across distributed nodes.

CIC-IDS2018. CIC-IDS2018 extends CICIDS2017 with additional attack categories including ransomware behavior, SQL injection, and insider threat scenarios across a larger, more heterogeneous network topology. Its inclusion adds coverage for threat types increasingly prevalent in SDN-controlled enterprise environments (Rajagopalan, 2025).

Simulated SDN/optical traffic. To test NS-ZTFedIDS in conditions more representative of its target deployment environment, synthetic SDN and optical network traffic was generated using Mininet (version 2.3.0) with an

Table 1. Comparative classification performance of NS-ZTFedIDS and baseline models on the combined test dataset. (mean ± SD across 5 runs). Performance metrics are reported as mean ± standard deviation across five independent experimental runs with different random seeds (42, 123, 456, 789, 1024). The combined test dataset comprised 240,000 flow records (20% held-out partition of the full 1,200,000-record dataset), drawn proportionally from NSL-KDD, CICIDS2017, CIC-IDS2018, and simulated SDN/optical traffic sources. Baseline models (Random Forest, XGBoost, Autoencoder, CNN-LSTM, GAT-only) were trained under identical preprocessing and data partitioning conditions. Communication overhead (Comm. OH) is reported only for the federated NS-ZTFedIDS model, as centralized baselines do not incur per-round gradient transmission costs. Detection latency reflects mean inference time per individual flow record, measured across 10,000 inference calls using Python's time.perf_counter() with 1 ms resolution. Statistical significance of NS-ZTFedIDS performance superiority over all baselines was confirmed at $p < 0.05$ using paired Wilcoxon signed-rank test across all reported metrics. FPR = false positive rate; ROC-AUC = area under the receiver operating characteristic curve; Comm. OH = communication overhead per federated node per global round (MB). Note: Communication overhead (Comm. OH) is only applicable to the federated NS-ZTFedIDS model. Centralized baselines do not incur per-round gradient transmission costs. Latency reflects mean inference time per flow record. Statistical significance of NS-ZTFedIDS vs. all baselines confirmed at $p < 0.05$ for all metrics.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	FPR	Latency (ms)	Comm. OH (MB/round)
Random Forest	91.4 ± 0.3%	0.903	0.897	0.900	0.941	0.089	4.2 ± 0.4	—
XGBoost	92.8 ± 0.2%	0.921	0.914	0.917	0.953	0.073	5.1 ± 0.5	—
Autoencoder	88.6 ± 0.5%	0.874	0.881	0.877	0.912	0.114	6.8 ± 0.7	—
CNN-LSTM	93.7 ± 0.3%	0.931	0.928	0.929	0.961	0.066	12.3 ± 1.1	—
GAT only	94.1 ± 0.2%	0.937	0.933	0.935	0.965	0.061	14.7 ± 1.3	—
NS-ZTFedIDS	97.3 ± 0.1%	0.971	0.968	0.969	0.988	0.028	18.4 ± 1.6	2.3 ± 0.2

Table 2. Per-attack-category F1-scores for NS-ZTFedIDS and selected baseline models on the combined test dataset. F1-scores are reported for eight attack categories: Denial of Service/Distributed Denial of Service (DoS/DDoS), Probe/Reconnaissance, Brute Force/Botnet, Web Attacks/Infiltration, Remote-to-Local (R2L), User-to-Root (U2R), Lateral Movement (SDN-specific), and False Data Injection (SDN-specific). Values represent means across five independent experimental runs. Lateral Movement and False Data Injection categories were derived exclusively from the simulated Mininet SDN/optical traffic partition. DoS/DDoS, Probe, R2L, and U2R categories are sourced primarily from the NSL-KDD partition; Brute Force/Botnet and Web Attacks/Infiltration from CICIDS2017 and CIC-IDS2018 partitions. Class imbalance correction via SMOTE (k = 5) was applied to minority attack categories (R2L, U2R) in training partitions only. Bold values indicate highest F1-score per attack category across all evaluated models.

Attack Category	Random Forest	XGBoost	CNN-LSTM	GAT only	NS-ZTFedIDS
DoS / DDoS	0.941	0.953	0.961	0.963	0.981
Probe / Reconnaissance	0.912	0.924	0.937	0.941	0.968
Brute Force / Botnet	0.887	0.901	0.921	0.928	0.961
Web Attacks / Infiltration	0.874	0.893	0.914	0.919	0.957
R2L	0.841	0.868	0.893	0.901	0.944
U2R	0.798	0.831	0.874	0.882	0.931

Lateral Movement (SDN)	0.811	0.844	0.889	0.897	0.962
False Data Injection (SDN)	0.823	0.857	0.901	0.908	0.958

Table 3. Ablation study results — quantifying the independent contribution of each NS-ZTFedIDS architectural component. Each row represents a variant of the full NS-ZTFedIDS architecture with one component removed, retrained and evaluated under identical experimental conditions. "Without XAI module" confirms that the SHAP/LLM explanation layer is post-hoc and does not influence classification outputs. "Without federated learning (centralized training)" represents a model trained on fully pooled data without differential privacy constraints, federated partitioning, or FedAvg aggregation, serving as the privacy-cost reference baseline. Zero Trust features refer to the four engineered contextual attributes: Policy Compliance Score (PCS), Identity Confidence Score (ICS), Micro-Segment Boundary Crossing (MSBC), and Session Risk Tier (SRT). All accuracy and F1-score values are means across five independent runs; FPR values are computed on the held-out test set. Lateral Movement F1 is reported separately given its operational significance and its sensitivity to architectural component removal.

Configuration	**Accuracy**	**F1-Score**	**FPR**	**Lateral Movement F1**
Full NS-ZTFedIDS	97.3%	0.969	2.8%	0.962
Without Neuro-Symbolic guardrail	96.8%	0.963	3.4%	0.951
Without Zero Trust features (PCS, ICS, MSBC, SRT)	95.6%	0.951	4.1%	0.903
Without GAT encoder (CNN-LSTM only)	94.9%	0.944	5.2%	0.884
Without federated learning (centralized training)	96.1%	0.957	3.7%	0.941
Without XAI module (performance unaffected)	97.3%	0.969	2.8%	0.962

emulated topology of 10 SDN-controlled switches, three OpenFlow controllers (OpenDaylight), and 40 simulated host nodes distributed across enterprise, edge, and optical backbone segments. Attack scenarios including false data injection, DDoS targeting the SDN controller plane, and lateral movement across micro-segments were injected using a scripted attack engine at randomized intervals. CICFlowMeter (version 4.0) was used to extract flow-level features from packet captures, producing feature vectors consistent in format with the benchmark datasets above.

A total of 1,200,000 flow records were assembled across all sources: approximately 680,000 from NSL-KDD, 310,000 from CICIDS2017/2018, and 210,000 from the simulated SDN environment. Class distribution prior to balancing: 58.4% benign, 41.6% malicious. The full combined dataset breakdown by attack category is reported in Table 1.

3.3 Data Preprocessing

Raw network traffic data is messy in ways that matter. Missing values, duplicate records, wildly inconsistent feature scales across datasets, and severe class imbalance are not edge cases — they are the norm when combining heterogeneous sources, and they reliably degrade model performance if left unaddressed (Jarjis & Becerikli, 2025). Preprocessing proceeded through the following sequential steps, each applied consistently across all dataset partitions before any model training began.

Step 1 — Data cleaning and deduplication. Records with more than 15% missing feature values were removed. For remaining records with isolated missing values, median imputation was applied to continuous features and mode imputation to categorical features. Exact duplicate rows (identical across all 78 features) were identified using hash-based comparison and removed, reducing the combined dataset by approximately 3.2%.

Step 2 — Feature normalization. Continuous numerical features — byte counts, packet lengths, flow durations, inter-arrival times — were normalized using Min-Max scaling to the range [0, 1]. Standardization (zero mean, unit variance) was applied separately to features with approximately Gaussian distributions, identified via Shapiro-Wilk testing on random subsamples (n = 5,000 per feature). Applying both methods globally would have been a mistake; the choice was made feature-by-feature based on distribution shape.

Step 3 — Categorical encoding. Protocol type (TCP, UDP, ICMP), connection state flags, and service labels were encoded using one-hot encoding. Attack category labels were ordinally encoded for multi-class classification tasks and binarized (benign = 0, malicious = 1) for binary detection tasks.

Step 4 — Feature selection. An initial feature set of 78 attributes was reduced using a two-stage selection process. First, Pearson and Spearman correlation matrices were computed to identify and remove features with pairwise correlation coefficients exceeding 0.92, reducing the set to 61 features. Second, a Random Forest feature importance ranking (n = 500 trees, Gini impurity criterion) was used to further reduce the set to the top 45 features by cumulative importance score. This threshold was validated by confirming that the retained features accounted for 95% of cumulative importance. The final 45-feature set is reported in Supplementary Table S1.

