electronics

MDPI

Article

Scenario-Aware Federated Intrusion Detection for V2X-Inspired Edge Security: Calibration, Heterogeneity, and Communication Analysis

Manuel J. Cabral S. Reis

Engineering Department and IEETA, University of Trás-os-Montes e Alto Douro, Quinta de Prados, 5000-801 Vila Real, Portugal; mcabral@utad.pt

Abstract

Vehicle-to-Everything (V2X) communication systems are becoming a foundational component of intelligent transportation systems, but their increasing connectivity also enlarges the cyberattack surface and raises important privacy and deployment challenges for intrusion detection. Conventional centralized intrusion detection systems can achieve strong predictive performance, yet they require aggregation of sensitive traffic data and may be difficult to deploy across distributed edge environments. This study presents a federated learning-based intrusion detection evaluation framework for privacy-aware and deployment-oriented security monitoring in V2X-inspired distributed environments. Rather than proposing a new detection architecture, the work focuses on a more rigorous and realistic assessment protocol for federated intrusion detection under scenario shift, client heterogeneity, threshold-sensitive operation, and communication constraints. The proposed approach combines group-based train/test partitioning to better reflect scenario separation, lightweight multilayer perceptron (MLP) models suitable for edge-side training, and explicit analysis of communication overhead and threshold calibration. Using the CICIDS2017 dataset as a controlled proxy benchmark, the study compares centralized baselines, local-only learning, FedAvg, and FedProx for binary intrusion detection under approximately IID and strongly non-IID client partitions. The experimental protocol uses equal training-set sizes across centralized and federated methods, an independent calibration set for threshold and checkpoint selection, and five independent random seeds, with results reported as mean $\pm$ standard deviation. The results show that the stricter group-based evaluation protocol substantially reduces performance compared with optimistic random-split evaluation, confirming the importance of scenario-aware validation for intrusion detection. Under the protocol, the best mean F1-score was obtained by FedAvg in the strong non-IID configuration, with an F1-score of 0.468 ± 0.028, followed closely by FedAvg under approximately IID partitioning with 0.463 ± 0.036. Centralized MLP, logistic regression, and random forest baselines achieved comparable but slightly lower F1-scores, indicating that federated learning remained competitive rather than clearly superior under this challenging setting. The analysis further shows that threshold calibration on an independent calibration set materially changes the operating point of the detectors, while validation-selected federated checkpoints generally occurred in later communication rounds within the tested 15-round budget. Communication analysis showed that the lightweight MLP required only approximately 72 kB per model update, corresponding to about 720 kB per federated round when both uplink and downlink traffic were counted for five clients. Overall, the findings support federated learning as a viable and communication-efficient direction for privacy-aware intrusion detection in distributed

Academic Editor: Emre Tokgoz

Received: 6 April 2026
Revised: 7 July 2026
Accepted: 10 July 2026
Published: 14 July 2026

edge-security settings, while also highlighting the need for cautious interpretation, native V2X validation, and future robustness analysis against compromised federated clients.

Keywords: federated learning; intrusion detection system; V2X security; edge computing; intelligent transportation systems; CICIDS2017; non-IID learning; cybersecurity

1. Introduction

The rapid evolution of intelligent transportation systems (ITS) and connected mobility is accelerating the deployment of Vehicle-to-Everything (V2X) communications as a foundational enabler of safer, more efficient, and increasingly autonomous transportation ecosystems. By allowing vehicles to exchange information with other vehicles, roadside infrastructure, pedestrians, and cloud-edge services, V2X supports a wide range of latency-sensitive and safety-critical applications, including cooperative awareness, collision avoidance, traffic coordination, and distributed perception [1,2]. However, this growing interconnectivity also significantly expands the cyber-attack surface of transportation networks, exposing them to a diverse range of threats such as denial-of-service attacks, spoofing, message falsification, Sybil attacks, impersonation, and malware propagation across highly distributed communication and computing layers [1,3].

In contrast to traditional enterprise or static industrial networks, V2X-enabled environments are inherently dynamic, geographically distributed, and operationally heterogeneous. Vehicles, roadside units (RSUs), edge nodes, and backend services interact under stringent constraints related to latency, intermittent connectivity, mobility, and privacy preservation. These characteristics make the design of effective cybersecurity mechanisms particularly challenging. Conventional signature-based protection mechanisms remain important but are often insufficient against evolving, previously unseen, or context-dependent attacks. As a result, intelligent intrusion detection systems (IDSs), particularly those based on machine learning (ML) and deep learning (DL), have become increasingly relevant as adaptive defensive layers for next-generation transportation infrastructures [4,5].

Recent years have witnessed substantial progress in AI-driven IDS research, with numerous studies demonstrating the effectiveness of supervised and deep neural models for detecting anomalous or malicious traffic patterns in IoT, Industrial IoT, and vehicular networking contexts [4,5]. Nevertheless, most existing approaches still rely on a centralized learning paradigm, where raw traffic or telemetry collected from distributed devices is transferred to a central server for model training and optimization. Although this strategy can yield strong predictive performance under controlled experimental conditions, it introduces several critical limitations in the context of V2X and edge-assisted ITS deployments. First, the transfer and aggregation of raw traffic data raise privacy and confidentiality concerns, especially when local observations may contain sensitive behavioral, location, or infrastructure-related information. Second, centralized training can incur non-negligible communication overhead, which is problematic in bandwidth-constrained or intermittently connected environments. Third, centralized datasets often fail to reflect the strong statistical heterogeneity naturally arising across geographically dispersed clients, such as vehicles or RSUs observing distinct traffic patterns, attack prevalence, and operational contexts [6,7].

Federated Learning (FL) has emerged as a compelling alternative to centralized ML in such settings. Rather than moving raw data to a central repository, FL allows multiple distributed clients to collaboratively train a shared global model by exchanging only model parameters or gradients with an aggregation server, thereby preserving data locality and reducing direct exposure of sensitive observations [6–8]. This paradigm is particularly

attractive for distributed cyber-defense, where privacy, communication efficiency, and local autonomy are essential. In IoT and network security research, FL-based IDSs have shown promising results as privacy-preserving alternatives to centralized approaches, including in scenarios involving CICIDS2017, ToN-IoT, and other cyber-physical or device-centric benchmarks. However, despite this progress, the literature still presents important limitations when viewed from a V2X and edge-computing perspective [9–14].

First, many FL-based IDS studies still evaluate performance under conventional random train-test splits, which may inadvertently overestimate generalization by allowing closely related traffic samples or scenario-specific artifacts to appear in both training and testing subsets. While such protocols remain common in benchmark-driven intrusion detection research, they can produce overly optimistic results that do not accurately reflect operational deployment conditions, especially in dynamic distributed environments. Second, relatively few studies explicitly examine the impact of realistic client heterogeneity under strongly non-identically distributed (non-IID) traffic partitions, even though such heterogeneity is expected to be the norm in real vehicular and edge-assisted infrastructures. Third, many published works focus primarily on raw classification accuracy, while providing less emphasis on communication-aware metrics, threshold calibration, and deployment-oriented trade-offs that are directly relevant to latency-sensitive cyber-defense in transportation systems [6,9,10]. Fourth, the statistical robustness of FL-based IDS experiments is not always sufficiently characterized, despite the sensitivity of federated training to random initialization, client partitioning, local optimization, and communication-round selection.

Motivated by these gaps, this paper revisits federated intrusion detection from a deployment-oriented and evaluation-focused perspective. The objective is not to introduce a new neural detection architecture, but rather to provide a more rigorous assessment of lightweight federated IDS behavior under scenario shift, client heterogeneity, threshold-sensitive operation, and communication constraints. The proposed approach uses a lightweight multilayer perceptron (MLP) as a distributed detection model and compares several centralized, local-only, and federated learning strategies, including logistic regression, random forest, XGBoost, centralized MLP, Isolation Forest, Local-only training, FedAvg, and FedProx. This design allows the study to distinguish between the contribution of the learning paradigm and the contribution of the evaluation protocol itself.

Although the present study uses CICIDS2017 as a reproducible benchmark rather than a native V2X dataset, the experimental design is intentionally structured to emulate key properties of realistic distributed deployments, namely scenario shift and client heterogeneity. This positioning is important because CICIDS2017 does not contain native DSRC, C-V2X, CAM/DENM, or other protocol-specific vehicular messages. Therefore, it should not be interpreted as a complete V2X traffic validation corpus. Instead, it is used as a controlled and widely adopted network-intrusion benchmark that enables reproducible analysis of methodological questions that are also relevant to V2X-inspired edge-security settings, such as distributed data ownership, heterogeneous local attack prevalence, privacy-preserving model training, and communication-aware deployment. In particular, beyond a conventional random split baseline, the evaluation includes a stricter group/scenario-based split in which training, calibration, and testing are performed on disjoint traffic capture files, thereby creating a more challenging and deployment-relevant generalization setting. Furthermore, the federated analysis considers both approximately IID clients and a strong non-IID configuration with substantially different local attack prevalences. The non-IID setting is intended to approximate a situation in which different edge nodes or monitoring locations observe distinct benign-to-malicious traffic ratios, as may occur across road segments, time periods, RSUs, local gateways, or security domains.

The experimental protocol also addresses several methodological issues that are critical for unbiased evaluation. First, centralized and federated methods are evaluated using consistent training-set sizes, so that direct comparisons are not confounded by different amounts of training data. Second, an independent calibration set is introduced for threshold selection and federated checkpoint selection, avoiding the use of the final test set for model operating-point tuning. Third, all main experiments are repeated over five independent random seeds, and the results are reported using mean values, standard deviations, confidence intervals, and paired statistical comparisons where appropriate. Fourth, communication cost is reported by accounting for both uplink and downlink model-transfer traffic, rather than only client-to-server upload volume.

The experimental results show that the evaluation protocol has a decisive impact on the reported performance. Under a conventional random split, centralized baselines achieve near-perfect F1-scores, suggesting an overly optimistic setting. In contrast, under the stricter group/scenario-based split, performance drops substantially for all models, highlighting the importance of scenario-aware evaluation. Under the multi-seed protocol, federated learning remains competitive with centralized and local-only baselines rather than clearly dominating them. In particular, FedAvg under strong non-IID partitioning achieved the highest mean F1-score among the evaluated methods, with an F1-score of 0.468 ± 0.028, followed closely by FedAvg under approximately IID partitioning with 0.463 ± 0.036. Local-only and centralized neural baselines achieved comparable but slightly lower performance, while Isolation Forest was less effective in this setting. These results indicate that lightweight federated learning can remain practically competitive under scenario-aware evaluation, but they also support a cautious interpretation of performance differences among the strongest supervised methods.

The communication analysis further shows that the lightweight MLP model remains suitable for communication-aware edge-oriented experimentation, with a serialized model size of approximately 72 kB. With five clients, this corresponds to approximately 720 kB of total model-transfer traffic per communication round when both uplink and downlink are counted, and approximately 11.06 MB over 15 communication rounds. Validation-selected federated checkpoints generally occurred in later rounds within the tested budget, indicating that additional communication rounds remained beneficial under the calibration protocol, although communication cost increased linearly with the number of rounds.

Taken together, these findings support the feasibility of federated learning as a privacy-aware and communication-efficient direction for distributed intrusion detection in edge-security settings inspired by V2X deployments. At the same time, the results should not be interpreted as demonstrating a production-ready V2X IDS. Native vehicular traffic datasets, protocol-aware feature engineering, compromised-client threat models, adversarial federated robustness, and real RSU or vehicular-edge deployments remain necessary future steps.

The main contributions of this work are summarized as follows:

1. We present a federated and edge-oriented intrusion detection evaluation framework for V2X-inspired distributed security settings, designed to preserve data locality while enabling collaborative model training across distributed clients. The framework is intentionally positioned as a deployment-oriented methodological study rather than as a protocol-specific production V2X IDS.
2. We introduce a more realistic evaluation protocol based on group/scenario-separated train-test splits, highlighting the limitations of conventional random partitioning for intrusion detection benchmarking in distributed environments. The protocol further includes an independent calibration set for threshold and checkpoint selection, thereby avoiding operating-point tuning on the final test set.

3. We compare centralized, local-only, and federated learning methods under a common binary detection setting, including logistic regression, random forest, XGBoost, centralized MLP, Isolation Forest, Local-only training, FedAvg, and FedProx.
4. We analyze both approximately IID and strongly heterogeneous non-IID client settings, explicitly examining how client-level attack-prevalence heterogeneity affects federated intrusion detection performance under a scenario-aware group split.
5. We strengthen statistical robustness by repeating the main experiments over five independent seeds and reporting mean performance, standard deviation, confidence intervals, and paired statistical comparisons where appropriate.
6. We show that threshold calibration is a practically relevant component of intrusion detection under scenario shift, and that validation-based checkpoint selection materially affects the reported federated operating point.
7. We complement classification metrics with communication-aware indicators, including model size, uplink traffic, downlink traffic, cumulative communication volume, communication rounds, and inference time, to support discussion of deployment feasibility in edge-assisted transportation and distributed security systems.

The remainder of this paper is organized as follows. Section 2 reviews the state of the art on AI-based intrusion detection, federated learning for distributed cyber-defense, and edge-assisted security architectures. Section 3 presents the proposed federated edge IDS framework. Section 4 formulates the learning and evaluation problem. Section 5 describes the experimental methodology, including dataset preparation, centralized baselines, local-only baselines, federated configurations, calibration procedures, and communication-aware metrics. Section 6 reports the experimental results. Section 7 discusses their methodological and deployment implications, limitations, and future research directions. Finally, Section 8 summarizes the main conclusions.

2. Related Work

This section reviews the state of the art relevant to the proposed work from five complementary perspectives: (i) cybersecurity challenges in V2X and Internet of Vehicles (IoV) environments, (ii) AI-driven intrusion detection systems (IDSs), (iii) federated learning for intrusion detection, (iv) edge-assisted security architectures, and (v) the specific research gap addressed in this paper. The section also clarifies the methodological positioning of the present study, particularly the distinction between protocol-specific V2X intrusion detection and V2X-inspired evaluation of distributed edge-security learning.

