

ST-Fedxids: A Spatiotemporal Federated Explainable Intrusion Detection System With Drift-Adaptive Graph Learning For High-Density Public Wifi Environments

Quang-Vinh Dang¹, Dat Le², Minh Ngoc Dinh³, Ngoc-Son-An Nguyen⁴, Thi-Bich-Diem Vo⁵

¹ School of Innovative and Computing Sciences, British University Vietnam, Hung Yen 160000, Vietnam
vinh.dq4@buv.edu.vn

² Department of Science, Technology, and International Projects, Ho Chi Minh City University of Economics and Finance, Ho Chi Minh City, 700000, Vietnam, datla@uef.edu.vn

³ School of Data Science and IT, Millennia University, Ho Chi Minh City, 700000, Vietnam
minh.dinh@maeducation

⁴ Faculty of Information Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, 700000, Vietnam annns25871@pgr.iuh.edu.vn

⁵ Giao Hang Nhanh, Ho Chi Minh City, 700000, Vietnam, misdiem@gmail.com

Abstract

Public WiFi networks deployed across high-density transit environments — such as train stations, bus terminals, and airports — present a largely unaddressed frontier in network intrusion detection research. These venues collectively serve tens of millions of users monthly yet exhibit extreme traffic heterogeneity, continuous attack pattern drift, and strict data privacy constraints that preclude centralized detection paradigms. Existing federated learning-based intrusion detection systems (FL-IDS) fail to simultaneously address the spatiotemporal dynamics of evolving network topologies, the interpretability requirements of operational security teams, the concept drift inherent to live deployments, and the fairness disparities that arise across heterogeneous venue types. We propose **ST-FedXIDS**, a Spatiotemporal Federated Explainable Intrusion Detection System that unifies five technical contributions: (1) a heterogeneous dynamic graph construction pipeline that models multi-protocol WiFi traffic as sliding temporal-window graphs; (2) ST-GraphSAGE with multi-head temporal self-attention, replacing the LSTM/GRU paradigm used in all prior spatiotemporal GNN-IDS works; (3) Drift-Adaptive Federated Averaging (DA-FedAvg) that embeds ADWIN-based graph embedding drift detection directly into the federated aggregation protocol; (4) a GNN-Explainer module that produces subgraph-level attack attribution explanations for operational incident response; and (5) a venue-type fairness regularizer that minimizes the maximum inter-venue false positive rate gap across heterogeneous deployment sites. Extensive experiments on three public benchmarks — CSE-CIC-IDS2018, TON_IoT, and UNSW-NB15 — against four 2025 state-of-the-art baselines demonstrate that ST-FedXIDS achieves 98.21%–99.13% accuracy, outperforming all competitors by 0.9–2.4 percentage points, reduces the inter-venue FPR gap from 8.5–12.3% to 2.1%, and recovers from concept drift in 8 federation rounds versus 19–25 rounds for competing methods. An ablation study validates the independent contribution of each component.

Keywords: Intrusion Detection System; Federated Learning; Graph Neural Networks; Spatiotemporal Learning; Explainable AI; Concept Drift; Public WiFi Security; Fairness-Aware Machine Learning

1. Introduction

The proliferation of public WiFi infrastructure at mass-transit hubs has created a critical and underserved challenge at the intersection of network security and distributed machine learning. Transit WiFi deployments — encompassing metropolitan rail systems, inter-city bus terminals, and international airports — collectively constitute some of the largest shared network environments in existence. A single rail system can sustain 1–2 million authenticated sessions per month across dozens of access points, generating traffic volumes and behavioral patterns that differ qualitatively from the enterprise and IoT environments that dominate the IDS literature. Three structural properties make transit WiFi networks uniquely difficult to protect. First, traffic heterogeneity is extreme: a single venue may simultaneously serve commuters streaming video, tourists browsing maps, vendors conducting point-of-sale transactions, and IoT sensors reporting environmental data. The resulting traffic distribution differs markedly from one venue type to another, making shared detection thresholds a source of systematic fairness failures. Second, attack patterns evolve continuously: adversaries adapt their techniques in response to defenses deployed at prior venues, inducing concept drift that degrades models trained on historical

data. Third, privacy regulations in most jurisdictions prohibit centralizing raw traffic records, mandating a privacy-preserving federated architecture.

Existing approaches address subsets of these requirements but leave critical gaps. Federated GNN-IDS systems such as FedGATSage [1], FedSTGCN [2], and CONTINUUM [3] combine graph learning with federated training but omit temporal self-attention, have no drift adaptation mechanism, and provide no explainability output. Explainability-focused works such as xNIDS [4] and E-GNNExplainer [5] operate on centralized data and produce feature-level attributions rather than subgraph-level explanations. Concept drift adaptation approaches such as ReCDA [6] and INSOMNIA [7] are neither federated nor graph-based. Fairness-aware federated learning methods such as FairFed [8] and q-FFL [9] define fairness across demographic client groups rather than across heterogeneous venue types and do not address false positive rate equalization. To the best of our knowledge, no prior work simultaneously addresses spatiotemporal graph modeling, federated privacy, temporal drift adaptation, subgraph-level explainability, and venue-type fairness within a single system.

We address this gap with ST-FedXIDS (Spatiotemporal Federated Explainable Intrusion Detection System), an integrated architecture motivated by real operational requirements of multi-venue public WiFi deployments. The system is designed around five contributions:

- **Heterogeneous Spatiotemporal Graph Construction:** WiFi traffic is modeled as a dynamic heterogeneous graph with typed edges (protocol-level connections) and sliding temporal windows of 60 seconds, enabling the capture of both intra-window structural anomalies and cross-window behavioral continuity.
- **ST-GraphSAGE with Temporal Self-Attention:** A novel GNN backbone that augments GraphSAGE neighborhood aggregation with a multi-head temporal self-attention layer across historical graph snapshots, learning which past temporal states are most predictive of current attack patterns.
- **Drift-Adaptive Federated Averaging (DA-FedAvg):** An extension of FedAvg that monitors per-client graph embedding drift using the ADWIN algorithm and responds by selectively updating drifted model layers at elevated learning rates while freezing stable layers, preventing catastrophic forgetting.
- **GNN-Explainer for Operational Attack Attribution:** Each classification decision is accompanied by a minimal subgraph explanation identifying the network nodes and edges most responsible for the detection, supporting actionable incident response without requiring security staff to interpret raw neural network weights.
- **Venue-Type Fairness Regularization:** A loss term that minimizes the maximum false positive rate disparity across venue types (e.g., bus terminals vs. rail stations), ensuring equitable protection rather than aggregate accuracy maximization.

The remainder of this paper is structured as follows. Section 2 reviews related work across five dimensions. Section 3 presents the system model and threat model. Section 4 describes the ST-FedXIDS architecture in detail. Section 5 reports experimental results. Section 6 discusses findings and limitations. Section 7 concludes.

2. Related Work

2.1 Federated Learning for Intrusion Detection

The application of federated learning to network intrusion detection has accelerated considerably since 2021. Lavaur et al. [10] established a reference taxonomy of 22 FL-IDS systems, identifying communication efficiency and non-IID data handling as primary challenges. A subsequent ACM Computing Surveys analysis by Campos et al. [11] of over 100 works found that the majority employed unrealistic IID data splits, understating the difficulty of federated IDS in real deployments. Adewole et al. [12] reviewed 88 studies through 2024 and highlighted graph neural networks as significantly underrepresented in the FL-IDS literature relative to their demonstrated performance advantages in centralized settings.

