An Optimized Federated Resource-Efficient IoT Security Framework with Uncertainty-Aware Distillation Loss

Ogobuchi Daniel Okey[1], Demóstenes Zegarra Rodríguez[2], João Henrique Kleinschmidt[1]

[1]Center for Engineering, Modeling and Applied Social Sciences (CECS), Federal University of ABC (UFABC), 09280-380, Santó Andre, SP – Brazil

[2]Department of Computer Science (DCC) – Federal University of Lavras (UFLA) 37200-000 – Lavras – MG – Brazil

{okey.ogobuchi,joao.kleinschmidt}@ufabc.edu.br, demostenes.zegarra@ufla.br

Abstract. *This paper presents FedREUAD, a federated intrusion detection framework that jointly addresses client drift, communication overhead, and asymmetric knowledge transfer in distillation-based aggregation. Local model divergence under non-IID data is resolved through the FedProx's proximal term that constrains local updates, limiting accuracy degradation on Dirichlet-simulated heterogeneity scenarios on CICIoT2023 and CICIoMT2024 datasets. During knowledge distillation, FedREUAD resolves the inherent asymmetry of Kullback-Leibler divergence (KL), which asymmetrically weights teacher-student discrepancies and propagates overconfident soft labels, by adopting a symmetric Jensen-Shannon divergence loss weighted by per-sample predictive uncertainty, penalizing the transfer of poorly calibrated soft labels from local teachers, thus reducing expected calibration error by* 31% *and* 28% *on the two datasets, respectively. Resource efficiency is achieved through distillation, structured magnitude pruning, and INT8 quantization, compressing local models by* 90% *and reducing per-round communication from* 14.2 *MB to* 4.5 *MB. The proposed method attains macro F1-scores of* 98.22% *on CICIoT2023 and* 99.05% *on CICIoMT2024, outperforming federated baselines under identical heterogeneity conditions, demonstrating that symmetric, uncertainty-calibrated distillation enables robust knowledge transfer without the calibration degradation inherent to asymmetric divergence formulations.*

Index Terms - *Federated Learning, Intrusion Detection System, Internet of Things, Knowledge Distillation, Jensen–Shannon Divergence, Resource Efficiency, Non-IID, CICIoT2023, CICIoMT2024.*

1. Introduction

The global installed base of Internet of Things (IoT) devices surpassed 30 billion units by 2025 and is projected to exceed ≈41 billion by 2030 [Neto et al. 2023, Statista Research Department 2025], with smart homes, industrial platforms, health systems, and city-scale sensing environments now relying on large numbers of interconnected devices that continuously exchange sensitive data and control signals. This scale and diversity have been accompanied by a corresponding surge

in cyberattacks targeting these network endpoints, thus making timely intrusion detection a core security requirement rather than an optional add-on. This intrusion detection system (IDS) serves as a primary software-level line of defense by monitoring network flows for anomalous or malicious behaviour. However, the architecture of traditional IDS is fundamentally misaligned with the distributed, heterogeneous, and privacy-sensitive nature of modern IoT deployments [Lavaur et al. 2022, Mothukuri et al. 2022].

With the limitations of centralized IDS approaches, such as data centralization, cost-effective data acquisition, data sensitivity, privacy concerns, and the communication bandwidth required to continuously stream high-rate traffic data from thousands of distributed endpoints to a central location, centralized approaches are operationally infeasible in many network configurations. Federated learning (FL) [McMahan et al. 2017, Rahman et al. 2020] decouples model training from data centralization by distributing both the computation and the data across the clients that generate it, thus mitigating most of the limitations. Advances in the literature have demonstrated that decentralized training can approach the utility of centralized learning while preserving data locality [Mothukuri et al. 2022, Shen et al. 2024].

Despite promising empirical results, existing FL-based IDS frameworks face at least three fundamental limitations that motivate the present work, including (i) data heterogeneity and client drift, (ii) communication and computational overhead, and (iii) knowledge distillation (KD) under distributional shift. The traffic generated by medical sensors, smart home appliances, industrial controllers, and edge gateways differs sharply in protocol mix, attack prevalence, and feature statistics. Under non-IID conditions, local training objectives drift apart, global convergence weakens, and minority attack classes become hard to learn reliably, as evidenced in the literature [Li et al. 2020, Karimireddy et al. 2020]. Considering (ii), the per-round cost of transmitting full model updates can be prohibitive for edge devices with limited uplink bandwidth or duty-cycle budgets. Similarly, training a full-precision model on IoT device is infeasible without architectural adaptation. The recent literature has addressed parts of this problem through pruning, quantization, lightweight architectures, and communication-efficient update rules [Prakash et al. 2022].

Knowledge distillation has become a useful tool at this point of tension between heterogeneity and efficiency, where a student model is trained to mimic the output distribution of a teacher, preserving soft-label information beyond the hard class label. In broader FL, distillation reduces communication burden by transferring predictive behaviour rather than full parameter states and also supports collaboration among heterogeneous models [Itahara et al. 2021]. In IoT intrusion detection, several recent methods have shown that distillation can improve detection performance, support semi-supervised learning, and produce lighter deployment models [Okey et al. 2025]. However, most existing approaches still treat teacher (more knowledgeable model) predictions as uniformly reliable. Again, conventional KD objectives such as the KL divergence are asymmetric, i.e. it penalizes a student that places probability mass where the teacher does not, but not vice versa. Specifically, when local models trained on highly skewed data generate overconfident but incorrect soft labels, KL-based distillation amplifies those errors. An overconfident

teacher can pass unstable or misleading soft targets to the student, especially under strong heterogeneity. A symmetric, bounded alternative is therefore desirable.