2.1. Cybersecurity in V2X and Internet of Vehicles Environments

Vehicle-to-Everything (V2X) communication is widely recognized as a key technological enabler of intelligent transportation systems, cooperative driving, and future autonomous mobility. However, its open, wireless, and highly distributed nature also makes it particularly vulnerable to cyber threats. Compared with conventional enterprise networks, V2X and Internet of Vehicles (IoV) environments are characterized by mobility, intermittent connectivity, dynamic topology changes, heterogeneous communication stacks, and strict latency constraints, all of which complicate the design of robust security mechanisms [1,3].

The literature identifies a broad range of attacks relevant to V2X systems, including denial-of-service and distributed denial-of-service attacks, Sybil attacks, impersonation, spoofing, false data injection, message tampering, replay attacks, and privacy-related tracking threats [1,3]. These attacks may affect not only communication availability and confidentiality, but also the integrity and timeliness of safety-critical information exchanged among vehicles, roadside units (RSUs), and backend services. False data injection and dummy data injection are particularly relevant in safety-critical cyber-physical environ-

ments because manipulated or fabricated information may affect downstream perception, coordination, or control decisions. Recent work on false or dummy data injection in related cyber-physical infrastructures further illustrates the need for detection approaches that can operate under distributed, heterogeneous, and time-sensitive conditions [15]. As a result, V2X security has increasingly moved beyond pure cryptographic protection and message authentication, motivating the integration of intelligent detection and monitoring mechanisms capable of identifying suspicious traffic or behavioral anomalies in near real time.

Several recent works have highlighted the importance of security architectures that combine vehicular communication, roadside processing, and edge/cloud coordination. This architectural trend is particularly relevant because many emerging V2X applications, such as cooperative perception and edge-assisted traffic intelligence, inherently distribute sensing and decision support across multiple local nodes [1,2]. Consequently, cybersecurity mechanisms for such environments must be not only accurate, but also scalable, privacy-aware, and resilient to statistical heterogeneity across geographically dispersed observation points. At the same time, protocol-specific V2X security evaluation requires datasets and features that capture vehicular communication semantics, such as DSRC, C-V2X, CAM/DENM messages, cooperative perception messages, or application-layer vehicular events. General network intrusion datasets do not directly provide this semantic layer. This distinction is important for interpreting studies, including the present one, that use general-purpose network intrusion benchmarks as controlled proxies for distributed edge-security evaluation rather than as complete V2X traffic corpora.

2.2. AI-Driven Intrusion Detection Systems

Intrusion detection systems have evolved from purely signature-based solutions toward hybrid and data-driven paradigms capable of identifying previously unseen attacks and context-dependent malicious behaviors. In recent years, machine learning (ML) and deep learning (DL) techniques have become central to this evolution, particularly in environments such as IoT, IIoT, cyber-physical systems, and vehicular networks, where attack patterns can be diverse and continuously changing [4,5].

Classical ML-based IDSs commonly rely on algorithms such as logistic regression, decision trees, random forests, support vector machines, and gradient boosting methods. These models often offer strong performance on tabular traffic datasets and remain competitive due to their interpretability, robustness, and relatively low deployment cost. Unsupervised or semi-supervised anomaly detection methods, such as Isolation Forest, are also relevant in intrusion detection because they can be trained primarily on benign traffic and used to flag statistically atypical observations without requiring exhaustive attack labels. In contrast, DL-based approaches increasingly employ multilayer perceptrons, convolutional neural networks, recurrent architectures, autoencoders, and, more recently, attention-based models to learn richer traffic representations and improve generalization under complex attack distributions [4].

The choice between classical ML, anomaly detection, and deep neural architectures is closely linked to the available feature representation and deployment constraints. For structured flow-level datasets such as CICIDS2017, tree-based methods, logistic models, gradient boosting, and compact MLPs remain strong and practical baselines. More complex architectures, including CNNs, LSTMs, Transformers, and graph neural networks, may be advantageous when spatial, temporal, sequence-based, or relational structure is explicitly available, but they also introduce additional computational and communication costs. For this reason, lightweight neural models remain relevant for edge-oriented and federated ex-

perimentation, particularly when the objective is to evaluate distributed learning behavior rather than to maximize architectural complexity.

Despite the significant reported progress in benchmark-driven IDS research, several methodological concerns remain. First, many studies report very high or near-perfect accuracy on popular datasets such as KDDCup99, NSL-KDD, CICIDS2017, and related benchmarks, but these results are often obtained under conventional random train-test splits that may not adequately capture scenario shift or realistic operational variation. Second, many papers emphasize overall accuracy while under-reporting class imbalance effects, false positives, and the precision–recall trade-off, which are particularly relevant in intrusion detection contexts. Third, comparatively fewer works explicitly discuss how decision thresholds, data partitioning choices, or deployment constraints influence the practical utility of the learned models. These limitations are especially important in dynamic and safety-critical environments such as V2X, where an apparently high benchmark score may not translate into robust field performance.

2.3. Federated Learning for Intrusion Detection

Federated Learning (FL) has emerged as a promising paradigm for distributed intrusion detection because it addresses one of the central tensions in cybersecurity analytics: the need to learn from diverse, distributed data sources while avoiding raw data centralization. The seminal FedAvg framework introduced in [6] established the practical feasibility of collaborative model training under decentralized data storage. Since then, FL has been increasingly studied in cybersecurity, especially in intrusion detection scenarios involving IoT, IIoT, and other distributed edge environments [7,8].

A growing body of work has specifically examined FL-based IDSs. Lazzarini et al. [9] provide a useful recent overview of FL for IoT intrusion detection and emphasize the privacy and scalability advantages of distributed training. Hamdi et al. [14] propose a federated learning-based IDS for IoT and reports encouraging results compared with centralized alternatives. Rashid et al. [11] investigate a federated approach for intrusion detection in industrial IoT networks, showing that collaborative distributed training can outperform isolated local models. Similarly, Amiri-Zarandi et al. [12] propose a trust-aware federated IDS architecture for IoT environments and highlight the importance of privacy-preserving collaboration under heterogeneous conditions. More broadly, Agrawal et al. [13] review the conceptual foundations, implementation challenges, and future directions of FL for IDS, identifying issues such as non-IID data, communication cost, poisoning risk, and model aggregation stability as central research challenges.

Beyond FedAvg, several federated optimization and aggregation strategies have been proposed to improve training stability under heterogeneous data. FedProx extends FedAvg by adding a proximal regularization term that discourages excessive deviation of local models from the current global model, making it particularly relevant in non-IID settings [16]. Other strategies, such as SCAFFOLD and FedNova, address client drift, variance reduction, and normalization of local update effects under heterogeneous client participation and local training schedules [17,18]. Although the present study does not aim to exhaustively benchmark all FL optimizers, it includes FedProx as a representative non-IID-oriented baseline in addition to FedAvg and local-only training. This choice strengthens the comparison while preserving the light-weight and reproducible character of the experimental framework.

Although these studies demonstrate the promise of FL for IDS, the literature still presents several limitations relevant to the present work. First, many FL-based IDS papers continue to rely on conventional random train-test partitions of benchmark datasets, which may overestimate generalization in distributed deployment scenarios. Second, while non-

IID data are widely acknowledged as a defining characteristic of federated settings, the actual degree of client heterogeneity is often only weakly explored or not systematically stressed. Third, communication-aware evaluation remains uneven: some works focus primarily on classification performance without explicitly quantifying communication load, update volume, or simple latency proxies. Fourth, many FL-based IDS evaluations report single-run results, even though federated performance may be sensitive to random initialization, client partitioning, local optimization order, and communication-round selection. Fifth, threshold calibration and validation-based checkpoint selection are not always treated as part of the experimental protocol, even though they can materially change the operating point of an intrusion detector. These gaps motivate a more deployment-oriented experimental design that combines realistic partitioning, explicit client heterogeneity, and communication-aware reporting.

It is also important to distinguish traffic-level intrusion detection from the security of the federated learning protocol itself. In practical V2X or edge deployments, clients such as RSUs, gateways, or local monitors may become compromised and submit malicious model updates. This raises important challenges related to poisoning, Byzantine behavior, inference leakage, and secure aggregation. These issues are highly relevant to production deployments, but they define a different threat model from the one evaluated in this paper. The present work focuses on traffic-level intrusion detection under benign FL participants, while adversarially robust FL aggregation and compromised-client defense are identified as important future extensions.

2.4. Edge-Assisted Security Architectures for Distributed Transportation and IoT Systems

Edge computing has become increasingly relevant in security-sensitive distributed systems because it allows data processing, decision support, and anomaly detection to be moved closer to the data source. In transportation and V2X contexts, edge nodes may be implemented at RSUs, roadside micro-data centers, vehicular gateways, or other local infrastructure capable of supporting low-latency analytics. This architectural model is particularly attractive for intrusion detection, as it reduces dependence on remote cloud processing and supports faster local reaction to suspicious traffic patterns.

From a systems perspective, edge-assisted IDS deployment offers several practical advantages. It can reduce communication overhead, improve responsiveness, support local autonomy during intermittent connectivity, and better align with privacy-preserving or regulatory constraints. At the same time, however, it introduces challenges such as limited computational resources, heterogeneous local data distributions, synchronization constraints, and the need for efficient coordination across distributed nodes. These characteristics make FL especially relevant, as it naturally complements edge-assisted architectures by enabling collaborative learning without requiring full traffic centralization [6–8].

In recent years, the combination of FL and edge-oriented security has gained increasing attention in IoT and cyber-physical system contexts. Existing studies suggest that this combination can provide a reasonable balance between privacy, scalability, and detection effectiveness, particularly when centralized data sharing is undesirable or infeasible [10–12]. Related fog- and edge-based AI cybersecurity architectures also emphasize intermediate processing layers between end devices and the cloud, where local or regional nodes perform detection, aggregation, or decision support closer to the data source. Such architectures are conceptually aligned with the motivation of federated edge IDS, although many existing works do not explicitly evaluate federated training under scenario-separated test conditions, independent threshold calibration, and statistically repeated non-IID client partitions. However, relatively fewer works explicitly frame their contributions in the context of V2X-inspired distributed deployments while simultaneously addressing realistic

partitioning effects and threshold-sensitive detection behavior. This gap is important because in transportation systems, the practical operating point of an IDS—rather than only its default-threshold score—may strongly influence false alarms, missed detections, and overall operational acceptability.

Communication-aware analysis is also central to edge-assisted deployment. In federated IDS, communication cost includes not only client-to-server upload traffic, but also server-to-client distribution of the global model. Reporting only upload volume can underestimate the total model-transfer cost. Therefore, deployment-oriented evaluation should account for both uplink and downlink communication, as well as model size, number of clients, number of communication rounds, and inference-time proxies.

2.5. Research Gap and Positioning of the Present Work

Based on the reviewed literature, five main gaps can be identified.

First, there is a methodological gap in the evaluation of AI-based IDSs for distributed cyber-defense. Many studies, including federated ones, still rely on conventional random train-test splits that may yield overly optimistic performance estimates. Such protocols are often insufficient for capturing the scenario shift expected in real-world distributed environments, including V2X-inspired deployments where edge nodes may observe substantially different traffic contexts over time and location.

Second, although non-IID data are widely recognized as a core challenge in FL, many published IDS studies either treat heterogeneity only superficially or do not systematically compare approximately IID and strongly heterogeneous client settings under a common experimental protocol. This limits the practical interpretability of the reported results.

Third, deployment-oriented aspects such as decision-threshold calibration, communication-aware indicators, and robustness under realistic data partitioning remain underexplored in comparison with pure classification benchmarking. Yet these aspects are directly relevant for latency-sensitive and false-alarm-sensitive intrusion detection in transportation and edge-assisted environments.

Fourth, statistical robustness is often insufficiently documented. Since federated IDS performance may vary across seeds, client assignments, and communication rounds, single-run results provide limited evidence for stable conclusions. Multi-seed evaluation, variability reporting, and paired statistical comparisons are therefore important for a more reliable assessment.

Fifth, the relationship between general-purpose network intrusion datasets and V2X-specific deployment claims is not always clearly stated. Native V2X datasets would provide stronger domain specificity, especially for protocol-aware attacks such as Sybil behavior, false vehicular messages, replay, and cooperative-perception manipulation. However, controlled benchmarks such as CICIDS2017 remain useful for reproducible methodological evaluation when their limitations are explicitly acknowledged.

The present work is positioned at the intersection of these gaps. Rather than aiming solely for maximal benchmark accuracy, this paper focuses on a more deployment-aware validation of an AI-driven federated intrusion detection framework. More precisely, it presents a V2X-inspired, edge-oriented evaluation framework for lightweight federated intrusion detection, rather than a protocol-specific production V2X IDS. Specifically, it combines: (i) a scenario-aware group-based evaluation protocol, (ii) centralized, local-only, and federated comparisons under a common binary detection setting, (iii) explicit IID and strong non-IID client configurations, (iv) independent threshold calibration and validation-based federated checkpoint selection, (v) multi-seed performance reporting with variability analysis, and (vi) communication-aware measurements that account for both uplink and downlink model-transfer cost. Although the experiments use CICIDS2017 as a

reproducible benchmark rather than a native V2X dataset, the proposed evaluation methodology is designed to emulate key properties of realistic edge-assisted distributed security environments and therefore to provide a more informative assessment of practical federated IDS feasibility. The resulting conclusions are intentionally cautious: the study evaluates whether lightweight FL can remain competitive under stricter deployment-inspired evaluation conditions, while leaving native V2X validation, adversarial FL robustness, and real edge/fog deployment experiments for future work.

3. System Model and Threat Model

This section formalizes the distributed V2X-inspired security scenario considered in this work. It first introduces the system model underlying the proposed edge-assisted federated intrusion detection framework, and then defines the threat model that motivates the intrusion detection task. Because the experimental validation uses a general-purpose network intrusion benchmark rather than native vehicular traffic, the model is intentionally defined at the architectural level of distributed edge-security monitoring, rather than at the level of a specific V2X protocol stack.

3.1. System Model

We consider a distributed cyber-defense architecture inspired by Vehicle-to-Everything (V2X) and edge-assisted intelligent transportation system (ITS) deployments. The environment consists of a set of geographically distributed edge nodes, denoted by $\mathscr{E} = \{\mathrm{E}_1, \mathrm{E}_2, \ldots, \mathrm{E}_\mathrm{K}\}$, where each node observes local communication or network traffic generated within its own operational context. In a practical V2X setting, such edge nodes may correspond to roadside units (RSUs), vehicular gateways, roadside micro-servers, or local security monitors associated with transportation infrastructure and connected vehicles. Each node collects local traffic records, extracts features, and performs local intrusion detection inference and/or local model training.