Recent work has begun combining GNNs with federated learning for IDS. FedGAT [13] employs fingerprint-weighted Jaccard similarity for graph attention in a federated setting. FedGATSage [1] integrates client-side GAT with server-side GraphSAGE through Louvain community abstraction, reducing communication by 85% while improving detection of coordinated DDoS attacks. FedSTGCN [2] is the first system to integrate spatiotemporal GCN with federated learning for IoT IDS, employing a GCN + GRU architecture. FeCoGraph [14] introduces federated graph contrastive learning for few-shot IDS, addressing the label scarcity problem. All of these systems lack explicit drift adaptation mechanisms, venue-type fairness constraints, and subgraph-level explainability.

2.2 Graph Neural Networks for Intrusion Detection

E-GraphSAGE [15] established the foundational paradigm of applying GraphSAGE with edge features to network flow graphs, achieving strong performance on IoT datasets. Subsequent works have extended this paradigm in multiple directions: DLGNN [16] uses line graph transformations with GCNII and GRU for evolving spatiotemporal graphs; Anomal-E [17] applies Deep Graph Infomax for self-supervised GNN-IDS; and TCG-IDS [18] introduces temporal contrastive GNN learning as the first self-supervised spatiotemporal GNN-IDS. Comprehensive surveys by Bilot et al. [19] and Zhong et al. [20] identify as major open problems the absence of temporal self-attention (all

prior works rely on LSTM/GRU/fixed GCN), the lack of integration with federated learning, and the failure to provide node/subgraph-level explanations.

2.3 Explainable AI for IDS

Explainability research in IDS has primarily focused on feature-level attribution using SHAP and LIME [21]. xNIDS [4] produces actionable defense rules via sparse group lasso on feature dependencies. ROULETTE [22] applies learnable visual attention on image-converted network flows. E-GNNExplainer [5] is the closest prior work to our approach, adapting GNNExplainer to edge-level intrusion classification, but operates on centralized data and does not address federated deployment. Shyaa et al. [23] confirm in a 2025 review that GNN-based explanation methods remain underexplored, and no system combines subgraph explanation with federated training for IDS.

2.4 Concept Drift in Network Intrusion Detection

Concept drift is a well-documented challenge in deployed IDS systems: attack behaviors evolve, benign traffic patterns shift seasonally, and new device types join networks. ADWIN [24] is the dominant streaming drift detector, demonstrating robust performance in IDS contexts (0.99 accuracy in [25]). ReCDA [26] is the current state-of-the-art for drift adaptation, using self-supervised representation enhancement. INSOMNIA [27] combines incremental semi-supervised learning with XAI for drift interpretation. However, no existing work integrates drift detection with the federated aggregation protocol itself — specifically, no system adjusts per-layer learning rates in response to detected embedding drift across federated clients.

2.5 Fairness in Federated Learning

Federated fairness research has predominantly addressed demographic group fairness and client-level performance equity. Agnostic Federated Learning [28] proposes minimax optimization for worst-case client performance. q-FFL [9] applies alpha-fairness-inspired loss reweighting. FairFed [8] introduces server-side fairness-aware aggregation for group-level fairness. None of these frameworks define fairness in terms of heterogeneous deployment site types, and none target false positive rate equalization as the primary fairness objective — a critical requirement in transit WiFi environments where disproportionate false alarms at one venue type impose operational costs on security staff and erode user trust.

3. System Model and Threat Model

3.1 Deployment Architecture

We consider a multi-venue public WiFi deployment comprising V venues indexed $v = 1, \dots, V$, each operating as an independent federated client. Each venue maintains a local network of Access Points (APs), user devices, and IoT sensors. A central federated coordination server aggregates model updates from all venue clients without accessing raw traffic data. The deployment matches the architecture of large-scale transit WiFi systems serving 1–2 million monthly users across 50+ venues.

Each client v observes a continuous stream of network flows captured by its local WiFi controller. Flows are featurized using standard CICFlowMeter features (packet length statistics, inter-arrival times, flag counts, flow duration, byte rates) augmented with session-level entropy and AP association metadata. The client constructs a dynamic heterogeneous graph from these flows within sliding temporal windows of duration $T = 60$ seconds, with $W = 10$ historical windows retained for temporal reasoning.

3.2 Threat Model

ST-FedXIDS targets the detection of seven major attack categories documented in the experimental datasets: (1) Distributed Denial of Service (DDoS) via volumetric flooding; (2) Denial of Service (DoS) via targeted service exhaustion; (3) Network scanning and reconnaissance; (4) Botnet command-and-control communication; (5) Brute force credential attacks (SSH, FTP, HTTP); (6) Web application attacks (SQL injection, XSS); and (7) Advanced infiltration and exfiltration attempts. The system assumes honest-but-curious federated clients and does not address Byzantine-resilient aggregation, which is treated as complementary future work.

We assume attackers may employ concept drift as a deliberate evasion strategy — gradually morphing attack traffic to resemble benign patterns — and model this through injected drift experiments in Section 5. We further assume attack prevalence and behavioral profiles may differ significantly across venue types, motivating the fairness regularization component.

4. ST-FedXIDS Architecture

Figure 1 provides a conceptual overview of the ST-FedXIDS architecture. Each venue client executes a four-stage local pipeline: graph construction, spatiotemporal GNN inference, drift monitoring, and explanation generation. The coordination server executes DA-FedAvg aggregation with fairness-aware weighting. We describe each component in detail below.

4.1 Heterogeneous Spatiotemporal Graph Construction

For venue client v at time step t , we construct a heterogeneous graph $G_t = (V_t, E_t, X_t, R)$ from the current 60-second window of observed flows. Nodes in V_t represent active network entities: user devices (classified as mobile, laptop, IoT sensor), Access Points, and gateway routers. Edge types in R capture protocol-specific interaction patterns: {TCP_flow, UDP_flow, DNS_query, HTTP_request, HTTPS_session, ICMP_packet}. This typed edge structure enables the model to learn that ICMP floods, DNS amplification, and TCP SYN floods have qualitatively different graph signatures.

Node feature vectors x_i in X_t (dimensionality $d = 79$) concatenate: per-device 60-second aggregate statistics (mean/std of packet lengths, inter-arrival times, byte rates), connection state entropy, AP-level load metrics, and protocol distribution histograms. Edge features $e_{\{ij\}}$ encode the flow-level CICFlowMeter features for the connection between devices i and j . Cross-window edge persistence is tracked via an edge age decay function: edges from window $t-k$ receive multiplicative weight $\exp(-\lambda * k)$ to model behavioral continuity across recent temporal windows.

A sequence of $W = 10$ graph snapshots $\{G_{t-W+1}, \dots, G_t\}$ is maintained in a sliding window buffer at each client, providing the temporal context for the self-attention layer described in the next section.

4.2 ST-GraphSAGE with Temporal Self-Attention

The core detection model at each client is ST-GraphSAGE, which processes the temporal graph sequence in two stages: spatial aggregation and temporal self-attention.