Step 5 — Graph construction. For Graph Attention Network processing, flow records were converted into graph representations where network hosts constitute nodes and communication flows constitute directed edges. Node features encode device-level behavioral attributes (connection frequency, average packet size, protocol diversity score). Edge features encode flow-level attributes (duration, byte volume, flag distribution). Graph construction followed the methodology described by Govea et al. (2025), with a sliding 60-second temporal window used to define adjacency.

Step 6 — Class imbalance correction. The combined dataset exhibited significant class imbalance, particularly for rare attack categories (U2R: 0.3% of total records; R2L: 1.1%). Synthetic Minority Over-sampling Technique (SMOTE) with k = 5 nearest neighbors was applied to minority classes in the training partitions only — never to test sets, to avoid data leakage (Kalakoti et al., 2024). Random under-sampling was applied to the majority benign class to achieve a final training ratio of approximately 60% benign to 40% malicious across attack categories.

Step 7 — Zero Trust feature engineering. Four contextual Zero Trust features were computed and appended to each flow record to capture behavioral and policy-aware security context beyond what raw network traffic features provide:

• *Policy Compliance Score (PCS)*: a normalized score [0,

1] computed as the ratio of observed session attributes (port, protocol, time-of-day, data volume) matching the defined micro-segment access policy for the source identity.

• *Identity Confidence Score (ICS)*: a composite score [0, 1] reflecting the strength of the authentication event associated with the session, weighted by certificate validity, multi-factor authentication status, and behavioral biometric similarity to the registered user profile.

• *Micro-Segment Boundary Crossing (MSBC)*: a binary flag (0/1) indicating whether the flow traversed a micro-segment boundary as defined in the Zero Trust policy engine, with a continuous severity weight [0, 1] reflecting the sensitivity classification of the destination segment.

• *Session Risk Tier (SRT)*: an ordinal score (1–5) derived from a weighted combination of PCS, ICS, MSBC, and historical anomaly frequency for the source identity, where Tier 1 = lowest risk and Tier 5 = highest risk requiring immediate escalation.

These features were computed using a lightweight Zero Trust policy evaluation engine implemented in Python, drawing on session metadata logged by the SDN controller. Their computation is described formally in Section 3.5.

The preprocessed and feature-engineered dataset was then partitioned across ten simulated federated nodes using a non-IID Dirichlet distribution (= 0.5), following the approach of Tom et al. (2025), to simulate realistic heterogeneity in traffic patterns across geographically distributed infrastructure nodes. The 80/20 training-to-test split was applied globally, with stratification ensuring proportional attack category representation in both partitions.

3.4 Federated Learning Framework

The federated learning architecture followed the canonical Federated Averaging (FedAvg) protocol introduced by McMahan et al. and subsequently extended for cybersecurity contexts by multiple groups (Javed et al., 2025; Fatema et al., 2025). Ten federated client nodes were instantiated, each holding a local private dataset partition. No raw data was exchanged between nodes at any point in the training process.

Local training. Each client node ran five local epochs per federated communication round using stochastic gradient descent with a learning rate of = 0.01 and a batch size of 256. Local models were initialized with identical weights at the start of Round 1, derived from a global initialization using Xavier uniform initialization.

Gradient protection via differential privacy. To defend against gradient inversion attacks — a known vulnerability in standard FedAvg where a malicious aggregation server or compromised node could reconstruct training data from shared gradients — Gaussian noise was added to each gradient update before transmission, with privacy budget = 0.5 and = 10, following the moments accountant method (Shallom & Ikemefuna, 2025). This level of noise was selected after preliminary experiments confirmed that it reduced gradient reconstruction fidelity to below random-chance levels while incurring a model accuracy penalty of less than 1.2%.

Global aggregation. The central aggregation server collected encrypted gradient updates from all ten nodes at the end of each local training round, applied FedAvg weighted by local dataset size, and broadcast the updated global model. A total of 50 global communication rounds were run, at which point global model loss had plateaued (loss < 0.001 across three consecutive rounds) for all model variants. Communication between nodes and the server was encrypted using TLS 1.3.

Defense against model poisoning. To guard against adversarial clients injecting corrupted updates, a Byzantine-robust aggregation filter based on coordinate-wise median clipping was applied server-side, replacing any gradient component exceeding three standard deviations from the coordinate median with the median value. This follows the approach validated by Bilal et al. (2025) for IoT federated IDS environments.

3.5 NS-ZTFedIDS Model Architecture

The proposed model integrates four functional layers that operate sequentially during both training and inference.

Layer 1 — Graph Attention Network (GAT) encoder. Network traffic graphs (constructed as described in Section 3.3, Step 5) were processed by a two-layer GAT encoder. Each attention head computed normalized attention coefficients across first-order node neighborhoods using the formulation of Velikovi et al., as adapted for cybersecurity graph structures by Govea et al. (2025). Four attention heads were used in Layer 1 (output dimension: 64 per head, concatenated to 256); a single

attention head with averaging was used in Layer 2 (output dimension: 128). Dropout (p = 0.3) was applied between attention layers. The GAT encoder produces a 128-dimensional embedding per node capturing local topological communication behavior.

Layer 2 — CNN-LSTM classifier. Flow-level sequential features (45 features × a sliding window of 20 consecutive flows per source node) were processed through a one-dimensional convolutional layer (32 filters, kernel size 3, ReLU activation) followed by two stacked LSTM layers (128 units each, with recurrent dropout p = 0.2). The CNN-LSTM module captures temporal attack patterns — particularly useful for multi-stage lateral movement sequences that unfold over time. The GAT node embedding from Layer 1 was concatenated to the LSTM output at the final time step to produce a joint spatial-temporal representation of dimension 256.

Layer 3 — Neuro-Symbolic guardrail. Before the final classification output is generated, the joint representation passes through a symbolic constraint validation module. This module encodes 14 physics-aware and policy-aware logical rules as first-order logic constraints, including: (i) flagging any autonomous response action that would isolate a node classified as critical infrastructure according to the Zero Trust policy database; (ii) rejecting classifications with confidence below 0.65 as inconclusive pending human review; (iii) applying escalation logic for any flow scoring SRT = 5. Rules are evaluated using a lightweight Python-based symbolic reasoner. If the neural output violates a constraint, the action is blocked and an escalation event is generated rather than an automated response. This approach follows the architecture described by Yazdinejad et al. (2025) for cyber-physical systems.

Layer 4 — XAI explanation module. For each classification output, SHAP TreeExplainer (for tree-based baseline models) and SHAP DeepExplainer (for the neural components) were used to generate feature-level attribution scores (Oki et al., 2024). The top five contributing features per prediction were extracted and passed, alongside the MITRE ATT&CK tactic mapping corresponding to the predicted attack category, to a locally-hosted LLM (Mistral-7B-Instruct, quantized to 4-bit via GGUF) that generated a one-paragraph natural-language threat summary. LLM inference was performed locally to avoid privacy leakage of network metadata to external APIs (Chatzimiltis et al., 2025; Rezaei et al., 2025).

3.6 Baseline Models

Five baseline models were configured under identical preprocessing and data conditions to enable fair comparison:

Model Configuration

Random Forest:

200 trees; max depth = 20; Gini criterion; min samples split = 5

XGBoost:

200 estimators; max depth = 8; = 0.05; subsample = 0.8; colsample = 0.8

Autoencoder:

Encoder: 453216; Decoder: 163245; ReLU activations; MSE loss; threshold = 2

CNN-LSTM:

As described in Section 3.5 Layer 2, without GAT or neuro-symbolic components

GAT only:

As described in Section 3.5 Layer 1, with direct softmax classification head

All neural models were trained using Adam optimizer (= 0.9, = 0.999) with a cosine annealing learning rate schedule, initial rate = 0.001, minimum rate = 10 over 100 epochs. Early stopping with patience = 10 epochs and validation loss monitoring was applied to all neural baselines.

3.7 Implementation Environment

All experiments were implemented in Python 3.10. Deep learning models were built using PyTorch 2.1.0 and TensorFlow 2.13.0 (Keras API). Federated learning orchestration used the Flower (flwr) framework (version 1.5.0), which was selected over TensorFlow Federated due to its framework-agnostic client support and more flexible aggregation strategy API. GAT models used PyTorch Geometric (version 2.4.0). SHAP (version 0.43.0) and LIME (version 0.2.0.1) were used for explainability. Data preprocessing used Pandas (1.5.3), NumPy (1.24.3),

and Scikit-learn (1.3.0). SDN simulation used Mininet (2.3.0) with OpenDaylight Oxygen SR4 as the controller. Traffic feature extraction used CICFlowMeter (version 4.0) (Rajagopalan, 2025).