A central coordination server, denoted by S, is responsible for orchestrating collaborative model updates across the edge nodes. Importantly, the server does not require access to raw traffic data. Instead, each edge node trains a local model on its own private dataset and periodically shares only model parameters or parameter updates with the server. The server then aggregates the received updates to produce a new global model, which is redistributed to the participating nodes in subsequent communication rounds. This learning paradigm is consistent with the federated learning model introduced by McMahan et al. [6] and further discussed in broader FL surveys [7,8].

Let $\mathscr{D}_\mathscr{k} = \left\{(x_i, y_i)\right\}_{i=1}^{n_k}$ denote the local dataset stored at edge node E_k, where $x_i \in R^d$ is a feature vector extracted from a traffic observation and y_i is the corresponding label. Depending on the task, y_i may represent either a binary class (e.g., benign vs. attack) or a multi-class attack category. In this paper, the primary comparative analysis focuses on the binary intrusion detection setting, which is particularly relevant for practical anomaly screening and operational deployment. The local datasets are not assumed to be identically distributed across clients. On the contrary, statistical heterogeneity is explicitly considered a core characteristic of the target environment, reflecting the fact that different edge nodes may observe substantially different traffic mixes, attack prevalence levels, or temporal patterns.

The global learning objective can therefore be expressed as the minimization of a weighted empirical risk over distributed client datasets:

$$\min_{\theta} F(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(\theta), \quad F_k(\theta) = \frac{1}{n_k} \sum_{i=1}^{n_k} l(f_\theta(x_i), y_i),$$

where $n = \sum_{k=1}^{K} n_k$, f_θ is the intrusion detection model parameterized by θ, and $l(\cdot)$ is the classification loss. This formulation captures both centralized and federated learning under a common objective: centralized learning approximates the direct optimization of $F(\theta)$ over pooled data, whereas federated learning estimates the same objective through local client updates and server-side aggregation.

The global learning objective is therefore to obtain a collaborative intrusion detection model f_θ, parameterized by θ, that can generalize across heterogeneous local conditions while preserving data locality. Operationally, the framework supports two complementary functions: (i) local real-time or near-real-time inference at the edge, where the current model is used to classify incoming traffic records, and (ii) periodic federated retraining or model refinement, where local nodes contribute to improving the shared model based on newly observed data.

This system abstraction is intentionally deployment-oriented rather than protocol-specific. In other words, the proposed framework does not depend on a single V2X communication stack or a single ITS standard. Instead, it captures the architectural properties that are most relevant from a cybersecurity perspective: decentralized data generation, local edge observability, privacy-sensitive traffic logs, heterogeneous operating conditions, and the need for collaborative learning under limited data sharing. This abstraction is also compatible with broader IoT and cyber-physical system security contexts, which is why a reproducible benchmark such as CICIDS2017 can be used as an experimental proxy while still supporting meaningful conclusions about edge-assisted distributed IDS design [9,13]. However, because CICIDS2017 does not contain native DSRC, C-V2X, CAM/DENM, or other vehicular protocol messages, the conclusions are restricted to V2X-inspired distributed edge-security learning and should not be interpreted as direct validation of a protocol-specific V2X IDS.

3.2. Communication and Learning Assumptions

The proposed framework assumes a synchronous round-based federated learning process. During each communication round, the coordination server distributes the current global model parameters to a subset or all participating edge nodes. Each selected node performs local training for a fixed number of epochs using its own private dataset and then transmits the updated model parameters back to the server. The server applies a weighted aggregation rule, typically based on local sample counts, to compute the next global model. This process continues for a predefined number of communication rounds or until convergence criteria are satisfied.

This round-based abstraction is adopted because it offers a clear and reproducible baseline for comparing centralized and federated training under controlled conditions. It also aligns with the standard FedAvg setting widely used in the literature [6]. In the experimental study, FedAvg is complemented with FedProx, which introduces a proximal regularization term during local optimization to reduce excessive client drift under heterogeneous data. These methods are used as representative lightweight FL baselines rather than as new algorithmic contributions. In practical deployments, asynchronous variants or partially participating clients may also be relevant, especially in vehicular and edge environments subject to intermittent connectivity or variable availability. However, these extensions are outside the primary scope of the present study and are therefore left for future work.

Two data heterogeneity regimes are explicitly considered in this paper. The first approximates an IID-like partition, in which the local class distributions across edge nodes are relatively balanced and similar. The second represents a strongly non-IID partition, in which client datasets exhibit markedly different class proportions and therefore emulate

heterogeneous local operating conditions. This distinction is particularly important in V2X-inspired settings, where different locations or operational periods may produce very different traffic profiles and attack frequencies. For example, some monitoring points may observe mostly benign mobility-related traffic, whereas others may be located near attack-prone gateways, congested infrastructure segments, or time windows with a higher prevalence of anomalous flows. The non-IID partition used in the experiments should therefore be interpreted as an operational abstraction of heterogeneous attack exposure rather than as a measured distribution from a specific vehicular deployment. By including both regimes, the proposed evaluation directly addresses one of the central challenges of practical federated intrusion detection [7,13].

In addition to classification performance, the system model also motivates deployment-aware evaluation dimensions. Since the model is intended for edge-assisted operation, relevant indicators include not only accuracy-oriented metrics but also communication cost, inference time, and threshold-sensitive detection behavior. These aspects are especially relevant in safety-critical or latency-sensitive contexts, where a model with slightly lower nominal accuracy may still be preferable if it yields lower communication overhead, lower inference latency, or a more favorable false alarm profile at the selected operating threshold. Accordingly, the evaluation reports model size, client-to-server uplink traffic, server-to-client downlink traffic, total communication volume, communication-round budget, and inference-time proxies. This avoids underestimating communication cost by considering only upload traffic.

The learning protocol further assumes that threshold selection and federated checkpoint selection are performed using a calibration set that is separate from the final test set. This assumption is important because the binary decision threshold directly controls the precision–recall trade-off of the IDS. Selecting the threshold on the test set would produce an optimistically biased operating point. In the methodology, the final test set is therefore reserved for unbiased evaluation after model, round, and threshold selection have been completed.

3.3. Threat Model

The threat model considered in this work is aligned with the broader cybersecurity challenges reported for V2X, IoV, and distributed intelligent transportation systems [1,3]. We assume that the monitored communication environment may be exposed to a mixture of benign and malicious traffic patterns, where malicious traffic corresponds to cyberattacks capable of degrading service availability, compromising communication integrity, or disrupting normal network operation.

At the network and transport levels, relevant attack categories include denial-of-service (DoS) and distributed denial-of-service (DDoS) traffic, brute-force attempts, scanning and probing activity, web-based exploitation attempts, botnet-related behavior, and other traffic patterns associated with unauthorized access or resource exhaustion. These attack families are consistent with the broad types represented in the CICIDS2017 benchmark and are also representative of the types of anomalous or malicious flows that may be relevant in distributed cyber-physical communication infrastructures. They do not, however, exhaust the space of V2X-specific attacks, such as Sybil behavior, false cooperative-awareness messages, manipulated vehicular state information, replay of vehicular messages, or attacks targeting cooperative perception. Those protocol- and semantics-aware threats require native vehicular datasets and feature representations, and are therefore outside the empirical scope of the present benchmark-based study.

From a systems perspective, the adversary is assumed to be external to the collaborative training infrastructure. That is, the primary objective of this paper is to detect

malicious network behavior in the monitored environment, rather than to defend the federated learning protocol itself against poisoning, Byzantine, or model inversion attacks. Accordingly, the federated aggregation server and participating edge nodes are assumed to follow the protocol correctly during training. This assumption allows the study to isolate and evaluate the core question addressed in the paper: whether federated edge-assisted learning can provide effective and practically meaningful intrusion detection under realistic data heterogeneity and deployment-oriented constraints.

This is an important scope clarification. The security of federated learning itself is a rich and active research area involving adversarial model poisoning, data poisoning, inference leakage, and secure aggregation. Although these issues are highly relevant to real deployments, incorporating them would substantially broaden the problem formulation and risk diluting the central contribution of the present study. Therefore, the proposed framework should be interpreted as a first-layer distributed IDS architecture that addresses traffic-level intrusion detection under privacy-preserving collaborative learning, rather than as a complete end-to-end secure federated defense stack. In practical V2X deployments, compromised RSUs, gateways, or vehicular edge devices could submit malicious model updates; defending against such behavior would require robust aggregation, anomaly detection over client updates, secure aggregation, or trusted execution mechanisms. These extensions are explicitly identified as future work.

3.4. Detection Tasks Considered in This Work

To support a practical and reproducible evaluation, two intrusion detection tasks are considered.

The first task is binary intrusion detection, where the model classifies each traffic observation as either benign or malicious. This task is the primary focus of the paper because it aligns naturally with operational edge screening, where the first requirement is often to distinguish normal from suspicious traffic as reliably as possible. Binary detection is also well suited for analyzing threshold sensitivity, communication–performance trade-offs, and the impact of client heterogeneity under federated learning. The experimental comparison, including centralized baselines, local-only training, FedAvg, and FedProx, is therefore anchored in this binary task.

The second task is coarse multi-class intrusion categorization, where traffic is mapped into a reduced set of broader attack families rather than a large number of fine-grained labels. This task is included as a complementary perspective to assess whether the proposed framework can support richer attack semantics beyond anomaly screening. However, given the stronger sensitivity of multi-class learning to class coverage, partitioning effects, and unseen-category issues, the multi-class analysis is treated as secondary and exploratory in this study. In particular, the claims do not rely on multi-class performance, because scenario-aware group splitting may create label-coverage mismatches that make multi-class results difficult to interpret without additional task-specific controls.

This distinction is particularly relevant for the present experimental setup. Since scenario-aware or group-based splits may produce label distribution mismatch between training and testing subsets, binary detection provides a more robust baseline for fair centralized-versus-federated comparison. Multi-class categorization remains valuable, but its interpretation requires additional caution whenever some attack categories are underrepresented or absent in specific partitions. Future work should revisit this task using native V2X datasets or controlled vehicular attack corpora in which attack categories are consistently represented across training, calibration, and testing scenarios.

3.5. Scope and Practical Interpretation

The system and threat models defined above intentionally strike a balance between realism and experimental tractability. The framework is motivated by V2X and edge-assisted ITS deployments, but the experimental evaluation is carried out on a reproducible benchmark dataset rather than a native vehicular packet capture corpus. This design choice is deliberate. At the present stage, the main scientific objective is not to claim a production-ready V2X detector tied to a specific protocol stack, but rather to demonstrate a robust and reproducible methodology for evaluating federated intrusion detection under realistic distributed learning constraints.

Accordingly, the present work should be interpreted as a deployment-oriented methodological study with V2X-inspired relevance. The proposed architecture captures the essential properties of distributed edge-based cyber-defense—data locality, privacy constraints, client heterogeneity, collaborative model aggregation, and threshold-sensitive operation—and evaluates them under a controlled benchmark protocol. This positioning is important because it provides a sound foundation for future extensions involving native vehicular datasets, protocol-aware feature engineering, adversarial federated robustness, and real-time embedded deployment on actual roadside or vehicular edge hardware.

The practical interpretation of the results should therefore be cautious. Competitive performance under CICIDS2017 group-based evaluation indicates that lightweight FL can remain viable under a difficult scenario-shift benchmark, but it does not prove operational readiness for real V2X deployments. A production-oriented system would require validation with vehicular communication traces, protocol-aware features, realistic mobility and connectivity patterns, compromised-client defenses, and deployment on actual or emulated RSU/edge hardware.

4. Proposed Federated Edge IDS Framework

This section presents the proposed federated edge-oriented intrusion detection evaluation framework for distributed V2X-inspired communication security. The framework combines edge-assisted traffic monitoring with federated learning (FL) in order to support privacy-preserving collaborative model training across geographically distributed nodes. While the experimental validation is conducted using a reproducible benchmark dataset, the framework is designed from a deployment-oriented perspective and targets the architectural characteristics expected in distributed intelligent transportation and edge-assisted cyber-physical communication environments. Consistent with the positioning of this work, the framework is intended as a methodological and evaluation-oriented architecture rather than as a protocol-specific production V2X IDS.

4.1. Framework Overview

The proposed framework consists of three main layers: (i) distributed traffic observation and feature extraction at the edge, (ii) local model training and inference at edge nodes, and (iii) centralized coordination for global model aggregation. This organization is conceptually aligned with emerging edge-assisted security architectures in intelligent transportation systems, where traffic monitoring may occur at roadside units (RSUs), vehicular gateways, roadside micro-servers, or other local edge devices capable of observing network or communication flows [1,3].

At the edge layer, each node continuously receives or stores locally observed traffic records and transforms them into structured feature vectors suitable for machine learning. These feature vectors can be derived from packet captures, flow-level aggregations, or other network telemetry sources depending on the deployment context. In the present study, the feature space is instantiated through the preprocessed CICIDS2017 representation,

which provides a reproducible tabular approximation of traffic observations. Because CICIDS2017 is not a native V2X dataset, this feature space is used as a controlled proxy for distributed network-intrusion evaluation rather than as a direct representation of DSRC, C-V2X, CAM/DENM, or other vehicular protocol traffic.

Each edge node then performs two possible operations. First, it applies the current local copy of the intrusion detection model to classify incoming traffic in near real time. Second, during federated retraining cycles, it uses its local dataset to refine the shared model without transmitting raw data outside the local environment. Instead, only model parameters or parameter updates are exchanged with a central coordination server. The server aggregates the received local updates using federated optimization strategies such as FedAvg and FedProx and redistributes the resulting global model in subsequent rounds [6–8,16].

This architecture is designed to balance three objectives that are especially relevant in distributed transportation security: privacy preservation, communication efficiency, and robustness to statistical heterogeneity. By retaining traffic data at the edge, the framework reduces direct exposure of potentially sensitive local observations. By exchanging only compact model parameters, it avoids full traffic centralization. By training across multiple clients with different local distributions, it aims to improve robustness under realistic deployment variability. The framework also explicitly includes local-only baselines, centralized reference models, threshold calibration on an independent calibration set, validation-based checkpoint selection, multi-seed evaluation, and communication accounting for both uplink and downlink model-transfer traffic.