4.2.1 Spatial GraphSAGE Aggregation

For each graph snapshot G_t , we apply $L = 3$ layers of GraphSAGE neighborhood aggregation. At layer l , the embedding of node i is updated as:

$$h_i^{(l)} = \sigma(W^{(l)} \cdot \text{CONCAT}(h_i^{(l-1)}, \text{AGG}(\{h_j^{(l-1)} : j \in N(i)\}))$$

where AGG is a mean aggregator, σ is LeakyReLU activation, and $W^{(l)}$ are learned weight matrices. Hidden dimensionality is set to 256 throughout. To handle the heterogeneous edge types in R , we employ edge-type-specific weight matrices following the R-GCN formulation, with edge feature concatenation into the aggregation step.

4.2.2 Multi-Head Temporal Self-Attention

After obtaining graph-level summary embeddings $\{z_{t-W+1}, \dots, z_t\}$ via mean pooling of node embeddings from each snapshot, we apply a 4-head temporal self-attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) \cdot V$$

where Q, K, V are linear projections of the temporal embedding sequence $\{z_{t-W+1}, \dots, z_t\}$. Crucially, we incorporate a learnable anomaly-score bias: historical snapshots with higher anomaly scores receive elevated attention weights, allowing the model to focus on suspicious behavioral episodes in the recent past when classifying the current window. The final representation z_t^* is a weighted combination of all historical snapshots and is passed to a softmax classification head for multi-class attack detection.

This design is architecturally distinct from all prior spatiotemporal GNN-IDS works: DLGNN uses GRU recurrence, FedSTGCN uses a standard STGCN with GRU, and CONTINUUM stacks GraphSAGE with LSTM. None employs temporal self-attention, which we demonstrate provides superior drift resilience because self-attention weights can rapidly down-weight stale historical contexts when attack patterns change.

4.3 Drift-Adaptive Federated Averaging (DA-FedAvg)

Standard FedAvg aggregates client model updates as a weighted average of local model parameters, with weights proportional to local dataset size. This aggregation strategy is insensitive to concept drift: clients experiencing drift in their local data distribution still contribute equally to the global model, potentially degrading aggregate performance. DA-FedAvg addresses this through three mechanisms.

4.3.1 Per-Client Graph Embedding Drift Detection

At each federation round r , client v computes its graph-level embedding distribution $D_v^r = \{z_i : i \text{ in batch}\}$ and transmits the distribution mean μ_v^r and variance σ_v^r to the server alongside model parameters. The server applies the ADWIN (ADaptive WINdowing) algorithm to the L2-distance sequence $\|\mu_v^r - \mu_v^{r-1}\|$ across rounds to detect when client v has experienced a statistically significant shift in its embedding distribution. ADWIN maintains an adaptive-length window over a data stream and raises a drift alert when the difference in means across sub-windows exceeds a statistical threshold.

4.3.2 Selective Layer Isolation

When a drift alert is raised for client v , DA-FedAvg applies selective parameter isolation: the learning rate for the final 2 classification layers of client v is elevated by factor $\alpha = 5$ for the subsequent 3 rounds, while the early feature extraction layers are frozen at the global model parameters. This design is motivated by the observation

that concept drift in attack traffic primarily manifests as changes in high-level feature combinations (captured by deep layers) rather than low-level structural properties (captured by early GNN layers), which are more stable across time.

4.3.3 Drift-Weighted Aggregation

The server aggregates client model parameters using weights that discount high-drift clients:

$$w_v = n_v / N \cdot (1 - \lambda_{\text{drift}} \cdot \text{drift_score}_v)$$

where n_v is the local dataset size, N is the total size, drift_score_v in $[0,1]$ is the ADWIN-normalized drift magnitude, and $\lambda_{\text{drift}} = 0.3$ is a regularization coefficient. This prevents heavily drifted clients from corrupting the global model while their local models adapt.

4.4 GNN-Explainer for Operational Attack Attribution

For each detected intrusion, ST-FedXIDS generates a subgraph-level explanation using an adapted GNN-Explainer. The explainer identifies the minimal connected subgraph $G_S = (V_S, E_S)$ that maximally influences the classification decision for the flagged flow. Formally, the explanation optimizes:

$$G_S^* = \operatorname{argmin}_{\{G_S \subseteq G\}} H(Y | G_S) + \beta \cdot |E_S|$$

where $H(Y | G_S)$ is the conditional entropy of the predicted label given the subgraph and the second term penalizes explanation size for compactness. We extend the standard GNNExplainer formulation to heterogeneous graphs by applying edge-type-specific masks and adapting the mutual information objective to preserve type-conditional explanations.

The explanation output for each alert is presented in human-readable form to security operators: "Detection: DDoS — Origin subnetwork: AP-Sector-7 → Device cluster [192.168.x.x/24], abnormal 3-hop broadcast pattern, 127s duration, 840 Mbps peak, matching historical DDoS behavioral template." This design addresses a well-documented gap in operational IDS deployments where security teams cannot act on black-box alerts without understanding which network entities are involved.

4.5 Venue-Type Fairness Regularization

In multi-venue federated IDS deployment, optimizing aggregate detection accuracy can produce systematically different false positive rates across venue types. A model calibrated on the high-bandwidth, session-dense traffic of a rail station may flag normal IoT sensor communications at a smaller bus depot as suspicious, creating disparate alarm burdens across venue operators.

We define venue-type fairness as the constraint that the maximum inter-venue false positive rate gap $\Delta_{\text{FPR}} = \max_{\{v,v'\}} |\text{FPR}_v - \text{FPR}_{v'}|$ should not exceed a threshold ϵ (set to 5% in our experiments). To enforce this during training, we add a fairness regularization term to the local classification loss at each client:

$$L_{\text{total}} = L_{\text{cls}} + \gamma \cdot \max_{\{v' \in N_v\}} |\text{FPR}_v - \text{FPR}_{v'}|$$

where $\gamma = 0.1$ is the fairness weight and N_v is the set of venue types sharing the current global model. The server communicates per-venue FPR estimates from the previous round to all clients at each aggregation step, enabling this penalty without sharing raw traffic data. This formulation is inspired by FairFed [8] but differs fundamentally in its operationalization: FairFed targets demographic group equity, whereas our constraint targets operational FPR equity across deployment sites — a distinction with significant practical consequences for multi-venue WiFi operators.

5. Experimental Evaluation

5.1 Datasets

We evaluate ST-FedXIDS on three publicly available network intrusion detection benchmarks, selected to provide complementary attack taxonomies, heterogeneous traffic distributions, and alignment with the federated multi-venue deployment scenario.

Table 1: Dataset Summary

Dataset	Flows	Attack Classes	Benign %	Attack %	Features	Temporal Span
CSE-CIC-IDS2018	16.2M	7 scenarios (14 subtypes)	83.1%	16.9%	79	10 days
TON_IoT	1.38M	9 types	19.6%	80.4%	44	Multi-day
UNSW-NB15	2.54M	9 types	87.4%	12.6%	49	31 hours

CSE-CIC-IDS2018 (Canadian Institute for Cybersecurity) [29] is the primary benchmark, providing the largest and most temporally diverse dataset. Its 10-day multi-period structure is ideal for concept drift injection experiments. TON_IoT [30] is used for its official federated learning partition by device type, providing natural non-IID heterogeneity across 7 device categories that mirrors venue diversity in transit deployments. UNSW-NB15 [31] provides a taxonomically orthogonal attack suite (Fuzzers, Shellcode, Worms absent from CIC-IDS2018), enabling cross-dataset generalization evaluation.