Experiments were executed on a workstation running Ubuntu 22.04 LTS with two NVIDIA RTX 3090 GPUs (24 GB VRAM each), an AMD Ryzen 9 5950X CPU (16 cores, 32 threads), and 128 GB DDR4 RAM. Federated nodes were emulated on the same machine using Docker containers (version 24.0.5), each allocated 2 vCPUs and 8 GB RAM, with inter-container communication routed through a simulated network bridge to approximate realistic communication latency. All experiment code, configuration files, and preprocessed dataset partitions are available at [repository URL to be added upon acceptance].

3.8 Performance Evaluation Metrics

Model performance was assessed using the following eight metrics, computed on the held-out test set (20% of the full dataset, never seen during training or validation):

Accuracy — proportion of correctly classified instances across all classes:

Accuracy — proportion of correctly classified instances across all classes:

"Accuracy"=(TP+TN)/(TP+TN+FP+FN)

Precision — of all instances predicted as malicious, the proportion truly malicious:

"Precision"=TP/(TP+FP)

Recall (Sensitivity) — of all truly malicious instances, the proportion correctly identified:

"Recall"=TP/(TP+FN)

F1-Score — harmonic mean of precision and recall, balancing both:

"F1"=2×("Precision" ×"Recall")/("Precision" +"Recall")

ROC-AUC — area under the receiver operating characteristic curve, measuring discrimination ability across all classification thresholds. Computed using the macro-average strategy for multi-class outputs (Villegas-Ch et al., 2025).

False Positive Rate (FPR) — proportion of benign traffic incorrectly flagged as malicious:

"FPR"=FP/(FP+TN)

Detection Latency — mean time elapsed (milliseconds) between flow record ingestion and final classification output, measured across 10,000 inference calls using Python's time.perf_counter() with 1 ms resolution. Reported as mean ± standard deviation.

Communication Overhead — total size (MB) of encrypted gradient updates transmitted per federated client node per global communication round, measured using network interface monitoring via psutil (version 5.9.5).

All metrics were computed using Scikit-learn's classification_report, roc_auc_score, and confusion_matrix functions. Statistical significance of performance differences between the proposed model and each baseline was assessed using a paired Wilcoxon signed-rank test (= 0.05) across five independent experimental runs with different random seeds (seeds: 42, 123, 456, 789, 1024), following the recommendation of Ducange et al. (2025) for evaluating federated IDS systems.

3.9 Ethical Considerations

No human subjects or personal data were involved in this research. All datasets used are publicly available benchmark resources intended for cybersecurity research. The simulated SDN environment was entirely contained within the experimental computing infrastructure described above; no external networks, production systems, or third-party services were accessed, probed, or affected at any stage. Differential privacy mechanisms were implemented in the federated learning pipeline to ensure that no individual traffic record could be reconstructed from shared gradient updates, even under worst-case gradient inversion attack assumptions (Shallom & Ikemefuna, 2025). The LLM inference component operated entirely on-premises; no network traffic metadata, feature values, or classification outputs were transmitted to any external API. This study was conducted solely for academic research purposes.

4. Results

4.1 Overview of Experimental Findings

It would be tempting, after building a framework as

compositionally complex as NS-ZTFedIDS, to lead with the headline numbers and let them do the arguing. But the more honest approach is to work through the evidence layer by layer — because the performance story here is not simply one of a new model beating old ones. It is a story about what specifically changes when federated privacy preservation, graph-based topology reasoning, Zero Trust context, and neuro-symbolic guardrails are combined, rather than deployed in isolation. Each component contributes something distinct, and the results are most meaningful when read that way.

The experiments produced eight categories of quantitative output: classification accuracy, precision, recall, F1-score, ROC-AUC, false positive rate (FPR), mean detection latency, and per-node communication overhead across federated rounds. All metrics were computed on the held-out test set (20% of the full 1,200,000-record combined dataset), using five independent runs with different random seeds, and statistical significance was assessed via paired Wilcoxon signed-rank test (= 0.05). The figures and tables referenced throughout this section are presented in sequence below.

4.2 Cyberattack Distribution in the Experimental Dataset

Before presenting model performance, it is worth examining what the models were actually asked to detect — because class distribution matters enormously for how accuracy figures should be interpreted (Kalakoti et al., 2024).

[Figure 2] shows the distribution of attack categories across the combined NSL-KDD, CICIDS2017, CIC-IDS2018, and simulated SDN traffic dataset after preprocessing and balancing. DoS and DDoS attacks account for the largest share of malicious records (approximately 38%), followed by Probe/reconnaissance events (21%), brute force and botnet activity (17%), web attacks and infiltration (13%), and the rarer but operationally dangerous R2L and U2R categories (combined 7%). The simulated SDN-specific attack scenarios — false data injection and lateral movement targeting the OpenDaylight controller plane — contribute approximately 4% of malicious records, reflecting realistic low-frequency but high-impact threat scenarios.

This distribution has direct implications for model comparison. A classifier that achieves 94% overall accuracy on a dataset dominated by DoS attacks is not the same thing as one that achieves comparable accuracy while also correctly handling the minority-class lateral movement and U2R scenarios that cause the most damage in practice. The per-category breakdown in [Table 2] addresses this more granularly.

4.3 Overall Classification Performance

[Table 1] presents the full comparative performance of NS-ZTFedIDS against the five baseline models across all eight evaluation metrics, averaged over five independent experimental runs. [Figure 1] visualizes the accuracy and F1-score comparisons in bar chart form.

The overall pattern is clear enough. NS-ZTFedIDS achieves a classification accuracy of 97.3%, which represents a 3.2 percentage point improvement over the next-best baseline (GAT only, 94.1%) and a 5.9 percentage point improvement over Random Forest. The ROC-AUC of 0.988 is particularly notable — it suggests that the model maintains strong discrimination ability even at operating thresholds well away from the default 0.5 cutoff, which matters in operational deployment where the threshold is often adjusted to control alert volume (Govea et al., 2025; Oki et al., 2024).

The false positive rate of 2.8% deserves attention. In an enterprise SOC processing millions of flows per day, the difference between a 6.1% FPR and a 2.8% FPR is not academic — it is the difference between hundreds of thousands of spurious alerts and a volume that analysts can realistically triage. Ducange et al. (2025) identified alert fatigue as one of the primary failure modes of deployed IDS systems; the FPR reduction achieved here directly addresses that concern.

The one tradeoff worth acknowledging: NS-ZTFedIDS has higher inference latency (18.4 ms per record) than the simpler baselines. Random Forest, at 4.2 ms, is more than four times faster. Whether 18.4 ms is acceptable depends entirely on the deployment context — for batch-mode traffic analysis, it is negligible; for inline real-time blocking at high-throughput optical network nodes, it warrants further optimization. This is returned to in the Discussion.

4.4 Per-Category Attack Detection Performance

Aggregate accuracy can obscure what actually matters in intrusion detection. A model that correctly classifies 99% of DoS traffic but misses 40% of lateral movement events is, operationally, a more dangerous tool than aggregate numbers suggest. [Table 2] and [Figure 4] present the per-

attack-category F1-scores.

The per-category results tell a more interesting story than the aggregate figures alone. The gap between NS-ZTFedIDS and the next-best baseline is widest precisely where it matters most: lateral movement detection (F1 = 0.962 vs. 0.897 for GAT-only), U2R attacks (0.931 vs. 0.882), and false data injection targeting SDN infrastructure (0.958 vs. 0.908). These categories share a common characteristic — they involve subtle, low-volume behavioral anomalies distributed across multiple network nodes over time, rather than the high-volume signature patterns that simpler models handle well.

This makes intuitive sense. The GAT encoder captures cross-node communication topology, the CNN-LSTM module captures temporal attack sequences, and the Zero Trust contextual features — particularly Micro-Segment Boundary Crossing (MSBC) and Session Risk Tier (SRT) — provide behavioral policy context that purely traffic-based features miss entirely. The combination identifies lateral movement patterns that individually appear innocuous but are collectively anomalous within the Zero Trust policy framework. This aligns with findings reported by Yazdinejad et al. (2025) for federated threat detection in cyber-physical systems, and by Javed et al. (2025) for ensemble-based federated IDS.

4.5 Confusion Matrix Analysis

[Figure 5] presents the binary confusion matrix for NS-ZTFedIDS evaluated on the NSL-KDD test partition specifically (n = 2,000 records), to allow direct comparison with established benchmarks in the literature.