To provide a compact overview of the proposed methodology and the relationship between the centralized and federated evaluation stages, Figure 1 summarizes the end-to-end workflow adopted in this study, from dataset preparation and scenario-aware partitioning to threshold-calibrated performance assessment and communication-cost analysis.

4.2. Data Preparation and Feature Pipeline

Before training, the traffic data undergo a preprocessing pipeline designed to support both reproducibility and robustness. The raw CSV traffic captures are merged into a unified tabular dataset. Column names are normalized, inconsistent or corrupted label encodings are corrected, and semantically equivalent attack labels are harmonized into a consistent representation. Invalid numeric values, including infinite values and malformed entries, are converted into missing values and subsequently handled through median imputation applied to the numerical feature set.

From the cleaned dataset, three label views are constructed: (i) the original dataset label, (ii) a binary label for benign-versus-attack detection, and (iii) a coarse multi-class label for broader attack-family categorization. In this framework, binary intrusion detection is the primary task used for the main centralized-versus-federated comparison, as it provides the most stable basis for analyzing threshold sensitivity and scenario-aware generalization. The multi-class view is retained only as a complementary perspective because scenario-aware group splitting can create attack-category coverage mismatches that complicate direct interpretation.

Feature vectors are extracted by excluding metadata and label columns while retaining the numerical traffic attributes. In the experimental implementation, the final binary-learning feature space contains 78 numerical input features after preprocessing and label removal. These features are standardized using a training-set-fitted normalization transform before being used by the machine learning models. This step is especially important for neural network stability and also improves comparability across local clients in the federated setting.

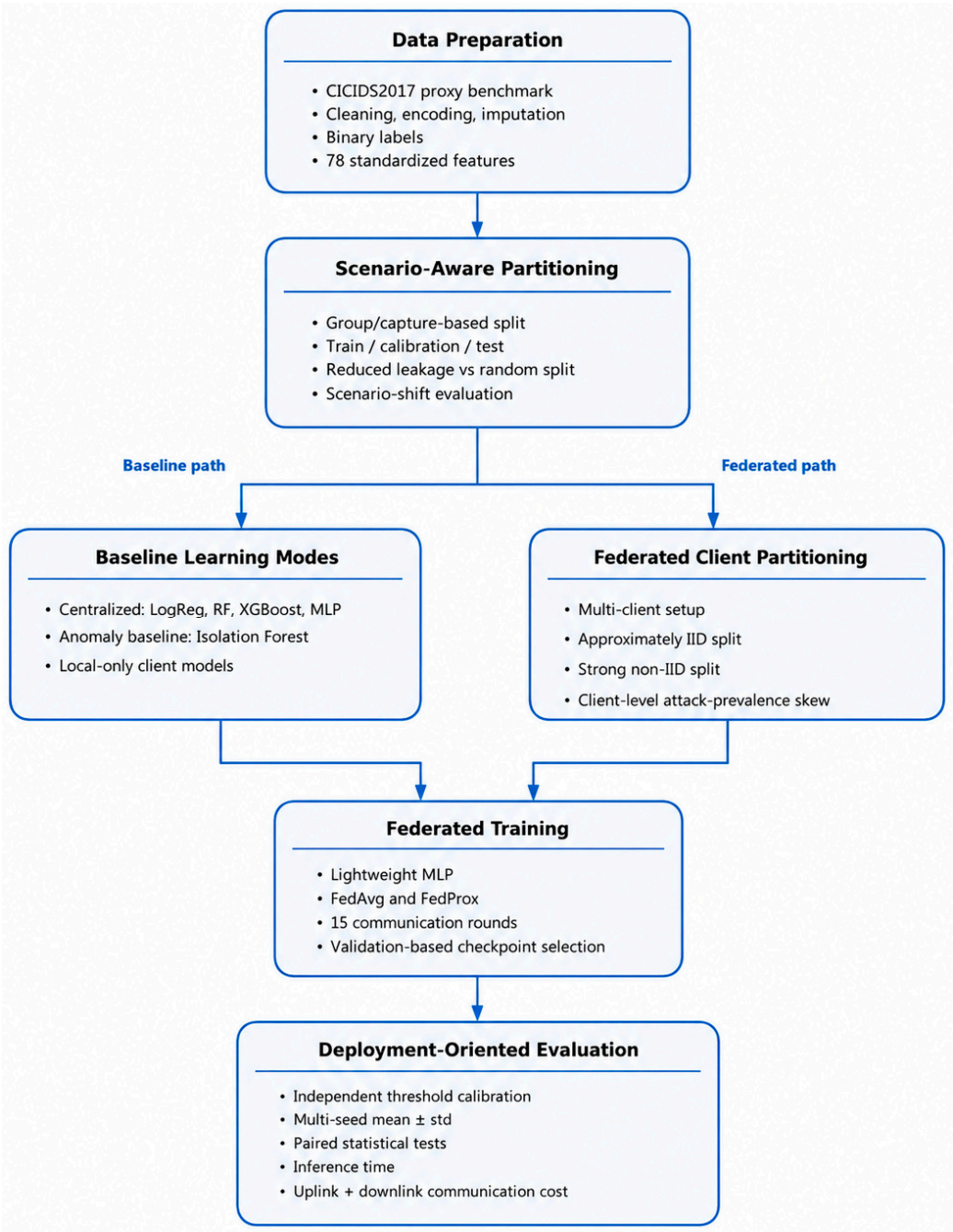

Figure 1. Overview of the proposed federated edge intrusion detection framework and evaluation workflow. The methodology combines dataset preprocessing, group-based scenario-aware rain/calibration/test partitioning, centralized baseline modeling, local-only client baselines, federated training with FedAvg and FedProx under IID and strong non-IID client partitions, independent threshold and checkpoint calibration, multi-seed statistical evaluation, and communication-cost analysis including both uplink and downlink traffic.

A key design aspect of the proposed framework is that the preprocessing and feature definition remain identical across centralized, local-only, and federated experiments. This ensures that the comparison focuses on the learning paradigm and the data partitioning regime rather than on differences in feature engineering. In practical deployments, protocol-aware or domain-specific feature extensions could be incorporated, especially for native V2X telemetry, but the present study intentionally prioritizes a controlled and reproducible comparison. Accordingly, the results should be interpreted as evidence about distributed

learning behavior under a scenario-aware network-intrusion benchmark, not as direct validation of protocol-specific vehicular feature engineering.

4.3. Centralized and Federated Learning Modes

To support a fair and informative assessment, the framework is evaluated under three complementary learning modes: centralized training, local-only training, and federated training.

In the centralized mode, all available training data are pooled into a single training set and used to fit a conventional supervised classifier. This mode serves as a reference baseline and approximates the standard benchmark-driven evaluation strategy frequently adopted in intrusion detection research. In the present study, the centralized baselines include logistic regression, random forest, XGBoost, and a multilayer perceptron (MLP), selected to cover linear, ensemble, boosting-based, and neural model families commonly used in IDS literature. An Isolation Forest anomaly-detection baseline is also included to provide a benign-only unsupervised reference point.

In the local-only mode, each client trains an independent model using only its own local data, without receiving aggregated information from other clients. This setting provides an important reference for determining whether federated collaboration offers advantages over isolated client-side learning. Local-only performance is evaluated both through the average client behavior and through the best client selected using the independent calibration set.

In the federated mode, the same global training data are partitioned into multiple client-local subsets, each representing the private observations of a distributed edge node. Local training is performed independently on each client, after which only the model parameters are sent to the coordination server for aggregation. This allows direct comparison between the centralized and federated paradigms under identical feature definitions and overall training data availability, while explicitly incorporating the impact of client-level distribution differences. In the evaluation, the centralized, local-only, and federated modes use consistent training-set sizes to avoid confounding methodological comparisons with different amounts of training data.

The use of these three learning modes is central to the methodological contribution of the paper. Rather than assuming that federated learning should necessarily outperform centralized learning under all conditions, the framework is designed to investigate when and how federated training remains competitive or even advantageous under more realistic evaluation settings, especially when data partitioning and threshold calibration are treated carefully.

4.4. Local Detection Model

The primary federated detection model used in this study is a lightweight multilayer perceptron (MLP). This choice is deliberate. The objective is not to maximize architectural complexity, but rather to evaluate whether a compact neural model can provide an effective balance between detection capability, training simplicity, and communication efficiency in an edge-assisted federated setting.

The implemented MLP uses a fully connected feed-forward architecture with two hidden layers. In the current experimental configuration, the network receives 78 standardized numerical input features, followed by hidden layers of sizes 128 and 64 with Rectified Linear Unit (ReLU) activations, and a final single-neuron output layer for binary classification. During training, the model is optimized using binary cross-entropy in logit space, while inference applies a sigmoid transformation to produce attack probabilities in the range $[0, 1]$.

This architecture is intentionally lightweight enough to support efficient parameter exchange during federated learning. In the experimental implementation, the resulting serialized model size is approximately 72 kB, which makes it suitable for communication-aware analysis. This compactness is particularly relevant in edge-assisted environments, where model-update overhead should remain limited compared with raw traffic transfer.

The centralized MLP baseline uses a conceptually comparable feed-forward architecture implemented through a standard machine learning pipeline. Although the centralized and federated MLP implementations are not required to be byte-identical at the code level, they are intentionally aligned in capacity and purpose, allowing a meaningful comparison of centralized versus distributed learning behavior under the same feature representation. The MLP is therefore used as a lightweight and reproducible edge-oriented model, not as a claim of architectural novelty over more complex CNN, recurrent, Transformer, or graph-based IDS models.

4.5. Federated Training Workflow

The proposed federated training workflow follows a synchronous round-based Federated Averaging (FedAvg) process. At the beginning of each communication round, the coordination server broadcasts the current global model parameters to the participating edge nodes. Each node then initializes its local model from the global state and performs local optimization on its own private dataset for a predefined number of local epochs. After local training, the updated model parameters are transmitted back to the server, which computes a weighted average of the local parameter tensors according to the number of local training samples per client.

Let θ_t denote the global model parameters at communication round t. Each client k receives θ_t, performs local training on $\mathcal{D}_k$, and returns an updated parameter vector $\theta_t^{(k)}$. The server then computes the next global model as:

$$\theta_{t+1} = \sum_{k=1}^{K} \frac{n_k}{\sum_{j=1}^{K} n_j} \theta_t^{(k)},$$

where n_k is the number of local training samples available at client k. This weighted aggregation rule is consistent with the standard FedAvg formulation introduced by [6].

To provide an additional federated baseline under heterogeneous client data, the framework also includes FedProx [16]. In FedProx, each client minimizes a local objective augmented with a proximal term that penalizes excessive deviation from the current global model:

$$\min_{\theta} F_k(\theta) + \frac{\mu}{2} |\theta - \theta_t|^2,$$

where μ controls the strength of the proximal regularization. When $\mu = 0$, the local objective reduces to the standard FedAvg local training objective. FedProx is included because it is specifically designed to improve stability under heterogeneous client distributions, which is relevant to the strong non-IID setting considered in this study.

In the current experimental study, the federated workflow is instantiated using a manual FedAvg/FedProx implementation with a fixed number of clients, one local epoch per round, mini-batch stochastic optimization, and a predefined number of communication rounds. This setup was intentionally chosen to provide a transparent and reproducible baseline rather than a heavily optimized federated system. The resulting configuration is therefore appropriate for methodological comparison and can later be extended with client subsampling, asynchronous updates, adaptive aggregation, or more advanced personalization strategies.

4.6. IID and Strong Non-IID Client Modeling

A central aspect of the proposed framework is the explicit modeling of client heterogeneity. To capture this effect, the federated experiments consider two distinct client partitioning regimes.

In the first regime, the training data are partitioned into approximately IID clients, where each client receives a subset of samples with class proportions broadly similar to the global training distribution. This setting provides a useful reference point and approximates the behavior of federated learning when local nodes observe statistically comparable traffic mixtures.

In the second regime, a strong non-IID configuration is constructed by deliberately assigning substantially different attack prevalences to different clients. In the current binary experiments, this is achieved by creating clients with progressively increasing malicious-traffic ratios, thereby simulating distributed edge nodes deployed in environments with different exposure levels to suspicious or malicious activity. This type of partitioning is particularly relevant for V2X-inspired deployments, where some nodes may monitor relatively benign traffic while others may be exposed to attack-heavy hotspots or atypical operating conditions. The specific non-IID construction should be interpreted as an experimental abstraction of heterogeneous attack exposure across monitoring locations or operational periods, rather than as a measured distribution from a real vehicular deployment.

The inclusion of both regimes is methodologically important. Many federated intrusion detection studies acknowledge non-IID data as a challenge, but relatively few explicitly compare approximately IID and strongly heterogeneous settings under the same controlled framework. By doing so, the proposed methodology allows the practical robustness of the federated model to be assessed more realistically. The analysis further avoids interpreting non-IID performance differences as definitive unless supported by the multi-seed variability and paired statistical comparisons.

4.7. Scenario-Aware Evaluation and Group-Based Splitting

Another key contribution of the proposed framework lies in the evaluation protocol. In addition to conventional random train-test partitioning, the framework introduces a stricter scenario-aware evaluation based on group- or capture-level separation. Under this protocol, the training and test sets are formed from disjoint traffic capture files, so that the model is evaluated on traffic scenarios not seen during training.

This design is especially important for intrusion detection research. Random splitting can produce highly optimistic results because closely related flows, session artifacts, or scenario-specific statistical signatures may appear in both the training and test subsets. In contrast, capture-level separation creates a more realistic generalization challenge by forcing the model to transfer across traffic contexts. In the present work, this evaluation protocol proved highly informative, as it revealed a substantial performance gap between optimistic random-split baselines and more realistic scenario-separated testing.

From a practical perspective, this protocol better approximates real deployment conditions in distributed edge-assisted security systems, where a model trained on past traffic from some locations or time periods must often generalize to different future traffic contexts. For this reason, the group-based split is treated as the primary realism-oriented benchmark in the paper, while the random split is retained mainly as a reference baseline.