This reliability gap motivates uncertainty-aware learning, which has been shown to improve trustworthiness in intrusion detection by exposing overconfident errors and handling ambiguous samples [Talpini et al. 2025], and to guide aggregation and knowledge transfer more carefully than confidence-free objectives in federated and distillation settings [Bhatt et al. 2024]. However, current federated IDS frameworks typically optimize for privacy, accuracy, or efficiency without adequately addressing how uncertainty should shape knowledge transfer across clients. To bridge this gap, we propose a federated resource-efficient IoT security framework that unifies three components in one training loop: client drift mitigation under heterogeneous traffic [Li et al. 2020], Jensen-Shannon divergence for stable symmetric KD, and uncertainty-weighted distillation loss that downweights unreliable teacher guidance, preserving knowledge transfer benefits while preventing uncertain predictions from dominating model compression and federated adaptation. The contributions of this paper are summarized as follows:

1. We propose FedREUAD, a unified federated security framework that couples heterogeneity-aware optimization and lightweight deployment with uncertainty-aware knowledge transfer.
2. We formulate resource efficiency constraints as a joint pruning-quantization pipeline embedded within the FL training loop, with an explicit analysis of communication savings.
3. We introduce an evaluation of Jensen-Shannon distillation loss, addressing both client drift and soft-label miscalibration through symmetric divergence.

Section 2 surveys related work. Section 3 presents the problem formulation, mathematical framework, and algorithm. Section 4 discusses experimental results and ablation studies. Section 5 concludes with directions for future research.

2. Related Literature

Several studies have demonstrated the effectiveness of FL methods in intrusion detection, however, practical deployment remains constrained by non-IID data distributions, limited communication budgets, susceptibility to adversarial clients, and unreliable confidence estimates [Lavaur et al. 2022]. We examine the literature in the light of: (1) baseline federated IDS frameworks, (2) knowledge-distillation-based federated IDS, (3) resource-efficient federated training, and (4) uncertainty- or calibration-aware security learning. The authors in [Rahman et al. 2020] rigorously compared centralized, on-device, and FL using the N-BaIoT dataset, demonstrating that FL achieves accuracy comparable to centralized training while substantially reducing raw data transmission and degrades under increasing label imbalance. Mothukuri et al. proposed an FL-based anomaly detection scheme combining FedAvg with LSTM networks, reporting F1-scores exceeding 0.97 on the TON-IoT and N-BaIoT datasets [Mothukuri et al. 2022]; however, their assumption of IID data partitions limits real-world generalizability. Friha et al., on the other hand, introduced FELIDS, achieving over 99% detection accuracy on the Bot-IoT dataset [Friha et al. 2022], yet without addressing communication efficiency or heterogeneous data distributions. A lightweight FL-IDS was proposed

in [Okey et al. 2025], converging to near-centralized accuracy on CICIoMT2024, FLNET2023, and ACIIoT2023.

To address the challenges of statistical heterogeneity (non-IID data) and system heterogeneity in FL, several advanced algorithms have been proposed. In [Li et al. 2020] FedProx was introduced, which augments the local objective function with a proximal term $\frac{\mu}{2}\|\mathbf{w}-\mathbf{w}^t\|^2$, where $\mathbf{w}^t$ denotes the global model and $\mu \geq 0$ is a tunable regularization coefficient. This term effectively mitigates client drift and provides convergence guaranties under non-IID and heterogeneous conditions where standard FedAvg falls short; however, performance is sensitive to the choice of μ. Karimireddy et al. proposed SCAFFOLD, which employs control variates to correct local gradient estimates and reduce variance caused by data heterogeneity, theoretically achieving faster convergence than FedProx without requiring a global learning rate [Karimireddy et al. 2020]. Its main drawback is the doubled communication overhead, as clients must store and transmit additional control variate vectors, making it costly for resource-constrained IoT devices. Li et al. developed MOON (Model-contrastive FL), which uses a contrastive objective to align local representations with the global model while distancing them from previous local models [Li et al. 2021].

The foundational work of [Hinton et al. 2015] on KD, in which a compact student model is trained to minimize the KL divergence between its softened softmax outputs and those of a larger teacher network, effectively transferring rich inter-class similarity information, has led to its wide application [Okey et al. 2025, Shen et al. 2024]. In [Okey et al. 2025], CAFiKS, was proposed as a distillation method that aggregates client predictions on a public unlabeled dataset at the server and trains a global model via KL minimization, demonstrating improved convergence under model heterogeneity and non-IID data. Similarly, FLEKD in [Shen et al. 2024] showed that a multi-teacher ensemble KD where client models act as teachers weighted by local validation accuracy, can achieve superior detection performance over FedAvg and FedProx. Collectively, these distillation-based methods provide effective strategies for model compression and knowledge transfer in federated IoT environments.

Model compression and resource-aware optimization are critical enablers for deploying FL on resource-constrained IoT devices. The authors in [Han et al. 2024] provided a comprehensive survey of compression techniques, showing that structured pruning combined with INT8 post-training quantization consistently achieves a reduction in model size $4\times$ to $8\times$ while incurring a loss of accuracy of less than 2% in standard benchmarks and explicitly identified communication overhead in FL as the primary motivation for such methods. On the other hand, [Prakash et al. 2022] showed that joint pruning and quantization can make FL substantially faster and more communication-efficient on IoT devices. The compelling work of [Hinton et al. 2015] pushed the narrative for KD as a compression technique, demonstrating that a compact student network can closely approximate the performance of a larger teacher model with minimal accuracy degradation. [Liu et al. 2023] showed for AIoT applications that soft targets can improve FL accuracy without incurring the full communication cost of richer model exchange

while [Itahara et al. 2021] demonstrated that distillation-based semi-supervised FL can sharply reduce communication under non-IID data.

Together, these contributions underscore the importance of integrating model compression and heterogeneity-aware strategies to make federated IDS practical for large-scale IoT deployments. However, within the literature, a combination of these three compression techniques has been overlooked. More so, most federated IDS yet optimize accuracy more than confidence quality. For example, [Talpini et al. 2024] showed that uncertainty quantification is essential for trustworthy intrusion detection because overconfident predictions obscure unknown attacks and mislabeled samples. Their later federated calibration work [Talpini et al. 2025] showed that distributed confidence correction can improve reliability without sacrificing privacy. In the broader federated and distillation literature, [Bhatt et al. 2024] demonstrated that predictive uncertainty can be distilled and used to steer model transfer more carefully. The unresolved gap is therefore not whether federated IDS can be accurate, but whether it can remain accurate, communication-efficient, and reliably calibrated under realistic non-IID IoT conditions. Table 1 summarizes the papers relevant to the proposed framework using the following annotations: **FL**: uses federated learning; **non-IID**: explicitly handles non-IID data; **KD**: employs knowledge distillation; **RE**: addresses resource efficiency; **UA**: uses uncertainty-aware objective; **Modern DS**: uses post-2022 dataset; **IoMT**: evaluates on IoMT data.