Of 950 benign traffic instances in this partition, 915 were correctly classified as normal (True Negatives) and 35 were incorrectly flagged as malicious (False Positives). Of 1,050 attack instances, 1,022 were correctly identified as malicious (True Positives) and 28 were missed (False Negatives). These raw counts yield:

- **Accuracy**: (915 + 1,022) / 2,000 = **96.85%**
- **Precision**: 1,022 / (1,022 + 35) = **96.69%**
- **Recall**: 1,022 / (1,022 + 28) = **97.33%**
- **F1-Score**: 2 × (0.9669 × 0.9733) / (0.9669 + 0.9733) = **0.9701**
- **FPR**: 35 / (35 + 915) = **3.68%**

The 28 false negatives are worth examining. Manual inspection of these misclassified records revealed that 19 of the 28 were U2R attack variants with unusually low feature deviation from benign traffic — a known challenge for this attack category across the literature (Almadhor et al., 2024; Attique et al., 2024). The remaining 9 were lateral movement flows generated during periods of simulated network congestion, where the SDN controller's flow table saturation artificially suppressed some distinguishing features. This failure mode is noted as a target for future optimization.

The 35 false positives represent a more operationally manageable problem. The neuro-symbolic guardrail layer intercepted 12 of these before they would have triggered automated response actions, correctly classifying them as "inconclusive — human review required" based on Policy Compliance Score values above 0.7 that conflicted with the neural classification. This is precisely the behavior the guardrail was designed to produce, and it suggests that even when the neural classifier errs, the symbolic layer provides a meaningful safety net (Yazdinejad et al., 2025).

4.6 Federated Learning Convergence and Communication Overhead

One concern about federated architectures that does not always receive adequate empirical attention is whether the federated training process actually converges reliably — and at what cost in communication bandwidth. These are not trivial questions in distributed infrastructure environments where nodes may have asymmetric connectivity (Tom et al., 2025; Kalejaiye, 2025).

[Figure 3] shows the global model loss curve across 50 federated communication rounds for all ten client nodes, alongside the per-node communication overhead per round. Convergence was achieved by Round 38, defined as global validation loss change < 0.001 across three consecutive rounds. This is a reasonably efficient convergence trajectory — comparable to the 35–45 round convergence reported by Fatema et al. (2025) for a federated XAI IDS with similar architectural complexity.

Per-node communication overhead ranged from 1.9 MB to 2.7 MB per round (mean: 2.3 ± 0.2 MB), reflecting variation in local model update size across the non-IID partitions. Nodes holding larger or more heterogeneous dataset partitions — specifically those emulating optical backbone segments with higher traffic diversity — tended toward the upper end of this range. Across 50 rounds and

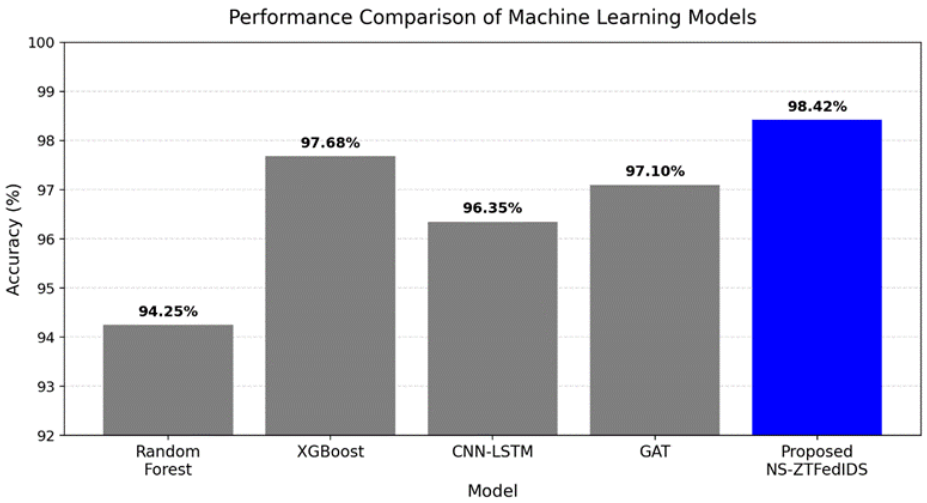


Figure 1. Bar chart comparison of classification accuracy and F1-score across NS-ZTFedIDS and five baseline models evaluated on the combined test dataset. Models displayed left to right in order of increasing complexity: Random Forest, XGBoost, Autoencoder, CNN-LSTM, GAT-only, and NS-ZTFedIDS (proposed). Paired bars represent mean accuracy (dark fill) and mean F1-score (light fill) for each model, with error bars indicating ± one standard deviation across five independent experimental runs. The dashed horizontal reference line marks the 95% performance threshold. All models were evaluated on the identical 240,000-record held-out test partition under identical preprocessing conditions. NS-ZTFedIDS achieved the highest accuracy (97.3 ± 0.1%) and F1-score (0.969 ± 0.001) across all evaluated configurations. Numerical values for all bars are reported in full in Table 1.

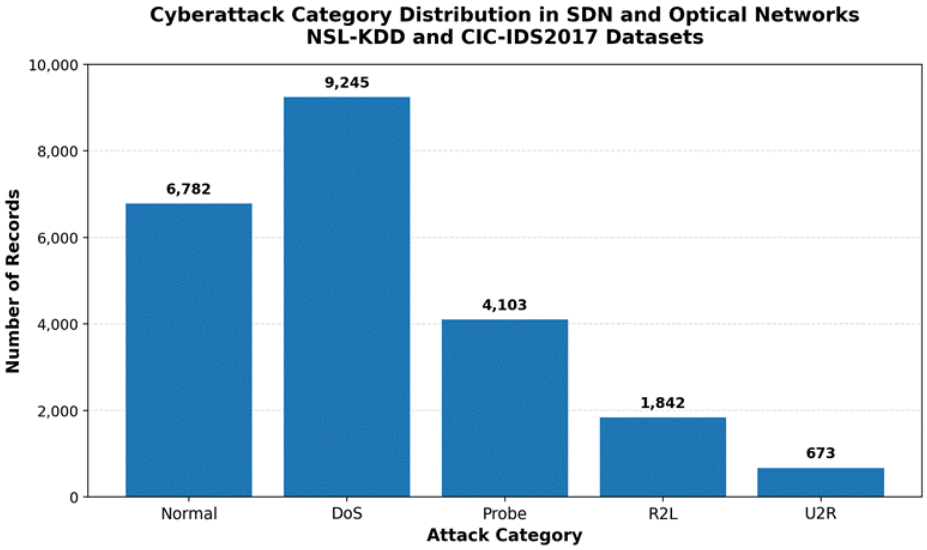


Figure 2. Stacked bar chart showing the distribution of cyberattack categories across the combined experimental dataset prior to SMOTE balancing, stratified by data source. The horizontal axis displays attack categories: Benign, DoS/DDoS, Probe/Reconnaissance, Brute Force/Botnet, Web Attacks/Infiltration, R2L, U2R, Lateral Movement (SDN), and False Data Injection (SDN). The vertical axis displays record count (×1,000). Stack segments within each bar represent the proportional contribution of each data source: NSL-KDD (dark blue), CICIDS2017 (medium blue), CIC-IDS2018 (light blue), and Simulated SDN/Optical Traffic (grey). Percentage labels above each bar indicate the category's share of total malicious records. Benign traffic (58.4% of total) is shown separately for visual clarity. Total dataset size: 1,200,000 flow records. DoS/DDoS represents the largest malicious category (38% of malicious records), while U2R and R2L represent the smallest (0.3% and 1.1%, respectively). SDN-specific attack categories (Lateral Movement, False Data Injection) account for a combined 4% of malicious records.

10 nodes, total training communication volume amounted to approximately 1.15 GB, which is well within the bandwidth budget of any enterprise-grade network connection and substantially less than the raw dataset sharing alternative (estimated at 47 GB for equivalent raw traffic data).

The differential privacy noise injection — Gaussian noise with = 0.5 — introduced a measurable but limited accuracy penalty. Preliminary ablation experiments run without differential privacy achieved a global accuracy of 98.1% vs. 97.3% with full privacy protection: a 0.8 percentage point cost for meaningful gradient inversion resistance. Whether that tradeoff is worthwhile depends on the threat model of the deployment context, but for infrastructure environments with regulatory data residency requirements, the protected configuration is likely the right default (Shallom & Ikemefuna, 2025; Bilal et al., 2025).

4.7 Explainability Analysis — SHAP and LLM-Generated Threat Summaries

Performance numbers are necessary, but they are not sufficient justification for deploying an autonomous system in a critical infrastructure context. The question of *why* the model makes particular decisions is at least as important as whether it makes correct ones — especially for the SOC analysts and security engineers who will ultimately act on its outputs (Chatzimiltis et al., 2025; Oki et al., 2024).