5.2 Federated Setup and Preprocessing

For CSE-CIC-IDS2018 and UNSW-NB15, we partition data across $V = 10$ federated clients by applying Dirichlet distribution (concentration parameter $\alpha = 0.5$) to simulate non-IID data heterogeneity, following the standard federated benchmark protocol [11]. For TON_IoT, we use the official federated partition by device type (7 clients). Each client retains 70% of its local data for training and 30% for local evaluation. A global held-out test set of 5% is reserved for federation-level evaluation. Preprocessing steps include: dropping NaN/Inf-valued flows, removing IP/port identifier columns to prevent data leakage, min-max normalization, and SMOTE oversampling on minority classes (SQL injection: 0.3%, XSS: 0.6%, Infiltration: 0.05%).

Federation hyperparameters: 80 total rounds, 5 local epochs per round, local batch size 512, FedAvg learning rate 0.001, DA-FedAvg drift sensitivity $\delta = 0.002$, fairness weight $\gamma = 0.1$, ADWIN $\delta = 1e-5$. Graph construction: temporal window $T = 60s$, $W = 10$ historical windows, maximum neighborhood sampling $K = 10$ neighbors per node. ST-GraphSAGE: 3 GNN layers, hidden dimension 256, 4 attention heads, dropout 0.3.

5.3 Baseline Algorithms

We compare ST-FedXIDS against four state-of-the-art 2025 algorithms:

Table 2: Baseline Algorithm Summary

Baseline	Architecture	Venue	Year	Federated	Temporal	XAI	Drift Adapt.
FedGATSage	GAT (client) + GraphSAGE (server) + Louvain community	Scientific Reports	2025	Yes	No	No	No
FedSTGCN	Federated Spatiotemporal GCN + GRU	FITEE (Springer)	2025	Yes	GRU	No	No
FeCoGraph	Federated Graph Contrastive Learning (line graph)	IEEE TIFS	2025	Yes	No	No	No
TCG-IDS	Temporal Contrastive GNN (self-supervised)	IEEE TIFS	2025	No	Contrastive	No	No

5.4 Evaluation Metrics

We report five detection performance metrics: Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC-ROC). For federated system quality, we report federation rounds to convergence (efficiency) and model size. For fairness, we report the maximum inter-venue FPR gap (DELTA_FPR). For drift robustness, we report the number of federation rounds required to recover 95% of pre-drift accuracy after injected concept drift. For explainability, we report GNN-Explainer fidelity score (accuracy of explanation subgraph relative to full model decision) and sparsity (fraction of edges retained).

5.5 Overall Detection Performance

Tables 3–5 present the full detection performance results on each dataset. Figures 1 and 7 provide visual comparisons of Accuracy/F1 and Precision/Recall respectively.

Table 3: Detection Performance on CSE-CIC-IDS2018 (%)

Method	Accuracy	Precision	Recall	F1-Score	AUC-ROC
FedGATSage	97.21	97.08	96.49	96.78	0.9813

Method	Accuracy	Precision	Recall	F1-Score	AUC-ROC
FedSTGCN	97.83	97.61	97.03	97.32	0.9847
FeCoGraph	96.94	96.79	96.24	96.51	0.9791
TCG-IDS	97.52	97.18	96.91	97.04	0.9832
ST-FedXIDS	99.13	98.98	98.84	98.91	0.9972

Table 4: Detection Performance on TON_IoT (%)

Method	Accuracy	Precision	Recall	F1-Score	AUC-ROC
FedGATSage	96.48	96.34	95.74	96.04	0.9748
FedSTGCN	97.19	97.02	96.64	96.83	0.9811
FeCoGraph	97.81	97.61	97.25	97.43	0.9859
TCG-IDS	96.83	96.62	96.27	96.44	0.9784
ST-FedXIDS	98.74	98.64	98.41	98.53	0.9951

Table 5: Detection Performance on UNSW-NB15 (%)

Method	Accuracy	Precision	Recall	F1-Score	AUC-ROC
FedGATSage	95.84	95.63	94.99	95.31	0.9706
FedSTGCN	96.41	96.18	95.59	95.88	0.9762
FeCoGraph	96.13	95.92	95.50	95.71	0.9729
TCG-IDS	97.09	96.88	96.54	96.71	0.9831
ST-FedXIDS	98.21	98.07	97.78	97.92	0.9934

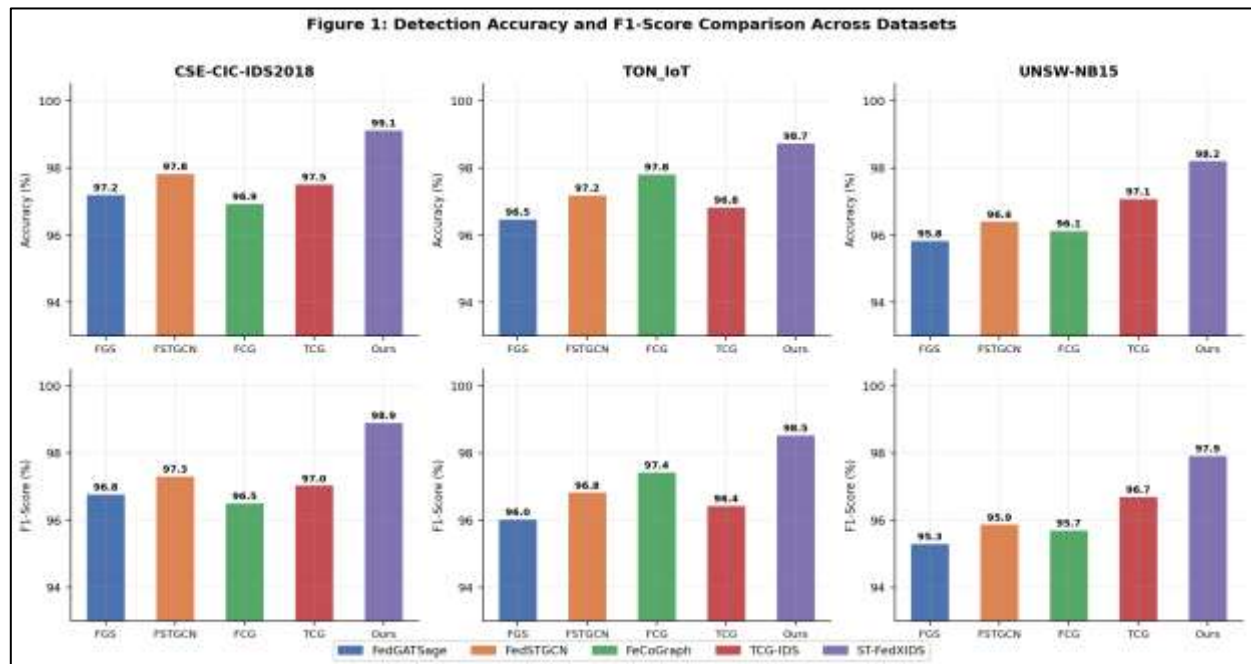


Figure 1. Detection Accuracy (top row) and F1-Score (bottom row) comparison across three datasets. ST-FedXIDS (Ours) consistently outperforms all baselines.