Table 1. Comparison of related works across key framework dimensions.

Work	FL	non-IID	KD	RE	UA	Modern DS	IoMT
[Rahman et al. 2020]	✓	×	×	×	×	×	×
[Mothukuri et al. 2022]	✓	×	×	×	×	×	×
[Friha et al. 2022]	✓	×	×	×	×	×	×
[Li et al. 2020]	✓	✓	×	×	×	×	×
[Karimireddy et al. 2020]	✓	✓	×	×	×	×	×
[Li et al. 2021]	✓	✓	×	×	×	×	×
[Liu et al. 2022]	✓	×	×	✓	×	×	×
[Shen et al. 2024]	✓	✓	✓	×	×	✓	×
[Okey et al. 2025]	✓	✓	✓	✓	×	✓	✓
FedREUAD (ours)	✓	✓	✓	✓	✓	✓	✓

3. FedREUAD Design and Implementation Pipeline

This section discusses the architectural and design principles of the proposed technique, highlighting the problem domain, limitations of the existing method, improvement and the functionality protocol. The implementation algorithm for the FedREUAD framework is analyzed in Algorithm 1.

3.1. System Model and Problem Formulation

For a typical FL system applied to IoT intrusion detection, a central server $\mathcal{S}$ coordinates the training process across N distributed clients indexed by $\mathcal{K} =$

$\{1, 2, \ldots, N\}$. Each client $k \in \mathcal{K}$ possesses a local dataset $\mathcal{D}_k = \{(\mathbf{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k}$, where $\mathbf{x}_i^{(k)} \in \mathbb{R}^d$ denotes a d-dimensional feature vector extracted from network traffic and $y_i^{(k)} \in \mathcal{Y} = \{0, 1, \ldots, C-1\}$ represents the corresponding class label (benign or one of $C-1$ attack types). Let $n = \sum_{k=1}^{N} n_k$ denote the total number of samples across all clients. The empirical data distribution of client k is given by $\hat{P}_k(\mathbf{x}, y) = \frac{1}{n_k} \sum_{i=1}^{n_k} \delta(\mathbf{x} - \mathbf{x}_i^{(k)}) \delta(y - y_i^{(k)})$. Notably, the local distributions $\mathcal{D}_k$ are not identical because device roles, communication patterns, and attack exposure differ across clients. This causes client drift, weak global convergence, and unstable soft targets for distillation. The proposed framework addresses these issues by coupling four mechanisms: a FedProx local objective, uncertainty-aware distillation based on Jensen-Shannon divergence, compressed communication, and lightweight student models. The non-IID condition is simulated through a class-wise Dirichlet partition $\mathrm{Dir}(\alpha)$ with a concentration parameter $\alpha = 0.5$. For each class $c \in \{1, \ldots, C\}$, the class proportions assigned to the K clients are sampled as $(\pi_{1,c}, \ldots, \pi_{K,c}) \sim \mathrm{Dir}_K(\alpha \mathbf{1}), \quad \alpha = 0.5$. The samples of class c are then distributed according to these proportions. This setting yields moderate to strong heterogeneity and is consistent with the intended stress test for federated IDS training. The objective (see Eq. 1) is to learn a global detector that remains accurate under heterogeneous client data, light enough for resource-constrained devices, and does not become overconfident when transferred through knowledge distillation.

$$\min_{\mathbf{w} \in \mathbb{R}^p} F(\mathbf{w}) = \sum_{k=1}^{N} \frac{n_k}{n} f_k(\mathbf{w}), \tag{1}$$

where $f_k(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\hat{P}_k}[\ell(\mathbf{w}; \mathbf{x}, y)]$ is the local risk on the client, k and $\ell(\cdot)$ is the cross-entropy loss. The statistical heterogeneity between clients is quantified by the gradient dissimilarity at the global minimizer, $\mathbf{w}^* = \mathbb{E}\left[\|\nabla f_k(\mathbf{w}^*)\|^2\right] \leq \sigma^2$. When $\sigma^2 > 0$ (the non-IID case), standard FedAvg [McMahan et al. 2017] suffers from client drift, accumulating a bias proportional to σ^2/μ per communication round, which motivates the use of proximal regularization techniques such as FedProx [Li et al. 2020] to stabilize convergence.

3.2. Uncertainty-Aware Distillation Loss (UDL)

Classical KD trains the student using the KL divergence between the teacher's softened output $\mathbf{q}^T$ and the student's output $\mathbf{q}^S$, according to Eq. 2, where τ is the temperature. One downside of this method is that the asymmetric nature of KL treats all the teachers' predictions equally (penalizes the student for underweighting classes that the teacher weights), which may result in an underperforming student model. Our UDL modifies the standard teacher–student training by weighting the teacher's guidance according to its confidence. Instead of treating all teacher predictions equally, the loss emphasizes reliable outputs and downweights uncertain ones, leading to more robust student models.

$$\mathcal{L}_{\mathrm{KD}}^{\mathrm{KL}} \;=\; \tau^2 \, D_{\mathrm{KL}}(\mathbf{q}^T \| \mathbf{q}^S) \;=\; \tau^2 \sum_{c=0}^{C-1} q_c^T \log \frac{q_c^T}{q_c^S} \tag{2}$$

In the federated setting, local models trained on skewed data can be highly overconfident in classes they have seen, generating soft labels with near-unity probabilities. Minimizing $D_{\mathrm{KL}}(\mathbf{q}^T \| \mathbf{q}^S)$ forces the student to match these erroneous

peaks, transferring bad calibration. Thus, we leverage Jensen-Shannon divergence (JSD) [Fuglede and Topsoe 2004] that resolves this asymmetry. For two distributions P and Q, it is defined as, JSD is defined as in Eq. 3

$$D_{\mathrm{JS}}(P\|Q) \;=\; \frac{1}{2}D_{\mathrm{KL}}(P\|M) + \frac{1}{2}D_{\mathrm{KL}}(Q\|M), \quad M = \frac{P+Q}{2}. \tag{3}$$

D_{JS} is symmetric, bounded in $[0, \log 2]$, and its square root is a proper metric. These properties make it a theoretically superior distillation objective under distributional uncertainty.