In the protocol, the group-based evaluation is extended from a simple train/test split to a train/calibration/test design. The calibration set is used for threshold selection and federated checkpoint selection, while the final test set is kept separate for unbiased performance reporting. This modification is essential because threshold tuning or best-round

selection on the test set would overestimate the practical operating-point performance of the IDS.

4.8. Threshold-Aware Edge Inference

The proposed framework treats edge inference as a probability-based decision process rather than a fixed-threshold hard classification rule. For binary intrusion detection, the model outputs an estimated attack probability $\hat{p}(x)$ for each observed traffic sample. A decision threshold τ is then applied such that:

$$\hat{p}(x) = \begin{cases} 1, \; if \; \hat{p}(x) \; \geq \; \tau \\ 0, \; \text{otherwise.} \end{cases}$$

In many benchmark-driven studies, the default threshold $\tau = 0.5$ is implicitly adopted without further analysis. However, in imbalanced intrusion detection settings, especially under scenario shift and heterogeneous distributed data, this default choice may be suboptimal. A threshold that is too conservative may yield high precision but poor recall, which can be operationally undesirable when suspicious traffic should be flagged aggressively enough to avoid excessive missed detections.

For this reason, the proposed framework explicitly incorporates threshold sensitivity analysis as part of the deployment-oriented evaluation. Rather than assuming that the default threshold is adequate, the framework assesses how the operating point affects the precision–recall trade-off and the resulting F1-score. This is particularly relevant for edge-assisted intrusion detection, where acceptable operating points may depend on local tolerance for false positives, safety requirements, or downstream analyst capacity.

In the evaluation, the decision threshold is selected exclusively on the independent calibration set and is then applied unchanged to the final test set. Consequently, the reported test-set metrics reflect a deployable calibration procedure rather than an oracle choice of the best threshold on the test data. This change also means that the selected threshold should not be interpreted as a universal value; instead, it is a model- and split-dependent operating point estimated from calibration data.

4.9. Communication-Aware Monitoring Metrics

Beyond conventional classification performance, the framework also tracks communication-aware and deployment-relevant indicators. These include the serialized model size, the aggregate upload volume per communication round, the aggregate download volume per communication round, the total bidirectional communication volume, the per-round execution time, and the local inference time. Such indicators do not fully characterize end-to-end deployment cost, but they provide useful first-order proxies for assessing whether the model remains practical in distributed edge environments.

For a model of serialized size (M) bytes and (K) participating clients, the client-to-server uplink cost per round is approximately (K M), while the server-to-client downlink cost is also approximately (K M) when the global model is redistributed to all clients. The total bidirectional communication per round is therefore approximated as

$$C_{\text{round}} \approx 2KM,$$

and the cumulative communication cost over (R) rounds is

$$C_{\text{total}} \approx 2KMR.$$

This accounting avoids underestimating communication requirements by considering only client uploads. In the present implementation, the lightweight MLP has a serialized

size of approximately 72 kB; with five clients, this corresponds to approximately 720 kB per communication round and approximately 11.06 MB over 15 rounds.

This perspective is important because a model with marginally higher classification performance may still be less attractive in operational settings if it requires significantly greater communication overhead or computational burden. Conversely, a lightweight model that remains competitive under realistic data heterogeneity may be preferable for practical deployment. The proposed framework therefore intentionally complements conventional detection metrics with system-oriented measurements, in line with the deployment constraints emphasized in federated learning and edge computing literature [6,7,13].

4.10. Practical Interpretation of the Proposed Framework

Taken together, the above components define a federated edge-oriented intrusion detection framework whose primary contribution is methodological and deployment-oriented. The framework does not claim to be a fully protocol-specific V2X production IDS, nor does it assume secure federated aggregation under adversarial model updates. Instead, it provides a reproducible and practically grounded architecture for evaluating whether privacy-preserving collaborative learning can remain effective under realistic distribution shift, client heterogeneity, and threshold-sensitive operation.

This positioning is intentional and important. In many real distributed security environments, the most relevant scientific question is not whether a highly optimized model can achieve a slightly higher benchmark score under favorable conditions, but whether a compact and privacy-aware collaborative model can maintain useful detection performance when evaluated under more realistic and operationally meaningful constraints. The proposed framework is designed precisely to answer that question. Accordingly, the framework should be viewed as a controlled methodological step toward federated edge IDS deployment, with native V2X datasets, protocol-aware feature extraction, compromised-client defenses, asynchronous operation, and real RSU or vehicular-edge implementation left as necessary future extensions.

5. Experimental Methodology

This section describes the dataset preparation process, the evaluation protocols, the centralized baseline models, the local-only baselines, the federated learning configurations, the calibration procedure, and the performance metrics used in the study. The experimental methodology was intentionally designed not only to compare centralized and federated learning under standard benchmark conditions, but also to assess their behavior under more realistic distribution-shift and client-heterogeneity scenarios. It should be noted that the objective is not to maximize architecture complexity, but to establish a transparent, statistically repeated, and communication-aware federated baseline. The experimental protocol uses consistent training-set sizes across centralized and federated methods, an independent calibration set for threshold and checkpoint selection, local-only and federated baselines, five independent random seeds, and communication accounting that includes both uplink and downlink traffic.

5.1. Dataset and Preprocessing Pipeline

The experimental analysis was conducted using the CICIDS2017 dataset [19], which is one of the most widely used public benchmarks for network intrusion detection research. Although CICIDS2017 is not a native V2X dataset, it provides a rich and diverse set of labeled benign and malicious traffic flows generated under realistic network activity and multiple attack scenarios. In the present work, CICIDS2017 is used as a reproducible proxy benchmark to emulate distributed edge-based intrusion detection conditions in a V2X-

inspired setting. Consequently, the dataset is used to study distributed intrusion detection methodology under scenario shift and client heterogeneity, not to claim direct protocol-level validation on DSRC, C-V2X, CAM/DENM, or other vehicular communication traces.

The raw CSV capture files were first consolidated into a unified tabular dataset. A preprocessing pipeline was then applied to improve consistency and reproducibility. The main preprocessing steps included: (i) normalization of column names, (ii) correction of inconsistent or corrupted label encodings, (iii) harmonization of semantically equivalent attack labels, (iv) conversion of infinite or malformed numeric values into missing values, and (v) median imputation of missing numerical features. This produced a clean feature matrix suitable for both conventional machine learning and neural network-based training.

From the cleaned dataset, three label representations were derived: the original label provided by the dataset, a binary label for benign-versus-attack detection, and a coarse multi-class label for broader attack-family categorization. The primary analysis in this paper focuses on binary intrusion detection because it provides the most stable and deployment-relevant basis for comparing centralized, local-only, and federated learning under realistic partitioning constraints.

After excluding metadata and label-related fields, the resulting binary-learning feature representation contained 78 numerical features. These features were standardized using a training-set-fitted normalization transform prior to model training. This normalization step was applied consistently across centralized, local-only, and federated experiments to ensure comparability.

5.2. Binary and Coarse Multi-Class Task Definitions

Two learning tasks were defined from the processed dataset.

The first and primary task is binary intrusion detection, where each flow is classified as either benign or malicious. This task reflects the most common operational requirement in edge-assisted intrusion detection, namely to distinguish suspicious traffic from normal traffic with sufficient reliability to support alerting or downstream investigation. All main quantitative comparisons reported in the study are therefore anchored in the binary task.

The second task is a coarse multi-class intrusion categorization problem, where the original attack labels are grouped into broader attack families. This task is included as a complementary and more exploratory extension of the binary setting. However, unlike binary detection, the multi-class setting is more sensitive to label coverage mismatch across train-test partitions, particularly under scenario-aware splitting. In practice, some attack categories may appear only in held-out capture files and therefore be absent from the training set. This issue is itself informative from a realism perspective, but it complicates direct centralized-versus-federated comparison. For this reason, the main conclusions of the present study are intentionally anchored in the binary detection results. Accordingly, the main claims do not rely on the exploratory multi-class results.

5.3. Random-Split Baseline Protocol

As a conventional reference baseline, a random train-test split protocol was first considered. Under this protocol, the full processed dataset is randomly partitioned into training and test subsets while approximately preserving the global class distribution. This setup corresponds to the standard evaluation strategy used in a large fraction of benchmark-based intrusion detection literature and provides an optimistic upper-bound reference for expected classification performance.

The random-split results are not treated as the primary realism-oriented benchmark in this paper. Instead, they serve as a calibration point that helps reveal how strongly performance may be overestimated when closely related traffic observations or scenario-

specific patterns are allowed to appear in both training and testing subsets. This contrast becomes particularly relevant when comparing the random-split results with the stricter scenario-aware protocol described next.

5.4. Scenario-Aware Group-Split and Calibration Protocol

To obtain a more realistic and deployment-relevant evaluation, a stricter group- or scenario-aware split was defined using disjoint original traffic capture files. Rather than randomly partitioning individual records, the training-side pool and the final test set were formed from different CICIDS2017 capture files, thereby ensuring that the final evaluation was performed on traffic scenarios not seen during training.

In the binary group-split experiments, the training-side pool was constructed from the following capture files:

- Monday-WorkingHours.pcap_ISCX.csv;
- Tuesday-WorkingHours.pcap_ISCX.csv;
- Wednesday-workingHours.pcap_ISCX.csv;
- Thursday-WorkingHours-Morning-WebAttacks.pcap_ISCX.csv;
- Friday-WorkingHours-Morning.pcap_ISCX.csv.

The final test set was constructed from the following disjoint capture files:

- Thursday-WorkingHours-Afternoon-Infilteration.pcap_ISCX.csv;
- Friday-WorkingHours-Afternoon-PortScan.pcap_ISCX.csv;
- Friday-WorkingHours-Afternoon-DDos.pcap_ISCX.csv.

This protocol intentionally creates a more difficult generalization setting by forcing the model to transfer across different capture sessions and attack compositions. From an operational standpoint, this better approximates a deployment in which a model trained on historical or geographically distinct traffic must generalize to new environments or future traffic conditions.

In the evaluation, the original training-side pool was further divided into a training subset and an independent calibration subset. The training subset was used to fit centralized, local-only, and federated models. The calibration subset was used only for threshold selection and, in the case of federated learning, validation-based checkpoint selection. The final test set remained fully held out until the final evaluation stage. This design avoids tuning the binary decision threshold or selecting the best communication round directly on the test data.

Because the full processed dataset is large, stratified subsampling was used in the main group-split binary experiments to keep the training, calibration, and testing workloads tractable while preserving class proportions within the corresponding parent pools. In the protocol, all main centralized, local-only, and federated comparisons used the same effective training-set size. Specifically, 300,000 samples were used for training, 60,000 samples were used for calibration, and 150,000 samples were used for final testing. This harmonized design ensures that centralized, local-only, and federated methods are compared under the same effective training, calibration, and testing sample budgets.

5.5. Centralized and Anomaly-Detection Baseline Models

To establish a robust centralized reference, four supervised models were evaluated under the centralized setting:

1. Logistic Regression (LogReg);
2. Random Forest (RF);
3. Extreme Gradient Boosting (XGBoost);
4. Multilayer Perceptron (MLP).

These models were selected to cover a representative spectrum of common IDS approaches, including linear classification, bagged decision-tree ensembles, gradient-boosted trees, and neural models. Together, they provide a useful baseline family for assessing whether the proposed federated approach remains competitive under both optimistic and realism-oriented evaluation conditions.

In addition, an Isolation Forest baseline was included as an unsupervised anomaly-detection reference. This model was trained using benign training samples and evaluated as a benign-only anomaly detector on the binary test task. Its inclusion allows comparison with a classical anomaly-detection approach that does not require fully supervised attack-label coverage during training.

Under the group-split binary protocol, the centralized models were trained on the 300,000-sample training subset, calibrated on the 60,000-sample calibration subset, and evaluated on the 150,000-sample final test subset. The results showed that centralized performance under scenario-aware splitting was substantially lower than under conventional random partitioning, thereby confirming that the stricter evaluation protocol provides a much more challenging and realistic assessment of generalization.

5.6. Local-Only Client Baselines

To determine whether federated collaboration provides benefits beyond isolated client-side learning, local-only baselines were added to the experimental protocol. In this setting, the global training subset was partitioned into client-local datasets using the same IID and strong non-IID partitioning strategies later used for federated training. Each client then trained an independent lightweight MLP using only its own local data, without receiving model updates from the other clients and without participating in server-side aggregation.

Two local-only summaries were reported. The first is the mean client performance, which describes the expected behavior of isolated local training across all clients. The second is the best client selected by calibration performance, which represents a more favorable local-only reference in which one client model is selected using the independent calibration set and then evaluated on the final test set. These baselines are important because they distinguish the value of federated collaboration from the performance that could be achieved by purely local models.

5.7. Federated Binary Learning Configuration

The primary federated learning experiments were implemented using a manual synchronous federated workflow combined with a lightweight binary multilayer perceptron (MLP). The experiments include both FedAvg and FedProx. FedAvg provides the standard weighted model-averaging baseline, whereas FedProx adds a proximal regularization term during local training to improve stability under heterogeneous client data. Both methods use the same model architecture, data partitions, optimization settings, calibration protocol, and communication-round budget.

A lightweight MLP was intentionally selected as the federated base learner because it offers a transparent, reproducible, and communication-aware neural baseline suitable for edge-side training. The objective of the present work is not to maximize architectural complexity, but to isolate the effects of scenario-aware partitioning, client heterogeneity, threshold calibration, checkpoint selection, and communication rounds under a controlled federated setup.

The federated binary MLP used 78 standardized numerical input features, followed by two fully connected hidden layers with 128 and 64 neurons, Rectified Linear Unit (ReLU) activations, and a final single-neuron output layer. Training used binary cross-entropy in logit space. The optimization was performed using the Adam optimizer with a learning rate

of 10^{-3}, one local epoch per communication round, and mini-batches of size 256. A total of 15 communication rounds were executed. The global model parameters were updated using weighted aggregation, where the contribution of each client was proportional to its local sample count.

For FedProx, the local training objective was augmented with a proximal term controlled by the parameter (μ), penalizing deviations from the current global model. This modification was introduced to provide a stronger federated baseline for non-IID settings while keeping the experimental design lightweight and reproducible.

The corresponding test set for federated evaluation consisted of the 150,000-sample stratified group-split test subset. For each seed, the best federated checkpoint was selected using the calibration set, and the selected checkpoint was then evaluated once on the final test set using the threshold selected on the same calibration set.