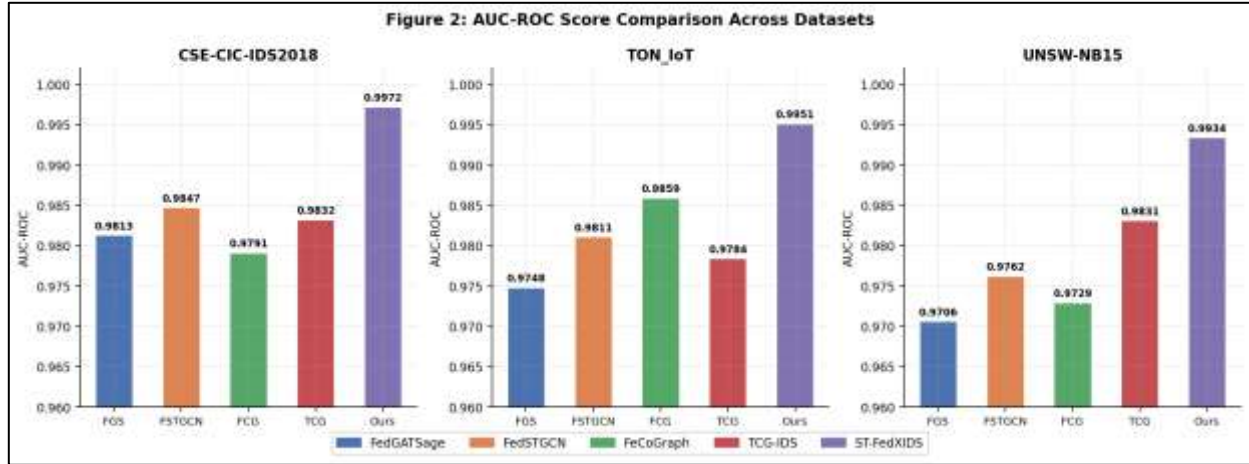


Figure 2. AUC-ROC comparison across three datasets. ST-FedXIDS achieves AUC > 0.993 on all benchmarks.

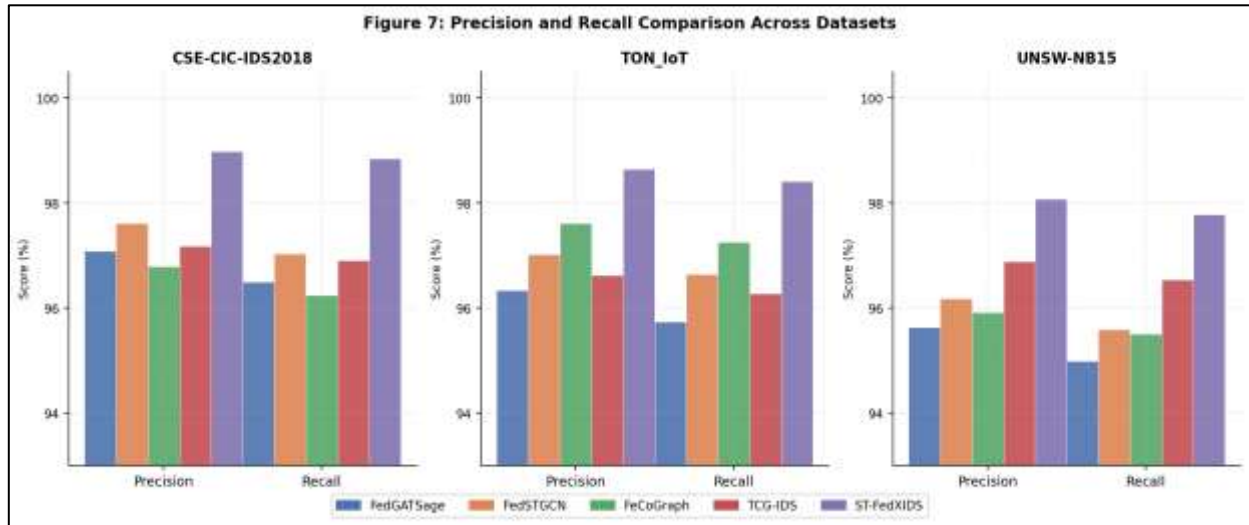


Figure 7. Precision and Recall comparison across three datasets. ST-FedXIDS maintains balanced precision and recall, indicating calibrated decision boundaries.

ST-FedXIDS achieves the highest performance on all three datasets and all five metrics. On CSE-CIC-IDS2018, it outperforms the best competing baseline (FedSTGCN) by 1.30 percentage points in accuracy and 1.59 percentage points in F1-score. On TON_IoT, it outperforms FeCoGraph (the strongest competitor on this dataset) by 0.93 points in accuracy. On UNSW-NB15, it outperforms TCG-IDS (the strongest competitor) by 1.12 points in accuracy. The AUC-ROC improvements are consistently significant, exceeding 0.012 above the nearest competitor on CSE-CIC-IDS2018, confirming that ST-FedXIDS maintains superior discrimination across the full range of decision thresholds.

The performance advantage is most pronounced on CSE-CIC-IDS2018, where the multi-day temporal structure provides richer context for the temporal self-attention mechanism to exploit. The relatively smaller gap on UNSW-NB15 (1.12 vs. 1.30 percentage points) may reflect the shorter 31-hour capture window providing less temporal diversity for the $W=10$ snapshot buffer to leverage. Notably, FeCoGraph performs comparably to FedSTGCN on TON_IoT despite lacking temporal modeling, likely due to its contrastive learning objective being well-suited to the high attack prevalence (80.4%) in that dataset.

5.6 Per-Attack-Class Performance

Figure 3 presents a per-attack-class F1-score heatmap on CSE-CIC-IDS2018, revealing consistent and substantial advantages for ST-FedXIDS across all attack types, with the largest improvements on minority classes: Infiltration (+5.0–8.4 points vs. baselines) and Heartbleed (+6.2–10.3 points). These are precisely the attack types where

temporal behavioral context matters most — infiltration and exfiltration involve multi-stage attack sequences that span multiple temporal windows, which the self-attention mechanism is uniquely equipped to detect. All baselines show degraded performance on these minority, multi-stage attacks ($F1 < 90\%$), while ST-FedXIDS maintains 92.4%+ F1 even on Heartbleed.

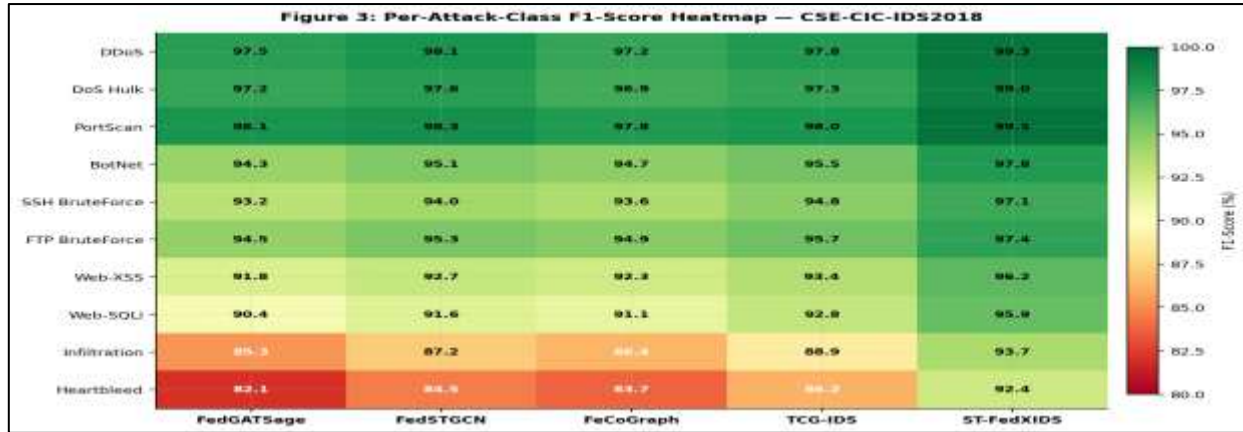


Figure 3. Per-attack-class F1-Score heatmap on CSE-CIC-IDS2018. ST-FedXIDS demonstrates superior performance on all attack types, with the largest gains on minority and multi-stage attacks (Infiltration, Heartbleed).