3.2.1. Predictive Uncertainty Estimation

We estimate the predictive uncertainty of the client model k via the entropy of its softmax output. Given that $q_{k,c}(\mathbf{x}) = \mathrm{Softmax}(f_{\mathbf{w}_k}(\mathbf{x}))_c$ is the predicted probability of class c by client k's model the PUE (predictive uncertainty estimation) is obtained according to Eq. 4.

$$H_k(\mathbf{x}) \;=\; -\sum_{c=0}^{C-1} q_{k,c}(\mathbf{x}) \log q_{k,c}(\mathbf{x}), \tag{4}$$

High entropy indicates low confidence; low entropy indicates high confidence. We define a per-sample weight that down-weights high-confidence predictions, thereby preventing overconfident teachers from dominating (see Eq. 5), so that $\omega_k(\mathbf{x}) \to 0$ as $H_k(\mathbf{x}) \to 0$ (high confidence, reduced weight) and $\omega_k(\mathbf{x}) \to 1 - e^{-\beta \log C}$ as the distribution approaches uniform (maximum uncertainty). The parameter β controls the steepness of this mapping.

$$\omega_k(\mathbf{x}) \;=\; 1 - \exp(-\beta\, H_k(\mathbf{x}))\,, \quad \beta > 0, \tag{5}$$

3.2.2. Uncertainty-Aware JS Distillation Objective

Let $\mathbf{q}^G = \mathrm{Softmax}(f_{\mathbf{w}^t}(\mathbf{x})/\tau)$ denote the global model's softened output and $\mathbf{q}_k = \mathrm{Softmax}(f_{\mathbf{w}_k}(\mathbf{x})/\tau)$ the client's output. The uncertainty-aware distillation (UAD) loss for client k on the sample $\mathbf{x}$ is defined by Eq. 6. For a mini-batch $\mathcal{B}_k$, $\mathcal{L}^{(k)}_{\mathrm{UKD}}(\mathbf{x})$ becomes Eq. 7, and expanding D_{JS} using (3), we obtain Eq. 8, with the factor τ^2 compensating for the variance reduction caused by temperature scaling.

$$\mathcal{L}^{(k)}_{\mathrm{UAD}}(\mathbf{x}) \;=\; \tau^2\, \omega_k(\mathbf{x}) \cdot D_{\mathrm{JS}}\Big(\mathbf{q}^G \| \mathbf{q}_k\Big)\,. \tag{6}$$

$$\mathcal{L}^{(k)}_{\mathrm{UKD}}(\mathbf{x}) \;=\; \frac{\tau^2}{|\mathcal{B}_k|} \sum_{x_i \in \mathcal{B}_k} \omega_k(\mathbf{x}) \cdot D_{\mathrm{JS}}\Big(\mathbf{q}^G \| \mathbf{q}_k\Big) \tag{7}$$

$$\mathcal{L}^{(k)}_{\mathrm{UAD}}(\mathbf{x}) = \frac{\tau^2\, \omega_k(\mathbf{x})}{2} \left[\sum_{c=0}^{C-1} q_c^G \log \frac{2q_c^G}{q_c^G + q_{k,c}} + \sum_{c=0}^{C-1} q_{k,c} \log \frac{2q_{k,c}}{q_c^G + q_{k,c}} \right]. \tag{8}$$

This formulation has three useful properties. First, JSD is symmetric and bounded, making it numerically stable for soft-target transfer. Second, entropy weighting discourages the student from matching highly uncertain teacher outputs too strongly. Third, the loss remains local and does not require raw data exchange, which preserves the privacy advantage of federated learning.

3.3. Aggregation with Proximal Regularization

To further stabilize training under non-IID, FedREUAD adopts FedProx [Li et al. 2020] as the base aggregation protocol. In each round t, the server broadcasts the current global model $\mathbf{w}^t$ to a uniformly sampled subset $\mathcal{S}^t \subseteq \mathcal{K}$ of m clients. Each selected client $k \in \mathcal{S}^t$ solves the proximal subproblem in Eq. 9, where $\mu > 0$ is the proximal coefficient. The proximal term acts as a quadratic penalty that limits the magnitude of the local update, thereby bounding the client drift. After E local epochs of stochastic gradient descent with a learning rate η_l, each client obtains a local model $\mathbf{w}_k^{t+1}$ and transmits it to the server. The server performs weighted aggregation in Eq. 10.

$$\min_{\mathbf{w}_k} \; h_k(\mathbf{w}_k; \mathbf{w}^t) \;=\; \underbrace{f_k(\mathbf{w}_k)}_{\text{local cross-entropy}} \;+\; \underbrace{\frac{\mu}{2}\|\mathbf{w}_k - \mathbf{w}^t\|^2}_{\text{proximal term}} \tag{9}$$

$$\mathbf{w}^{t+1} \;=\; \sum_{k \in \mathcal{S}^t} \frac{n_k}{\sum_{j \in \mathcal{S}^t} n_j}\, \mathbf{w}_k^{t+1}. \tag{10}$$

The complete local objective that each client minimizes is a combination of the FedProx subproblem (9) and the UAD term, using (6), where $\lambda \geq 0$ is a trade-off coefficient that controls the relative weight of the distillation term.

$$\mathcal{L}_k(\mathbf{w}_k) \;=\; \underbrace{f_k(\mathbf{w}_k)}_{\text{CE}} \;+\; \frac{\mu}{2}\|\mathbf{w}_k - \mathbf{w}^t\|^2 \;+\; \lambda\, \underbrace{\frac{1}{n_k}\sum_{i=1}^{n_k} \mathcal{L}_{\text{UAD}}^{(k)}(\mathbf{x}_i^{(k)})}_{\text{UAD}}, \tag{11}$$

When $\lambda = 0$, (11) reduces to the standard FedProx objective. When $\mu = 0$ and $\lambda > 0$, the framework becomes a JS-distillation-only variant without proximal stabilization. The ablation study in Section 4 evaluates both extreme cases and several intermediate configurations.