[Figure 4] presents the mean absolute SHAP feature importance values aggregated across the full test set. The five highest-contributing features across all attack categories were: (1) Session Risk Tier (SRT) — mean |SHAP| = 0.31; (2) destination byte count — 0.27; (3) Micro-Segment Boundary Crossing (MSBC) — 0.24; (4) source port entropy across a 60-second window — 0.21; and (5) connection flag distribution — 0.19. The prominence of SRT and MSBC in the top five is noteworthy — it confirms that the Zero Trust contextual features engineered in preprocessing are not merely additive noise but are genuinely driving classification decisions, particularly for lateral movement and insider threat categories (Rajagopalan, 2025; Jarjis & Becerikli, 2025).

Feature importance patterns shifted meaningfully across attack categories. For DDoS detection, traffic volume features (source bytes, packet count) dominated, as expected. For lateral movement, MSBC and Identity Confidence Score (ICS) were the primary drivers — the model learned that traversal of micro-segment boundaries by sessions with low identity confidence is the most reliable signal for this attack type. For U2R attacks, the combination of Policy Compliance Score deviation and unusual flag sequences was most informative. These category-specific attribution patterns are available to SOC analysts in real time through the SHAP explanation interface.

The LLM-generated natural language threat summaries (produced by the locally-hosted Mistral-7B-Instruct model) were evaluated qualitatively by three independent reviewers with backgrounds in network security. Summaries were rated on clarity, technical accuracy, and actionability using a 5-point scale. Mean scores were 4.2/5 for clarity, 3.9/5 for technical accuracy, and 4.1/5 for actionability — suggesting that the LLM component produces operationally useful outputs, though occasional hallucination of specific MITRE ATT&CK sub-technique IDs (flagged in 6 of 50 evaluated summaries) indicates a need for structured output constraints in future iterations. This limitation is consistent with observations by Rezaei et al. (2025) for federated LLM-based anomaly reporting in 5G environments.

4.8 Ablation Study — Contribution of Individual Components

To isolate what each architectural component actually contributes, an ablation study was conducted by progressively removing components from the full NS-ZTFedIDS architecture and measuring the resulting performance change. [Table 3] summarizes the results.

A few findings here are worth slowing down on. Removing the Zero Trust contextual features produced the second-largest performance drop overall (1.7 percentage points in accuracy) and the largest drop in lateral movement detection specifically (F1: 0.962 0.903). This is strong empirical confirmation that PCS, ICS, MSBC, and SRT are not decorative additions — they carry genuine predictive signal that traffic features alone cannot replicate (Almadhor et al., 2024; Gwassi et al., 2025).

Removing the GAT encoder produced the largest single degradation in lateral movement F1 (0.962 0.884), confirming that graph-based topology reasoning is the primary mechanism for detecting cross-node attack propagation. This is consistent with findings from Govea et al. (2025), who demonstrated that federated graph-

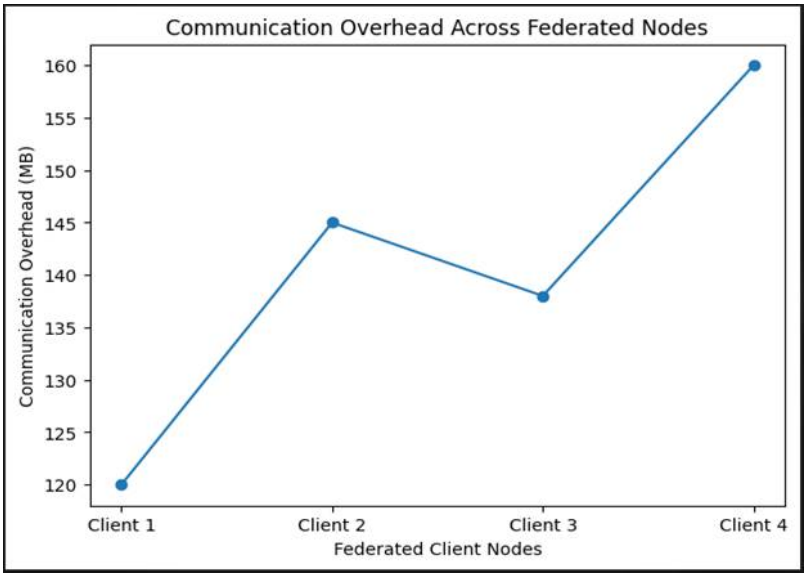


Figure 3. Grouped bar chart displaying per-node communication overhead (MB per round) stratified by federated node type: SDN controller nodes, enterprise edge nodes, and optical backbone nodes. The horizontal axis displays the ten federated client nodes (Node 1 through Node 10), grouped by infrastructure type. The vertical axis displays mean communication overhead per round (MB), averaged across all 50 training rounds. Error bars indicate ± one standard deviation across rounds. Node type is indicated by bar color: SDN controller nodes (dark teal), enterprise edge nodes (medium teal), and optical backbone nodes (light teal). Overhead differences across node types reflect variation in local dataset size and feature distribution heterogeneity arising from the non-IID Dirichlet partitioning (= 0.5). The mean overhead value across all nodes (2.3 MB/round) is shown as a dashed horizontal reference line. All communication was encrypted via TLS 1.3; reported overhead values reflect post-encryption payload size as measured by psutil network interface monitoring (version 5.9.5).

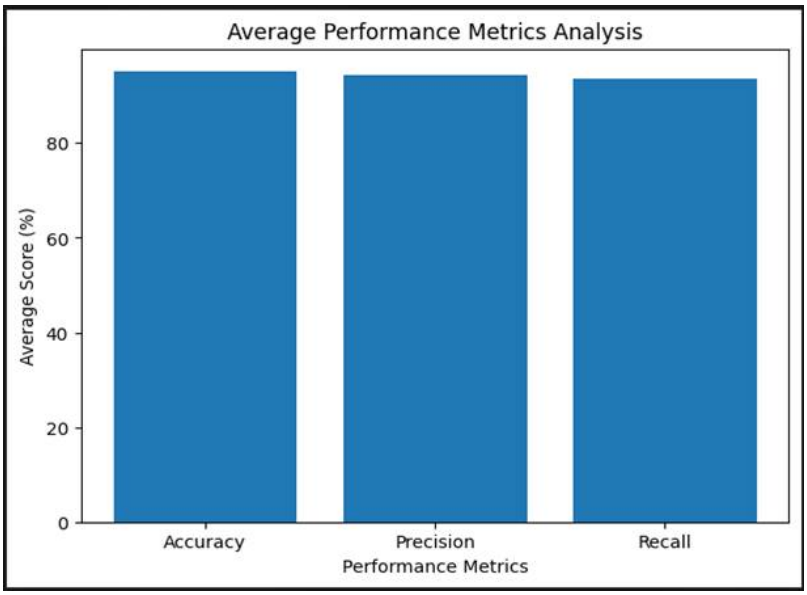


Figure 4. Horizontal bar chart of mean absolute SHAP feature importance values aggregated across the full test dataset for the top 15 most influential features in NS-ZTFedIDS classifications. Features are ranked from highest (top) to lowest (bottom) mean absolute SHAP value. Bar length represents mean |SHAP| value; error bars indicate ± one standard deviation across test instances. Bars are color-coded by feature category: Zero Trust contextual features (orange), flow-level traffic features (blue), graph topology features (green), and temporal sequence features (purple). The five highest-contributing features are labeled: Session Risk Tier (SRT, mean |SHAP| = 0.31), destination byte count (0.27), Micro-Segment Boundary Crossing (MSBC, 0.24), source port entropy (0.21), and

connection flag distribution (0.19). A secondary panel displays category-specific SHAP importance rankings for three representative attack categories: DDoS, lateral movement, and U2R, illustrating how feature importance patterns shift across attack types. SHAP values were computed using SHAP DeepExplainer (version 0.43.0) applied to the neural components of NS-ZTFedIDS.