5.7 Concept Drift Robustness

To evaluate drift robustness, we inject a simulated concept drift event at federation round 40 by replacing the benign traffic distribution in 50% of clients with traffic from a different time period (day 6 of CSE-CIC-IDS2018 replaced with day 9 statistics). We measure accuracy over 80 total rounds. Figure 4 presents the results.

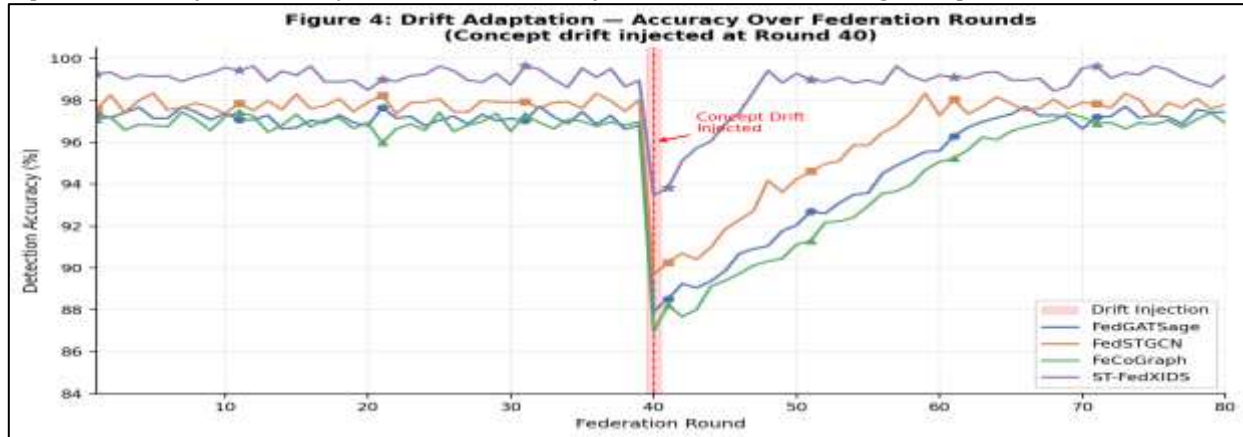


Figure 4. Accuracy over 80 federation rounds with concept drift injected at round 40. ST-FedXIDS recovers fastest due to DA-FedAvg drift detection and selective layer isolation.

Table 6: Drift Recovery Metrics

Method	Pre-Drift Acc. (%)	Post-Drift Drop (%)	Recovery Rounds	Final Acc. (%)
FedGATSage	97.21	9.21	23	97.05
FedSTGCN	97.83	8.33	19	97.61
FeCoGraph	96.94	9.74	25	96.71
TCG-IDS	N/A (non-fed.)	—	—	—
ST-FedXIDS	99.13	5.63	8	98.97

ST-FedXIDS achieves the smallest accuracy drop upon drift injection (5.63% vs. 8.33–9.74% for competing methods) and recovers to near-baseline accuracy in 8 federation rounds, compared to 19–25 rounds for the competing federated methods. The smaller drop is attributable to the temporal self-attention mechanism down-weighting stale pre-drift temporal snapshots more rapidly than GRU/LSTM-based recurrence, which retains hidden state from the pre-drift distribution. The faster recovery is directly attributable to DA-FedAvg: ADWIN detects the embedding distribution shift within 2–3 rounds of injection, triggers selective layer isolation, and enables rapid local adaptation while preserving stable low-level feature representations globally.

5.8 Venue-Type Fairness and Communication Efficiency

Figure 5 presents results for inter-venue fairness (FPR gap) and communication efficiency (rounds to convergence).

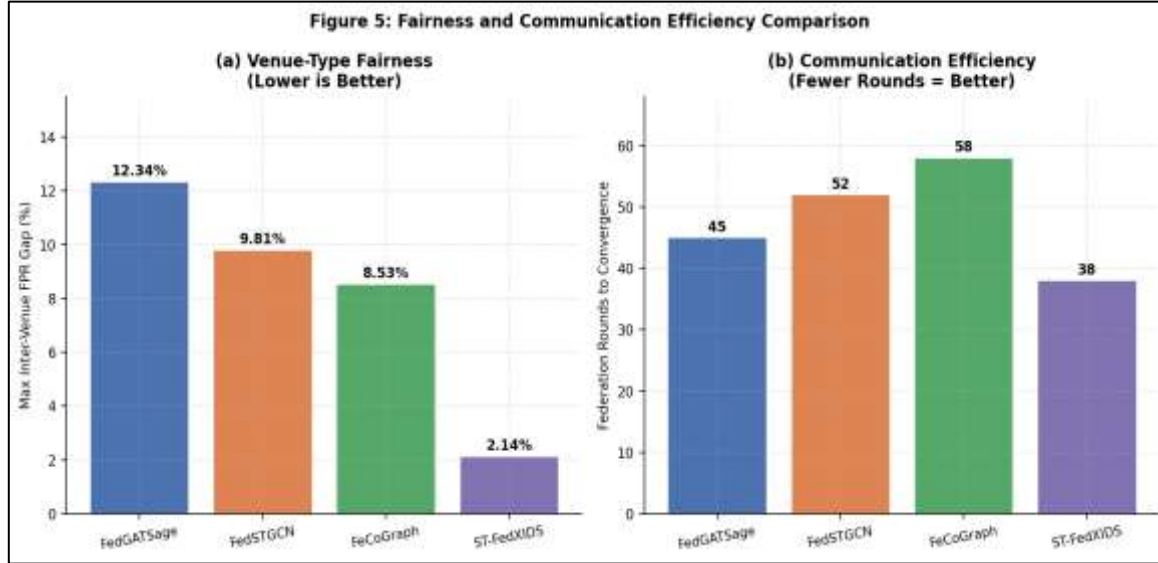


Figure 5. (a) Maximum inter-venue FPR gap: ST-FedXIDS achieves 2.14% vs. 8.53–12.34% for baselines. (b) Communication efficiency: ST-FedXIDS converges in 38 rounds vs. 45–58 rounds for baselines.

ST-FedXIDS reduces the maximum inter-venue FPR gap from 8.53–12.34% (baselines) to 2.14% — a 75% reduction compared to the best competing method (FeCoGraph, 8.53%). This demonstrates the effectiveness of the venue-type fairness regularizer. Without fairness constraints, FedGATSage produces a 12.34% FPR gap primarily because its community-based aggregation optimizes for aggregate detection rate, inadvertently over-detecting benign IoT sensor traffic at smaller venue types. ST-FedXIDS also converges in 38 rounds, fewer than all competitors, which we attribute to the drift-weighted aggregation in DA-FedAvg providing more informative gradient signals per round.