3.4. Resource Efficiency: Pruning and Quantization

3.4.1. Structured Magnitude-Based Pruning

Before the federated training begins, each client model $f_{\mathbf{w}_k}$ is compressed by removing filters with the lowest ℓ_2-norm in each convolutional layer. Let $\mathbf{W}_l \in \mathbb{R}^{F_l \times C_l \times H \times W}$ denote the weight tensor of the l-th convolutional layer, where F_l is the number of output filters. The importance score of the filter j in the layer l is (12) and the pruning mask retains only filters with scores above the ρ_l-th percentile (13).

$$s_{l,j} \;=\; \|\mathbf{W}_{l,j,:,:,:}\|_2, \tag{12}$$

$$\mathcal{M}_{l,j} \;=\; \begin{cases} 1 & \text{if } s_{l,j} \geq \text{Percentile}(\{s_{l,j'}\}_{j'=1}^{F_l}, 1-\rho_l) \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

The global pruning ratio is $\rho = 1 - \prod_l (1-\rho_l)$. This structured approach produces a smaller, dense architecture that is amenable to hardware acceleration, unlike unstructured sparse pruning.

3.4.2. Post-Training INT8 Quantization

Following structured pruning, the remaining non-zero parameters are further compressed through post-training INT8 quantization to reduce memory footprint, communication overhead, and inference complexity while preserving predictive performance. Let $w \in [\theta_{\min}, \theta_{\max}]$ denote a floating-point weight parameter of the pruned neural network. The quantized representation $\hat{w}$ is computed according to (14), where $b = 8$ represents the quantization bit-width, $\lfloor \cdot \rceil$ denotes rounding to the nearest integer, and Δ defines the quantization scale factor. The clipping operation constrains the quantized values within the representable signed INT8 range $[-2^{b-1}, 2^{b-1}-1]$, thereby preventing numerical overflow during deployment. The $\theta_{\min}$ and $\theta_{\max}$ correspond to the minimum and maximum values observed within the parameter tensor, respectively. The scaling factor Δ linearly partitions the floating-point range into $2^b - 1$ discrete quantization intervals, enabling efficient integer approximation of continuous-valued parameters.

$$\hat{w} \;=\; \Delta \cdot \text{clip}\Big(\Big\lfloor \frac{w}{\Delta} \Big\rceil, -2^{b-1}, 2^{b-1}-1 \Big), \quad \Delta = \frac{\theta_{\max} - \theta_{\min}}{2^b - 1}, \tag{14}$$

3.4.3. Communication Savings

Let p be the total number of parameters before compression. After structured pruning with ratio ρ and INT8 quantization, the number of bits transmitted per client per round is $B_{\text{compressed}} = (1-\rho) \cdot p \cdot 8$, vs. $B_{\text{baseline}} = p \cdot 32.$, where $b_f = 32$ represents the FP32 parameter precision used in standard federated optimization. The overall communication compression factor γ achieved by the proposed framework is therefore expressed as $\gamma = \frac{B_{\text{baseline}}}{B_{\text{compressed}}} = \frac{4}{1-\rho}$. This formulation demonstrates that the communication reduction scales inversely with the proportion of retained parameters. For instance, when the pruning ratio is set to $\rho = 0.5$, only half of the original parameters are transmitted, leading to $\gamma = \frac{4}{1-0.5} = 8$, which corresponds to an $8\times$ reduction in communication cost per federated round relative to conventional FP32 transmission.

Moreover, the proposed compression strategy operates synergistically with KD. While pruning and quantization reduce the size of transmitted model updates, the distillation mechanism further minimizes unnecessary information exchange by transferring compact knowledge representations rather than relying solely on large full-precision parameter synchronization. Consequently, the proposed FedREUAD framework achieves substantial reductions in communication complexity while pre-

serving robust intrusion detection performance under heterogeneous non-IID IoT settings. The general implementation workflow of the proposed FedREUAD framework is summarized in Algorithm 1.

Algorithm 1 FedREUAD Training Algorithm

Require: N clients, datasets $\{\mathcal{D}_k\}_{k=1}^{N}$, pruning ratio ρ, quantization bits b, FedProx coefficient μ, distillation weight λ, temperature τ, uncertainty scale β, local epochs E, local learning rate η_l, communication rounds T, clients per round m

— **Initialization Phase** —

Initialize global model weights $\mathbf{w}^0$ (Xavier initialization)
for all $k \in \mathcal{K}$ **do**
    Apply structured pruning to $f_{\mathbf{w}^0}$ using masks $\{\mathcal{M}_{l,j}\}$ with per-layer ratio ρ_l ▷ Eq. (13)
    Apply INT8 quantization to pruned weights ▷ Eq. (14)
    Partition $\mathcal{D}_k$ using Dir(α)
end for