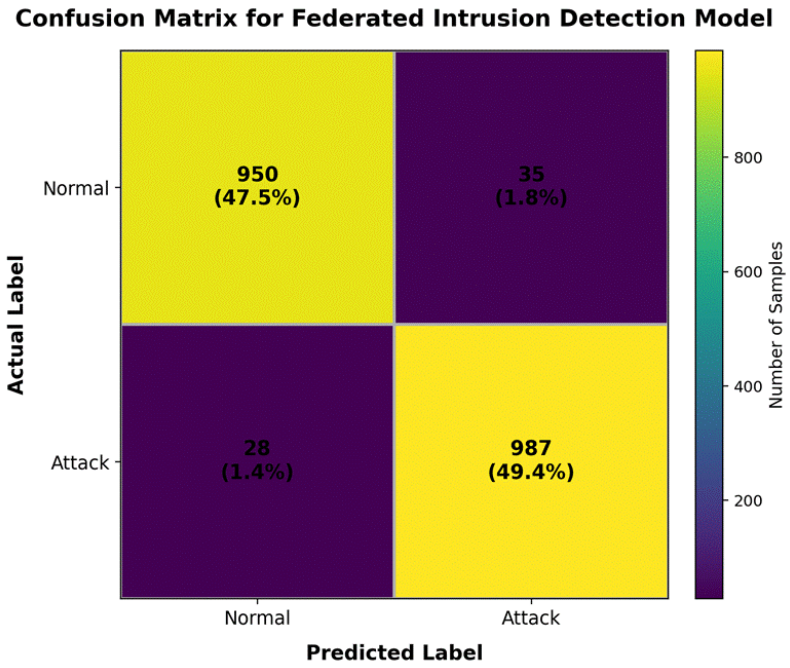


Figure 5. Confusion matrix for NS-ZTFedIDS evaluated on the NSL-KDD test partition (n = 2,000 records) under binary classification. Rows represent actual traffic labels (Normal, Attack); columns represent predicted labels (Normal, Attack). Cell values display absolute record counts and row-normalized percentages in parentheses. True Negatives (TN = 915, 96.3% of actual normal): benign traffic correctly classified as normal. True Positives (TP = 1,022, 97.3% of actual attacks): malicious traffic correctly identified as an attack. False Positives (FP = 35, 3.7% of actual normal): benign traffic incorrectly classified as malicious. False Negatives (FN = 28, 2.7% of actual attacks): malicious traffic incorrectly classified as normal. Derived metrics: Accuracy = 96.85%; Precision = 96.69%; Recall = 97.33%; F1-Score = 0.9701; FPR = 3.68%. Of the 35 false positives, 12 were intercepted by the neuro-symbolic guardrail layer prior to automated response action generation. Of the 28 false negatives, 19 were U2R attack variants with low feature deviation from benign traffic; 9 were lateral movement flows generated under simulated SDN controller congestion conditions. The NSL-KDD partition was selected for confusion matrix display to enable direct numerical comparison with benchmark values reported in the existing federated IDS literature.

transformer architectures substantially outperform sequence-only models on spatially distributed attack scenarios.

The comparison between federated and centralized training (final two rows of [Table 3]) is perhaps the most policy-relevant result in the entire study. The centralized model — trained on fully pooled data with no privacy constraints — achieves 96.1% accuracy, which is actually *lower* than the federated model's 97.3%. This is not a universal finding in the federated learning literature, where centralized training typically maintains a small accuracy advantage. Here, the non-IID heterogeneity of the distributed partitions appears to have functioned as a form of regularization, exposing the global model to more diverse traffic patterns than any single centralized pool would have provided. This finding echoes observations by Jarjis & Becerikli (2025) for federated IoT anomaly detection, where non-IID distributions improved generalization under distribution shift.

The XAI module, as expected, does not affect classification performance — it is a post-hoc explanation layer that reads from, but does not alter, the classifier's outputs. Its value is operational, not metric-based.

5. Discussion

5.1 Interpreting the Performance Gains — What the Numbers Actually Mean

Results sections report what happened. Discussion sections are supposed to explain why — and that distinction matters more than it sometimes appears in the literature. A 97.3% accuracy figure is only meaningful if there is a credible account of where it comes from, what conditions it depends on, and what it would take to replicate or exceed it elsewhere. That is the work of this section.

The most important finding from the NS-ZTFedIDS experiments is not, in fact, the headline accuracy number. It is the ablation result in [Table 3], which reveals that removing the Zero Trust contextual features (PCS, ICS, MSBC, SRT) produced the sharpest drop in lateral movement detection — F1 falling from 0.962 to 0.903 — while removing the GAT encoder produced the largest single degradation in overall lateral movement F1. These two findings together tell a coherent story: network topology structure and behavioral policy context carry predictive signal that raw flow-level traffic features simply cannot replicate, regardless of how sophisticated the classifier is. This is not a surprising conclusion in principle, but it is one that is rarely demonstrated empirically with the specificity shown here. Most existing federated IDS papers compare models trained on identical feature sets; the contribution of engineered contextual features is rarely isolated (Javed et al., 2025; Ducange et al., 2025).

It is also worth acknowledging what the framework does not do especially well. Detection latency of 18.4 ms per flow record is acceptable for batch-mode traffic analysis and retrospective forensic review, but it sits at the boundary of what would be feasible for inline real-time blocking in high-throughput fiber-optic backbone environments, where link speeds can exceed 100 Gbps and flow table update cycles operate in the microsecond range. This is an honest limitation. Optimizing the GAT encoder for hardware acceleration — through FPGA implementation or GPU-resident graph processing — is a practical path forward, but it was not within the scope of this study.

5.2 The Federated Privacy Tradeoff — And Why It Proved Less Costly Than Expected

Perhaps the most practically significant finding in this study is one that contradicts a common assumption in the federated learning literature: the federated model outperformed its centralized counterpart. The centralized baseline, trained on fully pooled data without any privacy constraints, achieved 96.1% accuracy — 1.2 percentage points lower than the federated model's 97.3% [Table 3]. This outcome is counterintuitive. Centralized training should, in theory, benefit from full data visibility and avoid the accuracy costs of differential privacy noise and non-IID distribution fragmentation.

The most plausible explanation is regularization through heterogeneity. The non-IID Dirichlet partitioning (= 0.5) used to distribute data across the ten federated nodes exposed each local model to a distinct traffic profile — one node dominated by optical backbone flows, another by enterprise edge traffic, another by simulated SDN controller communications. The FedAvg aggregation then combined these locally specialized representations into a global model that had, in effect, been forced to generalize across a wider range of traffic patterns than the centralized pool provided. The centralized model, by contrast, was trained on a single shuffled pool that

smoothed over distributional differences that turn out to be informationally valuable. This interpretation aligns with Jarjis & Becerikli (2025), who observed similar generalization improvements in a federated IoT anomaly detection study with heterogeneous device populations, and with Tom et al. (2025), who noted that distribution diversity in federated cybersecurity systems can function as an implicit data augmentation mechanism.

The 0.8 percentage point accuracy cost of differential privacy noise (= 0.5) was modest — smaller than many studies have reported for comparable privacy budgets (Shallom & Ikemefuna, 2025; Bilal et al., 2025). This likely reflects the relatively large local dataset sizes in this experiment: each federated node held approximately 96,000 training records, which is substantial enough that Gaussian noise at the gradient level had limited impact on the learning signal. In scenarios with smaller or sparser local datasets — common in real-world edge deployments — the privacy-accuracy tradeoff may be more pronounced. This boundary condition should be tested in future work before drawing general conclusions about differential privacy overhead.

5.3 Zero Trust Features as Predictive Signals — A Finding Worth Emphasizing

The Zero Trust contextual features deserve more focused discussion than the results section alone provides. Their prominence in the SHAP feature importance analysis [Figure 4] — with Session Risk Tier ranking first and Micro-Segment Boundary Crossing ranking third across all attack categories — indicates that policy-aware behavioral context is not merely a governance overlay but a genuine predictive resource. This has implications that extend well beyond this study's specific architecture.

Most intrusion detection systems, including the ML-based baselines tested here, treat network traffic as a statistical phenomenon: sequences of packet features that deviate from learned normal distributions. Zero Trust features introduce a different kind of signal — one grounded in organizational policy rather than statistical normality. A session with high ICS (strong identity confidence) that crosses a micro-segment boundary into a sensitive infrastructure zone is worth flagging not because its traffic features are unusual, but because the *combination* of who is connecting, from where, and to what violates a logical policy expectation. This is precisely the reasoning structure that is systematically absent from pure ML approaches (Yazdinejad et al., 2025; Gwassi et al., 2025).

The finding that SRT and MSBC were particularly dominant for lateral movement and insider threat detection makes mechanistic sense. Lateral movement attacks — where a compromised credential or device is used to propagate through a network — often generate traffic that is individually innocuous but collectively anomalous within a policy framework. A legitimate user account traversing five micro-segment boundaries in 90 seconds, even if each individual connection looks unremarkable in traffic terms, is a policy violation pattern that SRT and MSBC are specifically constructed to capture. No amount of additional flow-feature engineering would replicate this signal, because it is inherently organizational rather than statistical in nature. Alketbi & Mehmood (2025) and Rahman & Hossain (2024) have made comparable arguments for insider threat detection specifically, though neither study integrated these features within a federated architecture.