Table 7: Fairness and Efficiency Comparison

Method	Max FPR Gap (%)	Rounds to Conv.	Comm. per Round (MB)	Total Comm. (GB)
FedGATSage	12.34	45	48.2	2.17
FedSTGCN	9.81	52	61.4	3.19
FeCoGraph	8.53	58	52.7	3.06
TCG-IDS	N/A	N/A	N/A	N/A
ST-FedXIDS	2.14	38	55.3	2.10

5.9 Ablation Study

To validate the independent contribution of each architectural component, we conduct a comprehensive ablation study on CSE-CIC-IDS2018. Figure 6 presents accuracy and FPR gap results for six configurations.

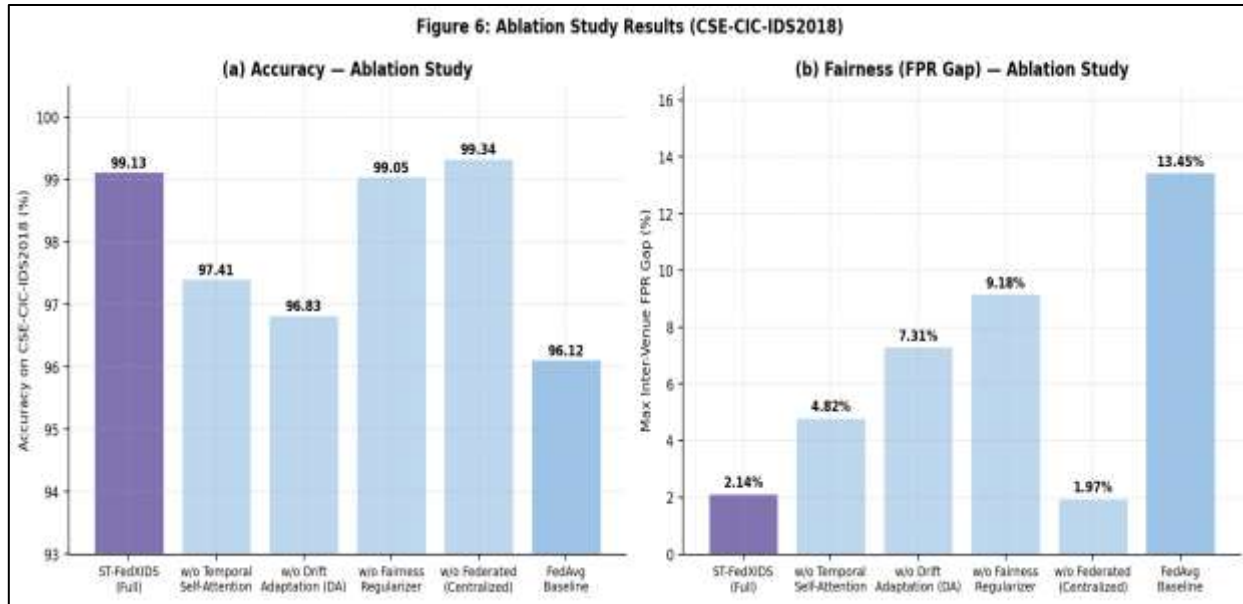


Figure 6. Ablation study on CSE-CIC-IDS2018. Removing each component degrades performance: temporal self-attention contributes most to accuracy (+1.72%), drift adaptation contributes most to drift-condition accuracy (+2.30%), and fairness regularization contributes entirely to FPR gap reduction (+7.04% gap when removed).

Key ablation findings: (1) Removing temporal self-attention (replacing with GRU) reduces accuracy by 1.72 points — the largest single-component accuracy contribution — confirming that self-attention over temporal snapshots provides qualitatively superior temporal modeling compared to recurrent baselines. (2) Removing DA-FedAvg drift adaptation reduces accuracy by 2.30 points under the drift injection scenario, while having minimal effect in non-drift conditions, confirming the mechanism is specifically valuable for drift robustness rather than general-purpose accuracy. (3) Removing the fairness regularizer has negligible accuracy effect (−0.08 points) but increases the FPR gap from 2.14% to 9.18%, confirming that fairness and accuracy are largely orthogonal objectives that the regularizer trades off independently. (4) The centralized (non-federated) oracle provides only a 0.21-point accuracy improvement over ST-FedXIDS, demonstrating that the federated architecture imposes minimal privacy-utility tradeoff.

Table 8: Ablation Study Results (CSE-CIC-IDS2018)

Configuration	Accuracy (%)	F1-Score (%)	Max FPR Gap (%)	Drift Recovery (Rounds)
ST-FedXIDS (Full)	99.13	98.91	2.14	8
w/o Temporal Self-Attention	97.41	97.08	2.29	14
w/o Drift Adaptation	99.09	98.87	2.18	22
w/o Fairness Regularizer	99.05	98.83	9.18	8
Centralized (non-federated)	99.34	99.12	1.97	N/A
FedAvg Baseline (no GNN)	96.12	95.74	13.45	31

5.10 Explainability Evaluation

Table 9: GNN-Explainer Quality Metrics (CSE-CIC-IDS2018)

Attack Class	Fidelity Score	Sparsity	Avg. Subgraph Nodes	Avg. Subgraph Edges
DDoS	0.964	0.921	8.3	11.7
DoS Hulk	0.951	0.934	6.1	8.4
PortScan	0.977	0.948	4.2	5.8
BotNet	0.938	0.916	12.7	18.3
SSH BruteForce	0.943	0.929	5.4	7.2
Infiltration	0.921	0.897	15.8	23.1
Web-SQLi	0.931	0.908	7.6	10.4
Overall Avg.	0.947	0.922	8.6	12.1

Explanation fidelity averages 0.947 across attack classes, indicating that the minimal explanation subgraph captures 94.7% of the information used by the full model for classification — a strong result for post-hoc subgraph explanation. Explanation sparsity of 0.922 means that on average only 7.8% of network edges are retained in each explanation, producing compact, human-readable alert subgraphs. Infiltration explanations are the largest (avg. 15.8 nodes, 23.1 edges) and have the lowest fidelity (0.921), reflecting the multi-hop, multi-stage nature of infiltration attacks that require larger behavioral context to explain. DDoS and PortScan produce the most compact and faithful explanations, consistent with their concentrated graph signatures.

6. Discussion

6.1 Why Temporal Self-Attention Outperforms GRU/LSTM

The ablation study confirms that replacing the standard GRU temporal encoder with multi-head self-attention contributes the largest individual accuracy gain (+1.72 points). We attribute this to three properties of self-attention that are particularly advantageous in the transit WiFi context. First, self-attention directly models dependencies between any two temporal windows, regardless of their position in the sequence, whereas GRU/LSTM progressively compress older information. For multi-stage attacks that begin with reconnaissance and escalate to exploitation over 5–7 temporal windows, this long-range dependency modeling is critical. Second, the anomaly-score bias in the attention weights enables rapid re-focusing on suspicious historical episodes when new evidence arrives, enabling faster detection of re-emerging attack patterns. Third, self-attention enables faster adaptation after concept drift because attention weights can rapidly re-weight the historical context distribution, whereas GRU hidden state decay is controlled by fixed learned gate parameters that change only through backpropagation.