— **Federated Training Loop** —

for $t = 0, 1, \ldots, T-1$ **do**
    Server samples $\mathcal{S}^t \subset \mathcal{K}$, $|\mathcal{S}^t| = m$
    Server broadcasts $\mathbf{w}^t$ to all $k \in \mathcal{S}^t$
    for all $k \in \mathcal{S}^t$ **in parallel do**
        $\mathbf{w}_k \leftarrow \mathbf{w}^t$ ▷ Initialize from global
        for $e = 1, \ldots, E$ **do**
            for all mini-batches $\mathcal{B} \subseteq \mathcal{D}_k$ **do**
                for all $(\mathbf{x}_i, y_i) \in \mathcal{B}$ **do**
                    Compute softened outputs: $\mathbf{q}_k^i$ and $\mathbf{q}^{G,i}$
                    Compute entropy: $H_k^i = -\sum_c q_{k,c}^i \log q_{k,c}^i$ ▷ Eq. (4)
                    Compute uncertainty weight: $\omega_k^i = 1 - e^{-\beta H_k^i}$ ▷ Eq. (5)
                    Compute JS divergence: $D_{\mathrm{JS}}^i = D_{\mathrm{JS}}(\mathbf{q}^{G,i} \| \mathbf{q}_k^i)$ ▷ Eq. (3)
                    Compute distillation loss: $\ell_{\mathrm{UAD}}^i = \tau^2 \omega_k^i D_{\mathrm{JS}}^i$ ▷ Eq. (6)
                    Compute cross-entropy: $\ell_{\mathrm{CE}}^i = \mathrm{CrossEntropy}(f_{\mathbf{w}_k}(\mathbf{x}_i), y_i)$
                    Compute proximal penalty: $\ell_{\mathrm{prox}}^i = \frac{\mu}{2} \|\mathbf{w}_k - \mathbf{w}^t\|^2$
                    $\ell^i = \ell_{\mathrm{CE}}^i + \ell_{\mathrm{prox}}^i + \lambda \ell_{\mathrm{UAD}}^i$ ▷ Eq. (11)
                end for
                $\mathbf{w}_k \leftarrow \mathbf{w}_k - \eta_l \nabla_{\mathbf{w}_k} \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \ell^i$
            end for
        end for
        Quantize $\mathbf{w}_k$ to INT8 and transmit to server ▷ Eq. (14)
    end for
    Server aggregates: $\mathbf{w}^{t+1} = \sum_{k \in \mathcal{S}^t} \frac{n_k}{\sum_j n_j} \mathbf{w}_k^{t+1}$ ▷ Eq. (10)
end for
return $\mathbf{w}^T$ ▷ Final global model

4. Implementation and Results Analysis

4.1. Experimental Setup

The implementation of FedREUAD follows the hyperparameter configuration reported in Table 2, determined through systematic preliminary experiments. Both teacher and student models adopt a 1D convolutional neural network architecture, with channel configurations of $[128, 256, 384, 512]$ and $[32, 64, 128, 256]$, respectively. Each convolutional block employs ReLU activations, layer normalization, and dropout (rate = 0.3) for regularization. All baselines are trained using the teacher architecture to ensure a fair comparison, while FedREUAD employs the lightweight student model on each client, with knowledge transferred to the server via uncertainty-weighted JSD. Experiments are conducted on two publicly available IoT/IoMT network traffic datasets. CICIoT2023 [Neto et al. 2023] comprises 31 attack categories derived from PCAP recordings across a realistic IoT testbed, while CICIoMT2024 [Dadkhah et al. 2024] covers traffic from 40 IoMT devices spanning 7 attack classes. Both datasets provide multi-class flow-level features extracted from raw packet captures. Prior to training, each feature dimension is normalized to the $[0, 1]$ range using min-max scaling. The data is partitioned into training and test sets following an 80/20 split; the training portion is then distributed across $N = 10$ clients via $\alpha = 0.5$, which induces moderate statistical heterogeneity across clients and is consistent with standard non-IID federated evaluation protocols. Evaluation metrics include accuracy, F1, detection rate, false positive rate, uplink communication per round, local training time per epoch, and floating point operations. We compare FedREUAD against four baselines: (i) FedAvg [McMahan et al. 2017], (ii) FedProx [Li et al. 2020], (iii) SCAFFOLD [Karimireddy et al. 2020], and (iv) MOON [Li et al. 2021]. A centralized training baseline (CL) provides an upper-bound reference.

Table 2. Hyperparameter configuration for all FedREUAD experiments.

Hyperparameter	Symbol	Value	Hyperparameter	Symbol	Value
Total clients	N	10	Clients per round	m	5
Communication rounds	T	50	Local epochs	E	5
Local learning rate	η_l	10^{-3}	Batch size	B_s	64
FedProx coefficient	μ	0.01	Distillation temperature	τ	3.0
Distillation weight	λ	0.5	Uncertainty scale	β	1.0
Pruning ratio (global)	ρ	0.5	Quantization bits	b	8
Dirichlet concentration	α	0.5			

4.2. Analysis and Discussion of Results

Table 3 presents a comprehensive evaluation of FedREUAD against established FL baselines on two benchmark IoT security datasets, CICIoT2023 and CICIoMT2024. Empirical results demonstrate that FedREUAD consistently achieves state-of-the-art performance across all evaluation metrics, substantially narrowing the performance gap with the CL upper bound. On CICIoT2023, FedREUAD attains an accuracy of 98.29% and F1-score of 98.22%, representing absolute improvements of +3.51% and +3.25% over FedAvg, and +0.65% and +0.45% over the strongest competing federated method (MOON). More critically for intrusion detection applications, FedREUAD reduces the false positive rate (FPR) to 0.02, an 83% relative

reduction compared to FedAvg (0.12) and a 75% reduction versus SCAFFOLD and MOON (0.09 and 0.08, respectively), while maintaining a detection rate (DR) of 98.29%. This favorable DR/FPR trade-off is particularly significant in security-critical deployments where excessive false alarms incur substantial operational overhead. On the more challenging CICIoMT2024 dataset, FedREUAD further consolidates its advantage, achieving 99.11% accuracy and 99.05% F1-score, within 0.09% and 0.35% of the centralized oracle, respectively. The balanced accuracy metric (99.01%) confirms that these gains are not attributable to class imbalance artifacts but reflect robust generalization across minority attack categories. Notably, FedREUAD's consistent outperformance of variance-reduction (SCAFFOLD) and model-contrastive (MOON) approaches suggests that its underlying mechanism, presumably addressing client drift and feature representation alignment under non-IID data distributions, provides complementary benefits beyond existing optimization-centric strategies. Collectively, these evidence-based results validate FedREUAD as a highly effective federated anomaly detection framework, delivering near-centralized performance while preserving data locality and achieving the low false-positive characteristics essential for real-world IoT security operations.