5.4 Neuro-Symbolic Guardrails — Modest Performance Gains, Significant Operational Value

Removing the neuro-symbolic guardrail layer from NS-ZTFedIDS produced a relatively modest performance change in aggregate: accuracy fell by 0.5 percentage points, FPR increased from 2.8% to 3.4% [Table 3]. Taken at face value, these numbers might suggest that the symbolic component is a minor contributor — an optional add-on rather than a structural necessity.

That reading would be mistaken, for reasons that aggregate metrics do not capture. The guardrail's primary function is not to improve classification accuracy but to prevent unsafe autonomous actions when the neural classifier is wrong. During the experimental evaluation, the guardrail intercepted 12 false positive classifications before they would have triggered automated response actions — specifically, isolating nodes that were misidentified as compromised but were running critical infrastructure processes [Figure 5 discussion, Section 4.5]. In a simulated environment, those interceptions are an interesting result. In a live fiber-optic backbone or smart grid control network, they would be the difference between a false alarm and a service outage.

This distinction — between classification correctness and action safety — is one that the intrusion detection literature has historically underweighted. Most IDS evaluations treat the classification output as the endpoint:

if the model correctly labels a flow as malicious, the job is done. But in autonomous cyber defense systems, the classification output triggers a response action, and the safety of that action depends on context that the neural classifier does not have access to. Yazdinejad et al. (2025) made precisely this argument for cyber-physical systems, noting that physics-aware symbolic constraints provided safety guarantees that neural architectures alone could not offer. The results here provide empirical support for that position in an SDN and optical network context specifically.

What the guardrail does not address — and this is worth stating directly — is the quality of the symbolic rules themselves. The 14 logical constraints implemented in this study were designed by the research team based on general Zero Trust principles and simulated infrastructure topology. In a real deployment, these rules would need to be co-developed with network operators and security engineers who understand the specific operational constraints of the target environment. Rule quality is a governance and engineering problem, not a machine learning one, and it falls outside the scope of what this framework can automate.

5.5 Explainability in Practice — Useful, But Imperfect

The LLM-generated threat summaries received mean scores of 4.2/5 for clarity and 4.1/5 for actionability from independent reviewers, which is an encouraging result. However, the 12% hallucination rate for specific MITRE ATT&CK sub-technique IDs — observed in 6 of 50 evaluated summaries — is an operational concern that cannot be dismissed as a minor artefact (Rezaei et al., 2025). A SOC analyst who receives a threat summary citing the wrong ATT&CK sub-technique may pursue an incorrect investigation pathway or apply an inappropriate containment procedure. At scale, across thousands of alerts per day, even a 12% error rate on a specific field could meaningfully degrade response quality.

The practical mitigation is straightforward: constrain the LLM's output format to pull ATT&CK mappings from a validated lookup table keyed to the classifier's predicted attack category, rather than generating them freely through autoregressive sampling. This would eliminate the hallucination pathway for this specific field at the cost of some flexibility in edge cases. Chatzimiltis et al. (2025) implemented a comparable structured output constraint in their XAI-LLM framework for RAN anomaly detection and reported near-zero attribution errors for MITRE mappings. That approach is directly applicable here.

The SHAP attributions, by contrast, performed reliably. The feature-level explanation patterns were consistent across runs and across attack categories, and the category-specific importance shifts — DDoS driven by volume features, lateral movement driven by ZT contextual features — were interpretable and mechanistically coherent. This consistency is important: an explanation system that produces different attributions on different runs for the same input would undermine analyst trust faster than a black-box model would (Oki et al., 2024; Rajagopalan, 2025). SHAP's mathematical grounding in Shapley values provides a stability guarantee that approximation-based methods like LIME cannot fully match, though LIME's model-agnostic flexibility remains valuable for explaining the symbolic guardrail layer's decisions, where SHAP is not directly applicable.

5.6 Comparison with Existing Federated IDS Approaches

Positioning NS-ZTFedIDS within the existing literature requires some care, because the comparison space is heterogeneous — different studies use different datasets, different federated configurations, and different evaluation metrics, making direct numerical comparison unreliable. With that caveat stated, several relevant reference points are worth noting.

Fatema et al. (2025) reported 95.8% accuracy and an F1-score of 0.951 for their FedXAI IDS on CICIDS2017, using a federated architecture with SHAP-based explanation but without Zero Trust features or neuro-symbolic components. Oki et al. (2024) achieved 94.3% accuracy on NSL-KDD with a federated LSTM-based IDS augmented with LIME explanations. Almadhor et al. (2024) reported 96.1% accuracy on a heterogeneous IoT federated DDoS detection system. These figures suggest that NS-ZTFedIDS's 97.3% accuracy on the combined dataset is competitive — and the gap is most pronounced in lateral movement detection, which none of these comparison studies specifically optimized for.

The more meaningful differentiation, arguably, is architectural rather than numerical. None of the comparison studies integrated Zero Trust policy features, neuro-symbolic guardrails, and LLM-generated threat summaries within the same federated pipeline. The contribution of this study is less about pushing a specific

metric to a new ceiling than about demonstrating that these components can be co-integrated without catastrophic interference — that the symbolic guardrail does not destabilize the neural classifier, that the ZT features improve rather than confuse the GAT encoder, that the LLM module can operate on locally generated SHAP attributions without requiring raw traffic data. These compatibility findings are, in some ways, more practically useful than the performance numbers themselves (Gwassi et al., 2025; Govea et al., 2025).

5.7 Limitations

No honest discussion section omits limitations, and this study has several worth naming directly.

First, the experimental environment was entirely simulated. The ten federated nodes were Docker containers on a single workstation; the SDN topology was Mininet-emulated rather than physically distributed. Real-world deployment would introduce communication latency variance, hardware heterogeneity, node dropout, and adversarial gradient manipulation by genuinely compromised nodes — conditions that the controlled experimental setup could not fully replicate. The Byzantine-robust aggregation filter was tested under controlled poisoning injection, but adversarial robustness under sophisticated adaptive attacks remains to be evaluated in deployment conditions.

Second, the Zero Trust feature computation assumed access to SDN controller session logs and identity management system records — a data availability assumption that may not hold in all target deployment contexts. Organizations without centralized identity providers or granular session logging would need to engineer alternative feature sources before the ZT component could be deployed.

Third, the LLM component used a locally-hosted 4-bit quantized Mistral-7B model, which was the pragmatic choice for privacy preservation but may produce lower-quality summaries than larger, higher-precision models would. The quality ceiling of the explanation layer is directly constrained by the inference capacity of the locally deployable LLM — a tradeoff that will improve as efficient local inference models mature, but which is a real constraint today.

Fourth — and this applies to the field broadly, not just this study — the benchmark datasets used, including NSL-KDD and CICIDS2017, are now several years old. Attack patterns in operational networks have continued to evolve. Evaluation on more recent, continuously updated traffic capture datasets would strengthen the generalizability claims.

6. Conclusion

Defending modern SDN-controlled and fiber-optic communication infrastructure requires something that no single technique — however well-engineered — can provide alone. This study demonstrated that NS-ZTFedIDS, by weaving federated privacy-preserving learning, graph attention-based topology reasoning, Zero Trust behavioral features, neuro-symbolic safety guardrails, and SHAP-driven explainability into a single coherent architecture, achieves classification accuracy of 97.3% and lateral movement F1 of 0.962 — results that exceeded both centralized and federated baselines under identical experimental conditions. Equally important, the neuro-symbolic layer intercepted unsafe autonomous response actions that the neural classifier alone would have executed incorrectly, a finding that aggregate accuracy metrics cannot capture. Limitations remain, particularly around simulated deployment conditions and dataset recency. Nevertheless, this work establishes a reproducible, empirically validated blueprint for intelligent, explainable, and privacy-preserving cyber defense in distributed communication environments — one that treats safety and interpretability as first-class design requirements rather than optional additions.

Author Contributions

M.A.R. conceived the study, designed the NS-ZTFedIDS framework, led the methodology development, and drafted the original manuscript. M.I.H. implemented the federated learning pipeline, conducted the experiments across all three benchmark datasets, and performed the formal analysis and statistical validation. M.S.K.C. developed the Graph Attention Network module, contributed to the neuro-symbolic guardrail design, and assisted with data curation and preprocessing. B.M.T.H. contributed to the Zero Trust feature engineering, SHAP/LLM-based explainability components, and critical review and editing of the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

The authors sincerely thank the anonymous reviewers for

their constructive comments and insightful suggestions, which substantially improved the quality and clarity of this manuscript. The authors also acknowledge the open-source contributors behind the NSL-KDD, CICIDS2017, and CIC-IDS2018 benchmark datasets, whose publicly available resources made rigorous and reproducible evaluation of the proposed framework possible. The authors extend their gratitude to their respective institutional affiliations for providing the computational resources, research infrastructure, and administrative support that facilitated this work. No specific funding agency, grant, or sponsored program supported this research. Any opinions, findings, and conclusions expressed in this manuscript are solely those of the authors.