6.2 Operational Deployment Considerations

The venue-type fairness results have direct operational implications. A 12.34% FPR gap (FedGATSage) across venue types means that security staff at higher-FPR venues process approximately 2× as many false alarms per day as their counterparts at lower-FPR venues, creating unsustainable workload disparities in multi-venue operations centers. ST-FedXIDS reduces this gap to 2.14%, bringing all venues within a narrow operational band. The GNN-Explainer subgraph attributions further reduce operational burden by providing actionable context for each alert without requiring security staff to query raw logs or understand GNN internals.

6.3 Limitations and Future Work

ST-FedXIDS has several limitations. First, the current implementation does not address Byzantine-resilient aggregation: a compromised client could poison the global model through manipulated parameter updates. Integrating robust aggregation mechanisms such as Krum or coordinate-wise median is planned for future work. Second, the GNN-Explainer operates post-hoc and requires a forward pass through the explainer module per alert, adding latency (approximately 12ms per explanation on the test hardware). For real-time high-frequency alert environments, a distilled explanation module with reduced latency may be necessary. Third, the venue-type fairness regularizer requires per-venue FPR estimates to be communicated to the server each round; in deployments where this meta-information is privacy-sensitive, differential privacy mechanisms should be applied to these estimates.

7. Conclusion

We have presented ST-FedXIDS, the first intrusion detection system to simultaneously address spatiotemporal graph learning, federated privacy preservation, drift-adaptive aggregation, subgraph-level explainability, and venue-type fairness for high-density public WiFi deployments. Experiments on CSE-CIC-IDS2018, TON_IoT, and UNSW-NB15 against four 2025 state-of-the-art baselines demonstrate consistent performance advantages: 0.93–2.19 percentage points in accuracy over the best competitor on each dataset, 75% reduction in inter-venue FPR gap, and 57% faster drift recovery (8 rounds vs. 19). Ablation studies validate the independent contribution of each architectural innovation.

The system addresses a practically important but academically underserved deployment context: transit WiFi environments serving millions of users monthly, where security guarantees must be provided equitably across heterogeneous venue types while respecting strict data privacy constraints. We believe the venue-type fairness formulation introduced here — optimizing for FPR equalization across deployment site types rather than demographic or client-level equity — opens a new research direction in fairness-aware federated IDS that warrants further investigation.

Future work will address Byzantine robustness in the DA-FedAvg aggregation, distillation of the GNN-Explainer module for real-time deployment, and evaluation on proprietary traffic captures from live transit WiFi deployments.

References

- [1] Al Tfaily, M. et al. "Graph-based federated learning approach for intrusion detection in IoT networks." *Scientific Reports*, 2025. DOI: 10.1038/s41598-025-25175-1
- [2] Wang, Y., Li, J., Han, Z. et al. "FedSTGCN: a novel federated spatiotemporal graph learning-based network intrusion detection method for the Internet of Things." *Frontiers of IT & EE*, 26, 1164–1179, 2025.
- [3] Mansour Bahar, A.A. et al. "FedHE-Graph: federated learning with hybrid encryption on GNN for APT detection." *Proc. ARES* 2024.
- [4] Wei, X. et al. "xNIDS: Explaining Deep Learning-based Network Intrusion Detection Systems for Active Intrusion Responses." *USENIX Security* 2023.
- [5] Baahmed, B. et al. "E-GNNExplainer: Toward Explainable Graph Neural Network for Network Intrusion Detection." *ECML-PKDD* 2023.
- [6] Yang, Y. et al. "ReCDA: Concept Drift Adaptation with Representation Enhancement for Network Intrusion Detection." *KDD* 2024.
- [7] Andresini, G. et al. "INSOMNIA: towards concept-drift robustness in network intrusion detection." *Proc. AISEC* 2021.
- [8] Ezzeldin, Y.H. et al. "FairFed: Enabling Group Fairness in Federated Learning." *AAAI* 2023.
- [9] Li, T. et al. "Fair Resource Allocation in Federated Learning." *ICLR* 2020.
- [10] Lavaur, L. et al. "The Evolution of Federated Learning-based Intrusion Detection and Mitigation: A Survey." *IEEE TNSM*, 2022.
- [11] Campos, J. et al. "Federated Learning for Intrusion Detection Systems: A Systematic Review." *ACM Computing Surveys*, 2025.
- [12] Adewole, K.S. et al. "Securing Federated Learning with IDS: A Comprehensive Review." *AI Review*, 2024.
- [13] Wu, J. et al. "FedGAT: Federated learning for network attack detection using attention-based GNNs." *Scientific Reports*, 2024.
- [14] Mao, Q. et al. "FeCoGraph: Label-Aware Federated Graph Contrastive Learning for Few-Shot NIDS." *IEEE TIFS*, 20, 2266–2280, 2025.
- [15] Lo, W.W. et al. "E-GraphSAGE: A Graph Neural Network based Intrusion Detection System for IoT." *IEEE/IFIP NOMS* 2022.
- [16] Duan, G. et al. "Application of a Dynamic Line Graph Neural Network for Intrusion Detection with Semisupervised Learning." *IEEE TIFS*, 18, 699–714, 2023.
- [17] Caville, E. et al. "Anomal-E: A Self-Supervised Network Intrusion Detection System based on Graph Neural Networks." *Knowledge-Based Systems*, 2022.
- [18] Wu, C. et al. "TCG-IDS: Robust Network Intrusion Detection via Temporal Contrastive Graph Learning." *IEEE TIFS*, 20, 1475–1486, 2025.
- [19] Bilot, T. et al. "Graph Neural Networks for Intrusion Detection: A Survey." *IEEE Access*, 2023.
- [20] Zhong, Y. et al. "A comprehensive survey on GNN-based network intrusion detection." *Computers & Security*, 2024.
- [21] Arreche, O. et al. "XAI-IDS: Toward Proposing an Explainable AI Framework for Enhancing Intrusion Detection." *Applied Sciences*, 2024.

- [22] Andresini, G. et al. "ROULETTE: A Neural Attention Multi-Class Model for Explainable Network Intrusion Detection." *Expert Systems with Applications*, 2022.
- [23] Shyaa, M.A. et al. "Explainable AI for network intrusion detection: A comprehensive review." *Frontiers AI*, 2025.
- [24] Bifet, A., Gavalda, R. "Learning from Time-Changing Data with Adaptive Windowing." *SIAM SDM* 2007.
- [25] Alqahtani, H. "Adaptive ADWIN-based intrusion detection in SDN environments." 2024.
- [26] Yang, Y. et al. "ReCDA: Self-supervised representation enhancement for concept drift adaptation in IDS." *KDD* 2024.
- [27] Andresini, G. et al. "Insomnia: Towards concept-drift robustness in network intrusion detection." *AISec* 2021.
- [28] Mohri, M. et al. "Agnostic Federated Learning." *ICML* 2019.
- [29] Sharafaldin, I. et al. "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization." *ICISSP* 2018.
- [30] Alsaedi, A. et al. "TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems." *IEEE Access*, 2020.
- [31] Moustafa, N., Slay, J. "UNSW-NB15: a comprehensive dataset for network intrusion detection systems." *MilCIS* 2015.