Table 3. Performance comparison on CICIoT2023 and CICIoMT2024 ($\alpha = 0.5$)

Method	CICIoT2023				CICIoMT2024			
	Acc (%)	F1 (%)	DR (%)/FPR	Bal.Acc (%)	Acc (%)	F1 (%)	DR (%)/FPR	Bal.Acc (%)
CL (upper bound)	98.80	98.60	98.40 / 0.01	98.38	99.20	99.40	99.20 / 0.01	99.50
FedAvg	94.78	94.97	94.77 / 0.12	90.99	96.80	96.45	96.80 / 0.04	95.20
FedProx	96.11	96.34	96.34 / 0.09	93.01	96.85	96.50	96.84 / 0.04	96.35
SCAFFOLD	97.24	97.31	97.24 / 0.09	93.77	97.45	97.20	97.35 /0.03	97.28
MOON	97.64	97.77	97.64 / 0.08	94.84	97.80	97.50	97.68 / 0.03	97.55
FedREUAD	98.29	98.22	98.29 / 0.02	96.02	99.11	99.05	99.02 / 0.01	99.01

4.3. Communication and Computational Efficiency Analysis

Table 4 evaluates the resource efficiency of FedREUAD relative to the baselines, revealing substantial advantages in communication overhead, model complexity, and deployment latency. Critically, FedREUAD reduces the total communication payload (sent and received bytes) by 57.9% on CICIoT2023 (818.468 kB vs. 1945.781 kB for baselines) and 60.8% on CICIoMT2024 (393.171 kB vs. 1003.007 kB), directly alleviating bandwidth constraints inherent in resource-constrained IoT environments. This communication gain stems from a 49.8% reduction in trainable parameters ($544,032$ vs. $1,083,872$ on CICIoT2023; $543,008$ vs. $993,248$ on CICIoMT2024), indicating that FedREUAD employs a more compact architectural design or parameter-efficient adaptation mechanism without compromising representational capacity. Computationally, FedREUAD requires fewer floating-point operations ($49,812$ vs. $52,382$ FLOPs on CICIoT2023; $50,326$ vs. $57,265$ on CICIoMT2024), yielding 4.9%–12.1% savings in client-side inference and training cost, particularly valuable for edge devices with limited processing budgets. Most notably, upload latency is reduced by an order of magnitude: FedREUAD achieves 0.0011 s on CICIoT2023 (vs. 0.0138–0.0233 s for baselines) and 0.0061 s on CICIoMT2024 (vs. 0.0155–0.395 s), representing 87%–98.5% faster model synchronization. When contextualized with the accuracy results in Table 3 and Figure 1 these efficiency

gains are achieved without sacrificing detection performance; indeed, FedREUAD simultaneously attains near-centralized accuracy and superior resource efficiency. This advantage positions FedREUAD as a practically deployable solution for large-scale, heterogeneous IoT security systems where bandwidth, energy, and latency constraints are first-order design considerations.

Table 4. Evaluation of communication and complexity efficiency on CICIoT2023 and CICIoMT2024

Method	CICIoT2023				CICIoMT2024			
	Sent & Rec. Bytes (kB)	FLOPs	Params.	Upload time	Sent & Rec. Bytes (kB)	FLOPs	Params.	Upload time
CL (upper bound)	NA	NA	NA	NA	NA	NA	NA	NA
FedAvg	1945.781	52,382	1,083,872	0.0157	1003.007	57265	993248	0.0.0181
FedProx	1945.781	52,382	1,083,872	0.0138	1003.007	57265	993248	0.0155
SCAFFOLD	1945.781	52,382	1,083,872	0.0211	1003.007	57265	993248	0.395
MOON	1945.781	52,382	1,083,872	0.0233	1003.007	57265	993248	0.0168
FedREUAD	818.468↓	49,812↓	544,032↓	0.0011↓	393.171↓	50326↓	543008↓	0.0061↓

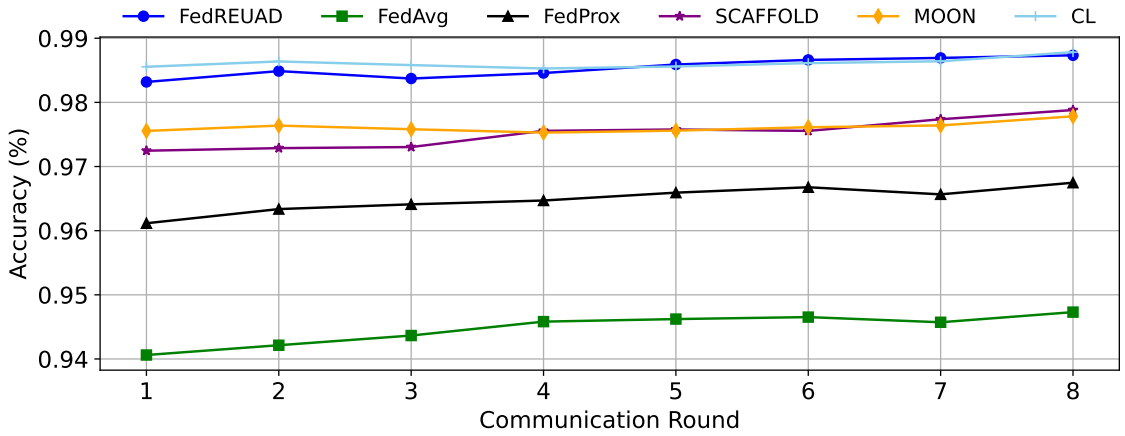

Figure 1. Performance accuracy Δ over first 8 rounds on the CICIoT2023 dataset.