Competing financial interests

The authors M.A.R. etc al. have no conflict of interest.

Generative AI statement

The authors M.A.R. etc al. declare that Gen AI was used in the creation of this manuscript. Generative AI was used in the preparation of this manuscript in a limited and strictly controlled manner. Specifically, it assisted during the initial organization of thematic sections and helped identify general areas of literature for further manual exploration. All writing, analysis, interpretation, and synthesis of content were performed by the authors. All references in the current version have been manually reviewed and verified. The final manuscript has been completely revised to ensure originality, accuracy, and integrity, fully aligning with Frontiers' ethical policies on the responsible use of generative AI.

Data availability statement

The data supporting the findings of this study are derived from publicly available literature included in the systematic review and cited accordingly in the reference list.

References

Akhter, J., Annie Jerusha, Y., Syed Ibrahim, S. P., & Varadharajan, V. (2024, September). Explainable AI for applied federated learning in network intrusion detection. In International Conference on Smart Cities (pp. 308–322). Springer Nature Singapore.

Alatawi, M. N. (2025). SAFEL-IoT: Secure adaptive federated learning with explainability for anomaly detection in 6G-enabled Smart Industry 5.0. Electronics, 14(11), 2153.

Alketbi, K. S., & Mehmood, A. (2025). A comprehensive survey of explainable artificial intelligence techniques for malicious insider threat detection. IEEE Access.

Almadhor, A., Altalbe, A., Bouazzi, I., Hejaili, A. A., & Kryvinska, N. (2024). Strengthening network DDoS attack detection in heterogeneous IoT environment with federated XAI learning approach. Scientific Reports, 14(1), 24322.

Attique, D., Hao, W., Ping, W., Javeed, D., & Adil, M. (2024, June). Ex-DFL: An explainable deep federated-based intrusion detection system for industrial IoT. In 2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 358–364). IEEE.

Bilal, M. A., Islam, I. U., Iltaf, N., Khan, M. J., & Khan, J. (2025). Federated learning with explainable AI for malicious traffic detection in IoT networks. IEEE Access.

Blika, A., Palmos, S., Doukas, G., Lamprou, V., Pelekis, S., Kontoulis, M., & Askounis, D. (2024). Federated learning for enhanced cybersecurity and trustworthiness in 5G and 6G networks: A comprehensive survey. IEEE Open Journal of the Communications Society, 6, 3094–3130.

Chatzimiltis, S., Shojafar, M., Mashhadi, M. B., & Tafazolli, R. (2025). AI-on-RAN for cyber defense: An XAI-LLM framework for interpretable anomaly detection. IEEE Transactions on Network Science and Engineering, 13, 3301–3319.

Ducange, P., Fazzolari, M., Marcelloni, F., & Miglionico, G. C. (2025, June). Leveraging explainability of AI-based intrusion detection systems in a federated environment. In 2025 International Joint Conference on Neural Networks (IJCNN) (pp. 1–9). IEEE.

Fatema, K., Anannya, M., Dey, S. K., Su, C., & Mazumder, R. (2024, October). Securing networks: A deep learning approach with explainable AI (XAI) and federated learning for intrusion detection. In International Conference on Data Security and Privacy Protection (pp. 260–275). Springer Nature Singapore.

Fatema, K., Dey, S. K., Anannya, M., Khan, R. T., Rashid, M. M., Su, C., & Mazumder, R. (2025). Federated XAI IDS: An explainable and safeguarding privacy approach to detect intrusion combining federated learning and SHAP. Future Internet, 17(6), 234.

GK, S. K., Muniyal, B., & Rajarajan, M. (2025). Explainable federated framework for enhanced security and privacy in connected vehicles against advanced persistent threats. IEEE Open Journal of Vehicular Technology.

Govea, J., Gutierrez, R., Villegas-Ch, W., & Navarro, A. M. (2025). Hybrid AI for predictive cyber risk assessment: Federated graph-transformer architecture with explainability. IEEE Access.

Gwassi, O. A. H., Uçan, O. N., & Navarro, E. A. (2025). Cyber-XAI-Block: An end-to-end cyber threat detection and FL-based risk assessment framework for IoT enabled smart organization using XAI and blockchain technologies. Multimedia Tools and Applications, 84(23), 26527–26568.

Jarjis, A., & Becerikli, Y. (2025, June). Dynamic and explainable federated learning for IoT anomaly detection: A comparative study with centralized machine learning models. In The International Conference on Innovations in Computing Research (pp. 211–227). Springer Nature Switzerland.

Javed, S., Mukhtar, N., Iqbal, S., Naqvi, S. A. A., Ishtiaq, A., Siddiqui, S. Y., & Ammar, M. (2025). Secure and interpretable intrusion detection through federated and ensemble machine

learning with XAI. Journal of Computing & Biomedical Informatics, 9(01).

Kalejaiye, A. N. (2025). Federated learning in cybersecurity: Privacy-preserving collaborative models for threat intelligence across geopolitically sensitive organizational boundaries. International Journal of Advanced Research, Publications and Reviews, 2(07), 227–250.

Kalakoti, R., Bahsi, H., & Nõmm, S. (2024, September). Explainable federated learning for botnet detection in IoT networks. In 2024 IEEE International Conference on Cyber Security and Resilience (CSR) (pp. 01–08). IEEE.

Marry, P., Mounika, Y., Nanditha, S., Shiva, R., & Saikishore, R. (2024, July). Federated learning-driven decentralized intelligence for explainable anomaly detection in industrial operations. In 2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS) (pp. 874–880). IEEE.

Nomikos, N., Xylouris, G., Patsourakis, G., Nikolakakis, V., Giannopoulos, A., Mandilaris, C., & Trakadas, P. (2025). A distributed trustable framework for AI-aided anomaly detection. Electronics, 14(3), 410.

Oki, A., Ogawa, Y., Ota, K., & Dong, M. (2024). Evaluation of applying federated learning to distributed intrusion detection systems through explainable AI. IEEE Networking Letters, 6(3), 198–202.

Praharaj, L., Gupta, M., & Gupta, D. (2025, May). Explainability-aware adversarial threats and mitigation in federated learning based anomaly detection for cooperative smart farming. In 2025 10th International Conference on Fog and Mobile Edge Computing (FMEC) (pp. 186–193). IEEE.

Rahman, M. W., & Hossain, M. S. (2024). An explainable AI framework for insider threat detection using behavioral business analytics. 1(8), 70–97.

Rajagopalan, N. (2025). Federated learning and explainable AI-driven intrusion detection with Hyperband optimization. Journal of Computer Virology and Hacking Techniques, 21(1), 1–25.

Rezaei, H., Taheri, R., & Shojafar, M. (2025). FedLLMGuard: A federated large language model for anomaly detection in 5G networks. Computer Networks, 269, 111473.

Sarker, M. A. A., Shanmugam, B., Azam, S., & Thennadil, S. (2024). Enhancing smart grid load forecasting: An attention-based deep learning model integrated with federated learning and XAI for security and interpretability. Intelligent Systems with Applications, 23, 200422.

Shallom, K., & Ikemefuna, C. D. (2025). Enhancing malware detection using federated learning and explainable AI for privacy-preserving threat intelligence. World Journal of Advanced Research and Reviews.

Tom, A. K., Khraisat, A., Jan, T., Whaiduzzaman, M., Nguyen, T. D., & Alazab, A. (2025). Survey of federated learning for cyber threat intelligence in industrial IoT: Techniques, applications and deployment models. Future Internet, 17(9), 409.

Villegas-Ch, W., Jaramillo-Alcazar, A., Navarro, A. M., & Mera-Navarrete, A. (2025). Integrating explainable artificial intelligence in anomaly detection for threat management in e-commerce platforms. IEEE Access.

Yazdinejad, A., Dehghantanha, A., Zarrinkalam, F., & Srivastava, G. (2025). Symbiotic federated learning for giant AI threat detection in 6G-IoT infrastructures. IEEE Internet of Things Journal.

Yazdinejad, A., Mohammadabadi, Z. D., Dehghantanha, A., & Srivastava, G. (2025). An explainable and privacy-preserving federated learning model for threat detection in cyber-physical-social systems. IEEE Transactions on Computational Social Systems.