5.8. IID and Strong Non-IID Federated Client Partitions

The federated experiments considered two client partitioning regimes.

In the approximately IID configuration, the 300,000-sample training subset was partitioned into five clients of approximately equal size. The client partitions were formed by shuffling the global training indices and splitting them into equal segments, which resulted in local datasets with broadly similar class proportions. This setting provides a reference point for federated learning when the local data distributions are relatively balanced.

To explicitly assess the effect of client heterogeneity, a strong non-IID binary configuration was also constructed. In this setting, the 300,000-sample training subset was repartitioned into five client-local datasets with deliberately different malicious-traffic prevalences. Because the global attack prevalence of the training subset was approximately 13.33%, the non-IID design used client attack ratios centered around this value rather than fixed low attack ratios. The target malicious ratios were approximately 7.33%, 10.33%, 13.33%, 16.33%, and 19.33%, respectively.

With five clients of 60,000 samples each, this corresponds approximately to the following class distributions:

- Client 0: 55,600 benign/4400 attack samples;
- Client 1: 53,800 benign/6200 attack samples;
- Client 2: 52,000 benign/8000 attack samples;
- Client 3: 50,200 benign/9800 attack samples;
- Client 4: 48,400 benign/11,600 attack samples.

The local model architecture, optimization settings, number of rounds, and aggregation strategy were kept consistent across the IID and strong non-IID federated configurations. This allows the observed performance differences to be attributed primarily to client heterogeneity rather than to architectural or optimization changes.

The strong non-IID design is particularly relevant to the target application domain because edge nodes in distributed transportation or cyber-physical systems are unlikely to observe identically distributed traffic. Instead, different nodes may operate under different attack exposure levels, traffic compositions, or environmental contexts. The strong non-IID experiment therefore provides a more realistic stress test of federated learning robustness. Nevertheless, this design should be interpreted as a controlled abstraction of heterogeneous attack exposure, not as a measured distribution from a real V2X deployment.

5.9. Threshold and Checkpoint Calibration

A dedicated threshold calibration procedure was conducted for the binary group-split experiments in order to assess whether the default decision threshold of 0.5 was appropriate. This analysis was motivated by the observation that several models exhibited

high precision but relatively low recall under the default threshold, suggesting an overly conservative operating point.

In the protocol, threshold selection was performed exclusively on the independent calibration set. For each trained model, a grid of candidate thresholds was evaluated on calibration predictions, and the threshold maximizing the calibration F1-score was selected. This selected threshold was then applied unchanged to the final test set. The test set was therefore not used for threshold tuning.

For federated models, the calibration set was also used to select the best communication-round checkpoint. Specifically, each round produced a candidate global model. Calibration performance was computed for each candidate checkpoint, and the best checkpoint was selected according to calibration F1-score after threshold optimization on the calibration set. The selected round and threshold were then carried forward to the final test evaluation. This procedure provides a realistic deployable calibration protocol by ensuring that the final test set is not used for threshold or checkpoint selection.

The selected thresholds should not be interpreted as universal IDS operating points. They are model-, seed-, and split-dependent values estimated from calibration data. In practical deployment, threshold selection would need to reflect operational preferences regarding false positives, missed detections, analyst workload, and safety requirements.

5.10. Performance Metrics and Statistical Reporting

The experiments were evaluated using a combination of classification-oriented, statistical, and deployment-aware metrics.

For binary intrusion detection, the primary classification metrics were:

1. Accuracy;
2. Precision;
3. Recall;
4. F1-score.

Among these, F1-score was treated as the most informative primary metric because the binary task remains class-imbalanced and because the balance between false positives and false negatives is particularly relevant in intrusion detection.

To strengthen statistical robustness, all main group-split experiments were repeated over five independent random seeds: 42, 123, 2026, 777, and 999. Results are reported as mean $\pm$ standard deviation across seeds. Confidence intervals and paired seed-level statistical comparisons were also computed where appropriate. This design addresses the sensitivity of federated learning to random initialization, client partitioning, stochastic optimization, and checkpoint selection.

For the federated experiments, additional system-level indicators were recorded, including:

1. Per-round execution time;
2. Serialized model size in bytes;
3. Client-to-server uplink volume per communication round;
4. Server-to-client downlink volume per communication round;
5. Total bidirectional communication volume;
6. Cumulative communication volume over the full round budget;
7. Approximate inference time on the test set.

These indicators do not constitute a full networked deployment benchmark, but they provide useful first-order proxies for communication and computational overhead in an edge-assisted setting. In particular, the communication accounting includes both upload and download traffic, avoiding the underestimation that would result from considering only client-to-server updates.

5.11. Implementation Details and Reproducibility Notes

The experiments were implemented in Python 3.11 using a combination of standard machine-learning and deep-learning libraries. The centralized baselines used scikit-learn 1.6.1, while the XGBoost baseline used XGBoost 2.1.4. The federated binary MLP experiments were implemented manually using PyTorch 2.8.0+cpu, with explicit client partitioning, local training loops, FedAvg/FedProx updates, and server-side weighted parameter aggregation.

To support reproducibility, the preprocessing pipeline, partitioning logic, calibration procedure, baseline training, and federated training procedure were implemented as separate scripts. Fixed random seeds were used where possible to reduce run-to-run variability. The study intentionally prioritizes a transparent and reproducible baseline implementation rather than a heavily optimized or framework-dependent federated stack.

It should also be noted that the coarse multi-class scenario-aware evaluation revealed an important practical issue: some attack categories may appear in the held-out test captures but not in the selected training captures, leading to unseen-label conditions. This effect is itself informative because it reflects a realistic limitation of strict scenario-separated evaluation. However, it also means that the multi-class analysis requires additional care, such as restricting evaluation to shared classes or redesigning the partitioning protocol. For this reason, the present paper places its primary emphasis on the binary intrusion detection findings, which provide the clearest and most robust centralized-versus-local-only-versus-federated comparison.

6. Results

This section presents the experimental results obtained under the scenario-aware group-based protocol described in Section 5. The analysis is structured around four methodological principles: consistent training-set sizes across learning modes, independent calibration-based threshold and checkpoint selection, multi-seed statistical reporting, and bidirectional communication-cost accounting. First, centralized, local-only, and federated models are evaluated using the same effective training-set size. Second, thresholds and federated checkpoints are selected using an independent calibration set rather than the final test set. Third, all main results are reported over five independent random seeds. Fourth, communication cost is reported as bidirectional model-transfer traffic, including both client-to-server uplink and server-to-client downlink. Together, these design choices provide a more conservative and methodologically defensible assessment of federated intrusion detection under scenario shift and client heterogeneity.

6.1. Centralized and Anomaly-Detection Baselines Under Scenario-Aware Evaluation

To establish reference points before the federated experiments, several centralized binary classifiers were evaluated on the CICIDS2017 binary task using the group-based train/calibration/test protocol described in Section 5. The objective of this stage was to determine how conventional centralized models behave under a stricter scenario-separated evaluation setting and to provide meaningful baselines for the subsequent local-only and federated comparisons.

Under this group-based split, the centralized models exhibited substantially more conservative performance than typically observed under conventional random partitioning. This confirms that capture-level separation introduces a strong distribution shift between training and testing conditions and therefore provides a more realistic assessment of generalization. The results also show a common pattern across several models: precision remained high, but recall was comparatively low, indicating that the classifiers tended to be conservative when predicting the attack class.

Table 1 summarizes the centralized supervised baselines and the Isolation Forest anomaly-detection baseline. Random Forest achieved the most stable centralized result, with an F1-score of 0.440 ± 0.002, very high precision, and low run-to-run variability. The centralized MLP achieved a slightly higher mean F1-score, 0.445 ± 0.058, but with greater variability across seeds. Logistic regression achieved a similar mean F1-score, 0.444 ± 0.040, with lower precision but slightly higher recall. XGBoost showed high precision but lower mean F1-score and greater variability in the runs. The Isolation Forest baseline was clearly less effective in this binary task, mainly due to low recall.

Table 1. Centralized and anomaly-detection baseline performance under the scenario-aware group-based protocol. Results are reported as mean ± standard deviation over five seeds.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.718 ± 0.013	0.757 ± 0.021	0.315 ± 0.038	0.444 ± 0.040
Random Forest	0.742 ± 0.001	0.986 ± 0.002	0.284 ± 0.001	0.440 ± 0.002
XGBoost	0.727 ± 0.027	0.984 ± 0.008	0.242 ± 0.076	0.384 ± 0.107
MLP	0.737 ± 0.017	0.902 ± 0.041	0.297 ± 0.049	0.445 ± 0.058
Isolation Forest	0.681 ± 0.011	0.712 ± 0.028	0.182 ± 0.038	0.288 ± 0.052

These centralized results support two conclusions. First, the scenario-aware protocol is substantially more challenging than random splitting and therefore provides a stricter benchmark for deployment-oriented IDS evaluation. Second, no single centralized model clearly dominates all others under this protocol. Random Forest is the most stable, MLP and logistic regression achieve similar mean F1-scores with higher variability, and XGBoost is affected by one lower-performing seed. For this reason, the subsequent comparison should be interpreted in terms of competitiveness and robustness under realistic constraints, rather than as a simple search for a single universally best classifier.

Figure 2 shows the mean F1-score with error bars across the centralized and anomaly-detection baselines.

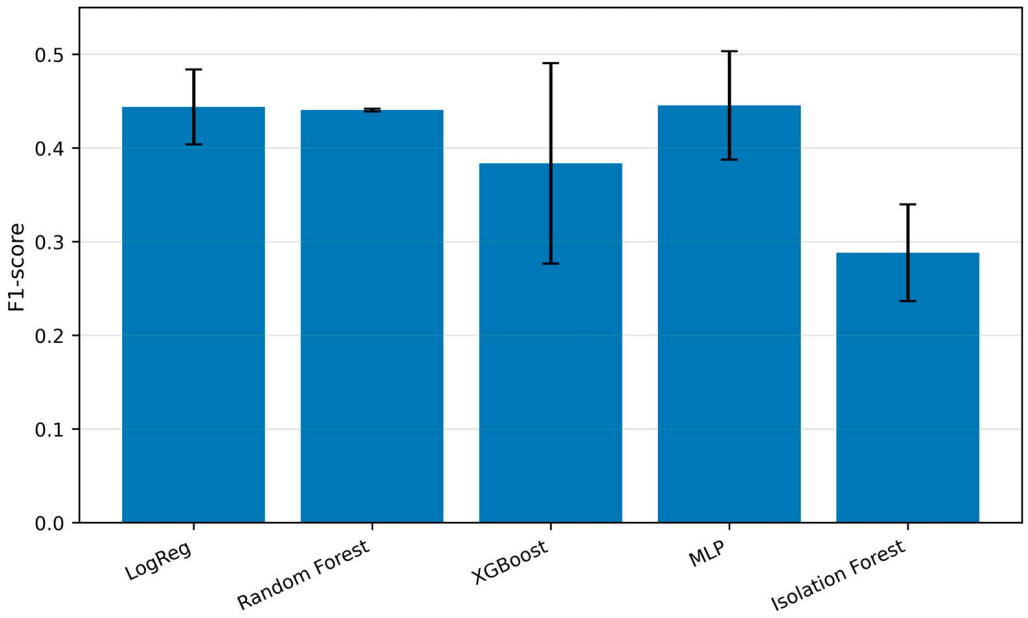

Figure 2. Mean final-test F1-score of the centralized and anomaly-detection baselines under the scenario-aware group-based protocol. Error bars denote one standard deviation across five random seeds.

6.2. Local-Only and Federated Binary IDS Performance

After establishing centralized references, the study evaluated local-only and federated binary intrusion detection using the same 300,000-sample training subset, 60,000-sample calibration subset, and 150,000-sample final test subset. The local-only baselines were included to determine whether collaborative learning provides benefits beyond isolated client-side training. The federated configurations included both FedAvg and FedProx, evaluated under approximately IID and strong non-IID client partitions.

Table 2 summarizes the full final-test comparison. The highest mean F1-score was obtained by FedAvg under the strong non-IID partition, with (F1 = 0.468 $\pm$ 0.028). FedAvg under the approximately IID partition followed closely, with (F1 = 0.463 $\pm$ 0.036). The best local-only strong non-IID client selected by calibration achieved (F1 = 0.461 $\pm$ 0.017), while the mean local-only strong non-IID performance was (F1 = 0.452 $\pm$0.018). These results indicate that federated learning remained competitive with centralized and local-only alternatives under the scenario-aware test protocol.

Table 2. Final-test performance of centralized, local-only, and federated methods under the scenario-aware protocol. Results are reported as mean $\pm$ standard deviation over five seeds.

Method	Learning Mode	Partition	Accuracy	Precision	Recall	F1-Score
FedAvg	Federated	Strong non-IID	0.734 $\pm$ 0.011	0.827 $\pm$ 0.055	0.328 $\pm$ 0.027	0.468 $\pm$ 0.028
FedAvg	Federated	IID	0.737 $\pm$ 0.013	0.863 $\pm$ 0.042	0.318 $\pm$ 0.033	0.463 $\pm$ 0.036
Local-only best client	Local-only	Strong non-IID	0.734 $\pm$ 0.008	0.844 $\pm$ 0.056	0.318 $\pm$ 0.018	0.461 $\pm$ 0.017
Local-only mean client	Local-only	Strong non-IID	0.732 $\pm$ 0.006	0.848 $\pm$ 0.015	0.309 $\pm$ 0.017	0.452 $\pm$ 0.018
MLP	Centralized	Centralized	0.737 $\pm$ 0.017	0.902 $\pm$ 0.041	0.297 $\pm$ 0.049	0.445 $\pm$ 0.058
Logistic Regression	Centralized	Centralized	0.718 $\pm$ 0.013	0.757 $\pm$ 0.021	0.315 $\pm$ 0.038	0.444 $\pm$ 0.040
Local-only mean client	Local-only	IID	0.732 $\pm$ 0.005	0.876 $\pm$ 0.010	0.299 $\pm$ 0.014	0.443 $\pm$ 0.016
Random Forest	Centralized	Centralized	0.742 $\pm$ 0.001	0.986 $\pm$ 0.002	0.284 $\pm$ 0.001	0.440 $\pm$ 0.002
FedProx	Federated	IID	0.733 $\pm$ 0.011	0.898 $\pm$ 0.054	0.289 $\pm$ 0.019	0.437 $\pm$ 0.025
FedProx	Federated	Strong non-IID	0.730 $\pm$ 0.008	0.896 $\pm$ 0.062	0.281 $\pm$ 0.002	0.428 $\pm$ 0.008
Local-only best client	Local-only	IID	0.729 $\pm$ 0.012	0.905 $\pm$ 0.057	0.273 $\pm$ 0.018	0.420 $\pm$ 0.026
XGBoost	Centralized	Centralized	0.727 $\pm$ 0.027	0.984 $\pm$ 0.008	0.242 $\pm$ 0.076	0.384 $\pm$ 0.107
Isolation Forest	Anomaly baseline	Centralized	0.681 $\pm$ 0.011	0.712 $\pm$ 0.028	0.182 $\pm$ 0.038	0.288 $\pm$ 0.052

The differences among the strongest supervised methods should be interpreted cautiously. Although FedAvg under strong non-IID partitioning achieved the highest mean F1-score, paired comparisons did not show a statistically strong advantage over the main centralized baselines. For example, the F1-score difference between FedAvg strong non-IID and Random Forest was positive in favor of FedAvg but did not reach conventional significance in the paired (t)-test ($p \approx 0.093$). Similarly, the differences between FedAvg and the centralized MLP or logistic regression were modest relative to the observed variability. Therefore, the most defensible conclusion is not that federated learning clearly outperforms centralized learning, but that lightweight FedAvg remains competitive under a substantially more demanding scenario-aware protocol while preserving data locality.

The comparison with local-only learning is also informative. The best local-only strong non-IID client selected by calibration achieved a mean F1-score close to FedAvg, suggesting that some local client models can perform well when selected using validation data. However, the federated models provide a shared global detector rather than relying

on a single isolated client. This distinction is important in distributed deployments where the goal is to learn from multiple observation points without centralizing raw traffic.

FedProx did not improve over FedAvg in the final experiments. Although FedProx is designed to mitigate client drift under heterogeneous data, the present configuration used one local epoch per round and a relatively compact MLP, conditions under which the additional proximal term may have provided limited benefit or may have constrained useful local adaptation. This result is valuable because it shows that adding a non-IID-oriented federated optimizer does not automatically improve IDS performance; the benefit of such methods depends on the data partitioning, local optimization schedule, proximal coefficient, and model capacity.

To provide a compact visual comparison across the evaluated learning paradigms, Figure 3 reports the mean final-test F1-score of the centralized, local-only, anomaly-detection, and federated methods under the scenario-aware group-based protocol.

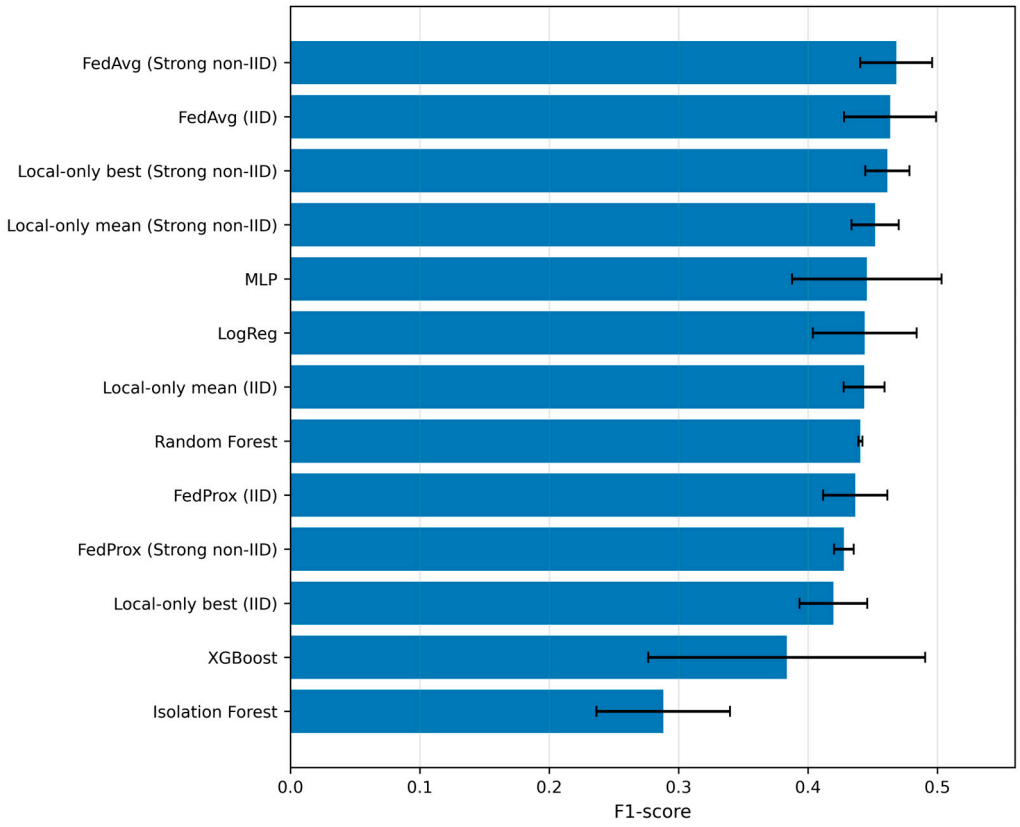

Figure 3. Mean final-test F1-score of centralized, local-only, and federated methods under the scenario-aware protocol. Error bars denote one standard deviation across five seeds.

6.3. Threshold Calibration and Federated Checkpoint Selection

Threshold calibration was a central component of the evaluation protocol. Threshold selection was performed exclusively on the independent calibration set, and the selected threshold was then applied unchanged to the final test set. This procedure avoids optimistic bias from test-set tuning and provides a more realistic operating-point selection process for deployment-oriented intrusion detection.

The selected thresholds varied across methods, partitions, and seeds, confirming that the operating point is not universal. The median selected threshold was 0.40 for FedAvg under strong non-IID partitioning and 0.57 for FedAvg under IID partitioning. The centralized MLP used a median selected threshold of 0.50, while logistic regression used 0.38

and Random Forest used 0.41. These values confirm that the operating threshold is model-, seed-, and split-dependent, and should therefore not be interpreted as a universal value.

Federated checkpoint selection was also performed using the calibration set. Under the final 15-round budget, the selected checkpoints generally occurred in later communication rounds rather than in the first few rounds. The median selected round was 13 for FedAvg under strong non-IID partitioning, 15 for FedAvg under IID partitioning, and 15 for both FedProx configurations. This finding clarifies the interpretation of the communication-round behavior: under the calibration protocol, additional rounds remained useful within the tested budget, although they increased communication cost linearly.

As can be seen in Table 3, the gap between calibration-selected performance and final test performance also reinforces the difficulty of the scenario-aware split. Even when thresholds and checkpoints are selected carefully using calibration data, the final test set remains challenging because it is drawn from disjoint capture files with different attack composition. This behavior is desirable from an evaluation perspective, as it prevents overly optimistic conclusions and better reflects the deployment problem of transferring IDS models to future or geographically distinct traffic scenarios.

Table 3. Calibration-selected operating points for the main federated configurations. Thresholds and checkpoints were selected on the calibration set and then evaluated on the final test set.

Method	Partition	Median Selected Threshold	Median Selected Round	Mean F1-Score
FedAvg	Strong non-IID	0.40	13	0.468 ± 0.028
FedAvg	IID	0.57	15	0.463 ± 0.036
FedProx	IID	0.45	15	0.437 ± 0.025
FedProx	Strong non-IID	0.45	15	0.428 ± 0.008

Figure 4 shows the distribution of selected communication rounds across seeds for FedAvg and FedProx under IID and strong non-IID partitions. As can be seen, the best validation-selected checkpoints generally occurred near the end of the 15-round budget.

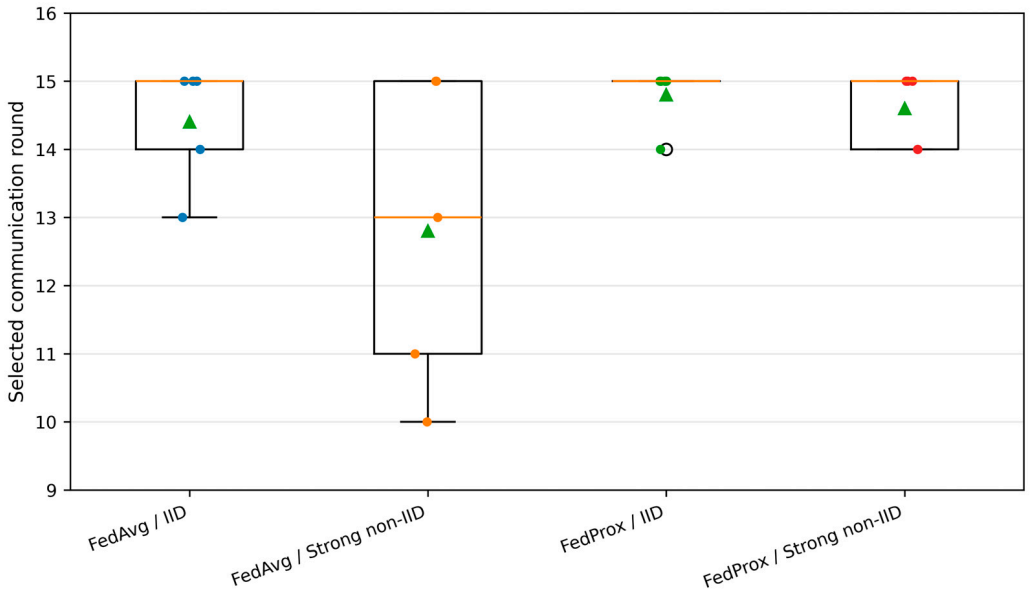

Figure 4. Calibration-selected federated checkpoint rounds across five seeds for FedAvg and FedProx under IID and strong non-IID client partitions. The selected checkpoints generally occurred in later rounds within the 15-round budget, indicating that additional communication remained beneficial under the protocol.

6.4. Communication Cost and Deployment-Oriented Interpretation

Communication cost is a central practical constraint in federated edge-security deployments. In the analysis, communication was counted bidirectionally. For each round, the server distributes the current global model to the clients, corresponding to downlink traffic, and the clients return updated models to the server, corresponding to uplink traffic. Reporting only upload traffic would underestimate the model-transfer cost.

The lightweight MLP used in the federated experiments had a serialized size of 73,732 bytes, approximately 72 KiB. With five clients, the uplink cost per round was 368,660 bytes and the downlink cost per round was also 368,660 bytes. Therefore, the total bidirectional communication cost was 737,320 bytes per round, approximately 720 KiB. Over a 15-round budget, this corresponds to 11,059,800 bytes, approximately 11.06 MB of total model-transfer traffic.

These communication requirements are modest for the compact MLP considered in this study, supporting its suitability as a lightweight edge-oriented baseline (Table 4). Nevertheless, the linear dependence on the number of clients and communication rounds remains important. Larger neural architectures, higher client counts, more frequent synchronization, or unreliable wireless links would increase the practical cost of deployment. Therefore, communication-aware evaluation should remain part of federated IDS assessment even when the model used in the initial study is compact.

Table 4. Communication cost of the federated MLP configuration with five clients.

Quantity	Value
Serialized model size	73,732 bytes
Number of clients	5
Uplink per round	368,660 bytes
Downlink per round	368,660 bytes
Total bidirectional cost per round	737,320 bytes
Total bidirectional cost over 15 rounds	11,059,800 bytes

Figure 5 reports cumulative bidirectional communication cost, showing the linear growth of total communication cost with the number of rounds and should preferably annotate the 15-round operating point used in the final experiments.

6.5. Practical Implications and Limitations

From a practical standpoint, the most important finding of this study is that lightweight federated learning remains competitive with centralized and local-only baselines under a substantially stricter evaluation regime than the conventional random split. The use of group-based scenario separation, explicit client heterogeneity, independent threshold calibration, validation-based checkpoint selection, and multi-seed reporting provides a more deployment-relevant assessment of federated IDS feasibility.

The results also emphasize the importance of cautious interpretation. FedAvg achieved the highest mean F1-score, particularly under strong non-IID partitioning, but the differences relative to the strongest centralized and local-only baselines were modest and generally not statistically decisive. Therefore, the evidence supports the viability and competitiveness of federated learning under privacy-aware edge constraints, rather than proving that federated learning is inherently superior to centralized learning.

The threshold-calibration results further show that IDS operating points must be selected carefully. The default threshold of 0.5 is not guaranteed to be optimal, particularly in imbalanced and scenario-shifted settings. At the same time, threshold tuning must be performed on a calibration or validation set rather than on the test set. The protocol

therefore provides a more realistic estimate of deployable performance, even though the final test performance remains substantially lower than optimistic random-split results.

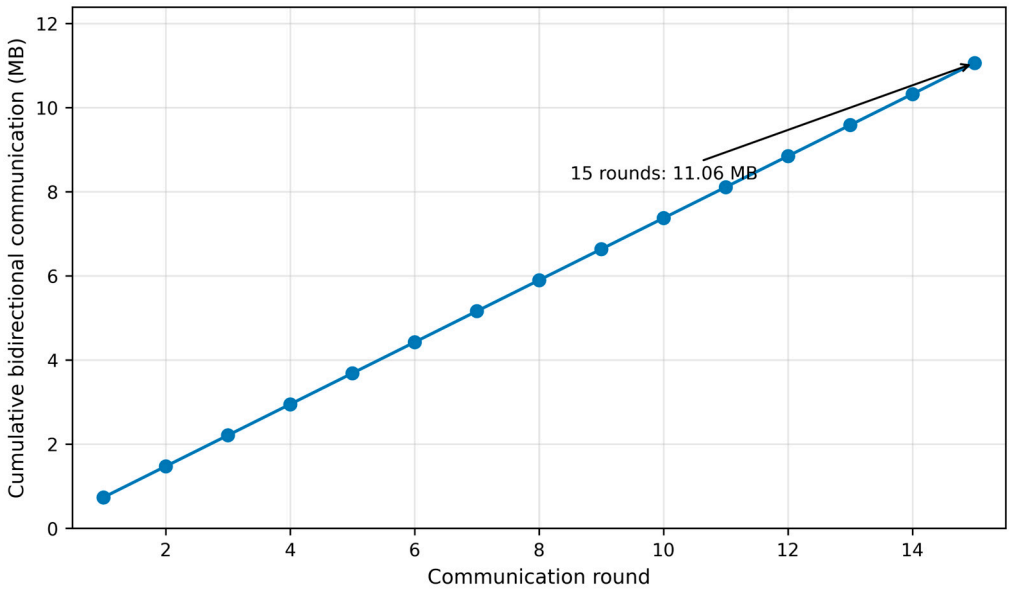

Figure 5. Cumulative bidirectional communication cost as a function of the communication-round budget for the federated MLP configuration with five clients. The cost includes both client-to-server uplink and server-to-client downlink model-transfer traffic.

The communication analysis shows that a compact MLP can be trained federatively with modest bidirectional model-transfer cost. However, this finding should not be generalized to larger architectures without additional analysis. CNNs, recurrent models, Transformers, and graph neural networks may provide richer representations in some IDS settings, especially when temporal or relational structure is available, but they would also increase computational and communication requirements. The present study intentionally uses a lightweight model to isolate the effects of scenario-aware evaluation, client heterogeneity, calibration, and communication accounting.

Several limitations remain. First, CICIDS2017 is a general network intrusion detection benchmark and does not contain native vehicular protocol traffic such as DSRC, C-V2X, CAM/DENM, or cooperative-perception messages. Therefore, the study should be interpreted as V2X-inspired rather than as direct V2X validation. Second, the threat model assumes honest federated participants and does not address malicious clients, poisoning attacks, Byzantine updates, or model inversion. Third, the federated protocol is synchronous and uses a fixed number of clients, whereas real vehicular and edge deployments may involve intermittent connectivity, partial participation, mobility, and variable client availability. Fourth, the multi-class analysis remains exploratory because strict scenario-aware splitting can produce unseen attack categories in the test set.

Globally, the results support federated learning as a viable direction for privacy-aware intrusion detection in distributed edge-security settings, provided that evaluation is aligned with realistic operating constraints. The main deployment-oriented recommendations are: (i) use scenario-aware data partitioning whenever possible; (ii) calibrate thresholds on validation data rather than on the test set; (iii) select federated checkpoints using independent validation criteria; (iv) report variability across multiple random seeds; and (v) account for both uplink and downlink communication when discussing deployment feasibility. These recommendations are modest in scope but directly actionable, and they provide a stronger methodological contribution than an optimistic benchmark comparison alone.

7. Discussion

7.1. Interpretation of the Main Findings

The findings of this study should be interpreted primarily from a methodological and deployment-oriented perspective. The objective was not to demonstrate a protocol-specific V2X intrusion detection system, but to investigate whether lightweight federated intrusion detection remains viable when evaluated under stricter conditions involving scenario shift, client heterogeneity, independent calibration, statistical repetition, and explicit communication accounting.

A central finding is that the evaluation protocol strongly affects the apparent performance of intrusion detection systems. Conventional random partitioning can produce highly optimistic estimates because related traffic flows and scenario-specific artifacts may be represented in both training and testing subsets. In contrast, the group-based protocol adopted in this study separates traffic captures across training, calibration, and testing, thereby creating a substantially more challenging generalization problem. The resulting performance reduction across all evaluated methods indicates that scenario-aware partitioning provides a more conservative assessment of how models may behave when transferred to previously unseen traffic contexts.

This observation has broader implications for intrusion detection research. High performance under randomly partitioned benchmark data does not necessarily imply robustness under temporal, geographical, or operational distribution shifts. In distributed V2X-inspired environments, models trained using traffic collected at some monitoring locations or operational periods may need to generalize to substantially different conditions. Consequently, the data-partitioning protocol should be regarded as a fundamental component of IDS evaluation rather than merely as an implementation detail.

Under this stricter protocol, lightweight federated learning remained competitive with centralized and local-only alternatives. FedAvg under the strong non-IID client partition achieved the highest mean F1-score of 0.468 ± 0.028, followed closely by FedAvg under approximately IID partitioning with 0.463 ± 0.036. The centralized MLP, logistic regression, random forest, and local-only baselines remained within a similar performance range. These differences should therefore be interpreted cautiously: the results support the feasibility and competitiveness of federated learning under privacy-aware distributed constraints, but they do not establish universal superiority over centralized learning.

The comparison between FedAvg and FedProx provides an additional practical insight. Although FedProx was included specifically because it was designed to mitigate optimization difficulties under heterogeneous client data, it did not outperform FedAvg in the final experimental configuration. This result does not imply that proximal regularization is generally ineffective. Rather, it illustrates that its benefit depends on the model architecture, local training schedule, client partition, proximal coefficient, and degree of heterogeneity. Federated optimization methods should therefore be evaluated empirically under the actual conditions of the target IDS scenario instead of being assumed to provide systematic improvements solely because they were designed for non-IID data.

7.2. Calibration, Checkpoint Selection, and Communication Implications

Threshold calibration emerged as another important methodological factor. The decision thresholds were selected using an independent calibration set rather than the final test set, thereby avoiding optimistic operating-point selection. The results confirm that the conventional threshold of 0.5 is not necessarily appropriate for imbalanced intrusion detection under scenario shift. At the same time, the selected thresholds varied across models, random seeds, and client partitions. The present experiments therefore do not support a universal threshold recommendation.

From an operational perspective, threshold selection should instead be treated as a validation-driven design choice. The preferred operating point will depend on the acceptable trade-off among false positives, missed attacks, analyst workload, downstream mitigation costs, and application-specific safety requirements. This is especially relevant for V2X-inspired systems, where the consequences of excessive false alarms may differ substantially from those of missed malicious activity.

The checkpoint-selection results provide a related insight into the role of communication rounds. Under the calibration-based protocol, the best federated checkpoints did not consistently occur during the first few rounds. Validation-selected checkpoints generally emerged later within the tested 15-round budget, indicating that additional communication remained useful in the evaluated setting. However, each additional round also increased communication volume. The practical implication is therefore not that federated IDS training should simply stop early, but that checkpoint selection should be driven by independent validation while being considered jointly with the available communication budget.

The communication analysis further showed that the lightweight MLP is compatible with communication-aware edge experimentation. The serialized model size was approximately 73,732 bytes, corresponding to approximately 737,320 bytes of bidirectional model-transfer traffic per round for five clients when both uplink and downlink are counted. Across the 15-round experimental budget, the cumulative communication volume was approximately 11.06 MB. This overhead is modest for the compact model considered here, but it scales linearly with the number of clients, communication rounds, and model size. Communication cost should therefore remain an explicit evaluation dimension when moving toward larger client populations or more complex neural architectures.

7.3. Limitations and Future Research

Several limitations constrain the scope of the present conclusions. First, CICIDS2017 is a general network intrusion detection benchmark and does not contain native V2X protocol traffic, including DSRC, C-V2X, CAM/DENM, cooperative perception messages, or vehicular application-layer semantics. The experiments should therefore be interpreted as a controlled study of V2X-inspired distributed edge-security learning rather than as direct validation of a production-ready V2X IDS.

Second, the threat model assumes honest federated participants. The study does not evaluate malicious clients, data or model poisoning, Byzantine updates, model inversion, inference leakage, or secure aggregation. These threats are particularly important in realistic distributed environments because compromised RSUs, gateways, or edge nodes may directly target the collaborative learning process.

Third, the federated experiments use a fixed synchronous configuration with five clients. Real deployments may involve larger client populations, partial participation, intermittent connectivity, mobility, asynchronous updates, and time-varying client availability. The communication measurements reported here should therefore be interpreted as controlled first-order indicators rather than as measurements from a complete networked deployment.

Fourth, the main conclusions concern binary intrusion detection. The coarse multi-class evaluation remains exploratory because strict scenario-separated partitioning may create unseen or underrepresented attack categories in the held-out test set. More robust multi-class and open-set protocols are needed before stronger claims can be made regarding attack-category identification under scenario shift.

Future work should address these limitations by extending the evaluation to native V2X and vehicular-edge datasets with protocol-aware traffic features and attack semantics. The binary task should also be expanded toward multi-class and open-set intrusion detec-

tion with explicit treatment of previously unseen attack categories. Additional federated optimization, aggregation, and personalization strategies—including FedProx variants, SCAFFOLD, FedNova, and robust aggregation methods—should be evaluated across larger and more heterogeneous client populations. Adversarial federated-learning threats, including compromised clients, poisoning, Byzantine behavior, and model-inference risks, should also be incorporated. Finally, real or emulated deployments involving RSUs, vehicular gateways, or fog nodes are necessary to characterize end-to-end latency, communication reliability, computational constraints, and operational feasibility more realistically.

8. Conclusions

This study investigated lightweight federated binary intrusion detection under a deployment-oriented evaluation protocol designed to account for scenario shift, client heterogeneity, independent threshold calibration, validation-based checkpoint selection, statistical variability, and communication cost. Using CICIDS2017 as a controlled proxy benchmark, the results showed that scenario-aware group-based partitioning produces a substantially more challenging and conservative assessment than conventional random splitting.

Under this stricter protocol, federated learning remained competitive with centralized and local-only alternatives. FedAvg achieved the highest mean F1-score in the strong non-IID setting (0.468 ± 0.028), closely followed by its approximately IID configuration (0.463 ± 0.036). These results support the feasibility of lightweight federated intrusion detection while also showing that its advantage over centralized learning is not universal. Independent threshold calibration and validation-based checkpoint selection materially affected the evaluated operating points, while the compact MLP required approximately 11.06 MB of cumulative bidirectional model-transfer traffic over the tested 15-round, five-client configuration.

Globally, the study shows that the credibility of federated IDS evaluation depends not only on predictive performance, but also on realistic data partitioning, unbiased calibration, multi-seed statistical reporting, and explicit communication accounting. The resulting framework provides a reproducible methodological baseline for future research toward native V2X validation, adversarially robust federated learning, larger and more dynamic client populations, and real or emulated edge deployments.

Funding: This research received no external funding.

Data Availability Statement: The dataset used in this study is publicly available from its official source, namely the CICIDS2017 intrusion detection benchmark. The implementation scripts, result tables, CSV summary files, figures, and the reproducibility package used to generate the reported results are publicly available at https://github.com/manuelcabralreis/federated-v2x-intrusion-detection (accessed on 9 July 2026) and archived on Zenodo at https://doi.org/10.5281/zenodo.19422283.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. El-Rewini, Z.; Sadatsharan, K.; Selvaraj, D.F.; Plathottam, S.J.; Ranganathan, P. Cybersecurity Challenges in Vehicular Communications. *Veh. Commun.* **2020**, *23*, 100214. [CrossRef]
2. Zhang, W.; Gao, Y.; Jiang, Z.; Mao, R.; Zhou, S. Calibration-Free Roadside BEV Perception with V2X-Enabled Vehicle Position Assistance. *Sensors* **2025**, *25*, 3919. [CrossRef] [PubMed]
3. Alnasser, A.; Sun, H.; Jiang, J. Cyber Security Challenges and Solutions for V2X Communications: A Survey. *Comput. Netw.* **2019**, *151*, 52–67. [CrossRef]
4. Tsimenidis, S.; Lagkas, T.; Rantos, K. Deep Learning in IoT Intrusion Detection. *J. Netw. Syst. Manag.* **2021**, *30*, 8. [CrossRef]
5. Ferrag, M.A.; Maglaras, L.; Moschoyiannis, S.; Janicke, H. Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study. *J. Inf. Secur. Appl.* **2020**, *50*, 102419. [CrossRef]

6. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A. Y Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*; PMLR: Cambridge, MA, USA, 2017; pp. 1273–1282.
7. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and Open Problems in Federated Learning. *Found. Trends Mach. Learn.* **2021**, *14*, 1–210. [CrossRef]
8. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated Machine Learning: Concept and Applications. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–19. [CrossRef]
9. Lazzarini, R.; Tianfield, H.; Charissis, V. Federated Learning for IoT Intrusion Detection. *AI* **2023**, *4*, 509–530. [CrossRef]
10. Hamdi, N. Federated Learning-Based Intrusion Detection System for Internet of Things. *Int. J. Inf. Secur.* **2023**, *22*, 1937–1948. [CrossRef]
11. Rashid, M.M.; Khan, S.U.; Eusufzai, F.; Redwan, M.A.; Sabuj, S.R.; Elsharief, M. A Federated Learning-Based Approach for Improving Intrusion Detection in Industrial Internet of Things Networks. *Network* **2023**, *3*, 158–179. [CrossRef]
12. Amiri-Zarandi, M.; Dara, R.A.; Lin, X. SIDS: A Federated Learning Approach for Intrusion Detection in IoT Using Social Internet of Things. *Comput. Netw.* **2023**, *236*, 110005. [CrossRef]
13. Agrawal, S.; Sarkar, S.; Aouedi, O.; Yenduri, G.; Piamrat, K.; Alazab, M.; Bhattacharya, S.; Maddikunta, P.K.R.; Gadekallu, T.R. Federated Learning for Intrusion Detection System: Concepts, Challenges and Future Directions. *Comput. Commun.* **2022**, *195*, 346–361. [CrossRef]
14. Heidari, A.; Khonsari, A.; Rastegar, S.H. An Energy-Efficient Privacy-Preserving Framework for Intrusion Detection in the Internet of Vehicles. *Comput. Electr. Eng.* **2026**, *132*, 111003. [CrossRef]
15. Wang, X.; Geng, Y.; Luo, X.; Guan, X. Detection of Dummy Data Injection Attacks by Using Particle Swarm Optimization-Attention Temporal Graph Convolutional Network Model in Power System. *Eng. Appl. Artif. Intell.* **2026**, *171*, 114259. [CrossRef]
16. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated Optimization in Heterogeneous Networks. *Proc. Mach. Learn. Syst.* **2020**, *2*, 429–450.
17. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; Suresh, A.T. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the 37th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2020; pp. 5132–5143.
18. Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; Poor, H.V. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*; NeurIPS Proceedings: La Jolla, CA, USA, 2020; pp. 7611–7623.
19. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018)*; Science and Technology Publications: Setúbal, Portugal, 2018; pp. 108–116.