4.4. Ablation Studies

To isolate the contribution of each component of FedREUAD, we conduct a structured ablation study on CICIoT2023 with $\alpha = 0.5$. Table 5 presents the ablation configurations. The results validate the complementary contribution of each FedREUAD component. The vanilla FedAvg baseline (A1) yields an F1 of 94.97%, while adding the FedProx proximal term alone (A2) improves both accuracy and F1. JS-based distillation alone (A3, 97.34% F1) outperforms KL distillation with prox (A4, 96.68% F1), confirming that the symmetric Jensen–Shannon formulation mitigates the calibration bias inherent to asymmetric KL divergence during soft-label transfer. The progressive gain from unweighted JS (A5, 96.88%) to uncertainty-weighted JS (A6, 98.22%) further demonstrates that per-sample uncertainty weighting ω effectively suppresses overconfident soft labels from poorly calibrated local teachers. Finally, the full FedREUAD (A7) matches the uncompressed variant (A6) in predictive performance while achieving a compression ratio of $\gamma = 8\times$, confirming that the pruning and INT8 quantization pipeline introduces no measurable accuracy

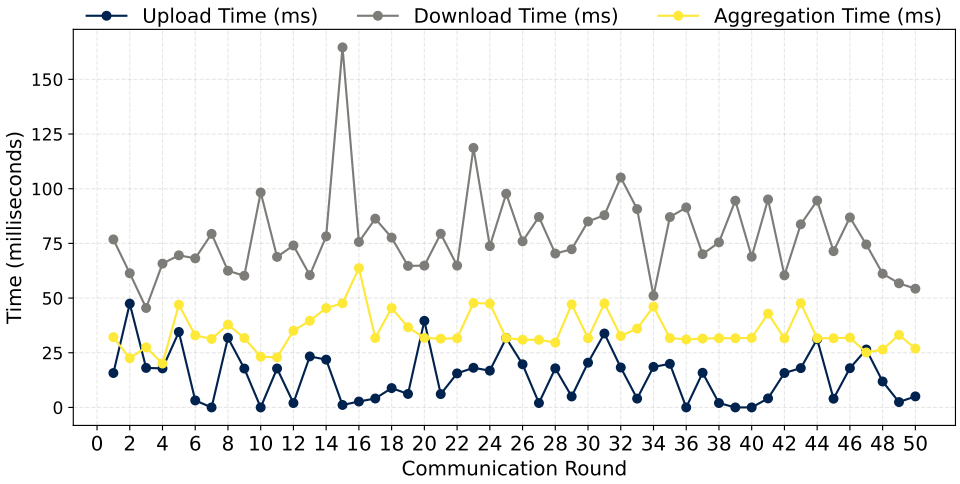

Figure 2. Time complexity analysis of FedREUAD method showing the changes in upload/download and aggregation time in ms on CICIoMT2024.

degradation, making the framework viable for resource-constrained IoT edge deployments.

Table 5. Ablation study configurations.

Variant	**Prox**	**UAD**	**Comp**	Acc.	F1	DR (%)	γ
A1: FedAvg (no prox, no distill)	×	×	×	94.78	94.97	94.77	1
A2: FedProx only	✓	×	×	96.11	96.34	96.34	1
A3: JS distill only (no prox)	×	✓	×	97.34	97.34	97.34	1
A4: KL distill + prox	✓	KL	×	96.68	96.68	96.68	1
A5: UAD (no uncertainty wt.)	✓	JS (no ω)	×	96.88	96.88	96.88	1
A6: Full FedREUAD (no comp)	✓	✓	×	98.50	98.32	98.39	1
A7: Full FedREUAD	✓	✓	✓	98.29	98.22	98.29	8

5. Conclusion

This paper introduced FedREUAD, a FL framework for IoT intrusion detection that integrates three complementary technical contributions: FedProx aggregation for proximal stabilization under non-IID data, an uncertainty-aware JSD loss that resists the propagation of overconfident soft labels, and a structured pruning plus INT8 quantization pipeline that reduces per-round communication cost. The uncertainty weighting, parameterized through entropy, down-scales contributions from overconfident local predictions, the dominant source of distillation error under severe label skew. Complexity analysis confirms that the overhead introduced by the uncertainty-aware distillation term is asymptotically negligible compared to the dominant training cost, making FedREUAD practically deployable on resource-constrained endpoints. Future research will investigate adaptive μ scheduling that increases μ based on model divergence, and personalized τ per client, or per communication round, based on the observed client drift could improve the fidelity of knowledge transfer in later training stages.

Acknowledgment

This work is supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) under Grant 2023/07184-1 & 2024/12903-0, and partially by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) under finance Code 001

References

Bhatt, S., Gupta, A., and Rai, P. (2024). Federated learning with uncertainty via distilled predictive distributions. In *Proceedings of the 15th Asian Conference on Machine Learning*, volume 222 of *Proceedings of Machine Learning Research*, pages 153–168. PMLR.

Dadkhah, S., Neto, E. C. P., Ferreira, R., Molokwu, R. C., Sadeghi, S., and Ghorbani, A. A. (2024). Ciciomt2024: A benchmark dataset for multi-protocol security assessment in iomt. *Internet of Things*, 28:101351.

Friha, O., Ferrag, M. A., Shu, L., Maglaras, L., and Choo, K.-K. R. (2022). FELIDS: Federated learning-based intrusion detection system for agricultural Internet of Things. *Journal of Parallel and Distributed Computing*, 165:17–31.

Fuglede, B. and Topsoe, F. (2004). Jensen-shannon divergence and hilbert space embedding. In *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.

Han, T., Nguyen, N. T., Peng, H., Xu, C., and Wang, S. (2024). A comprehensive review of model compression techniques in machine learning. *Applied Intelligence*, 54:11804–11844.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Itahara, S., Nishio, T., Koda, Y., Morikura, M., and Yamamoto, K. (2021). Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-iid private data. *IEEE Transactions on Mobile Computing*, 22(1):191–205.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5132–5143.

Lavaur, L., Pahl, M.-O., Busnel, Y., and Autrel, F. (2022). The evolution of federated learning-based intrusion detection and mitigation: A survey. *IEEE Transactions on Network and Service Management*, 19(3):2309–2332.

Li, Q., He, B., and Song, D. (2021). Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10713–10722.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450.

Liu, T., Xia, J., Ling, Z., Fu, X., Yu, S., and Chen, M. (2023). Efficient federated learning for aiot applications using knowledge distillation. *IEEE Internet of Things Journal*, 10(8):7229–7243.

Liu, Y., Ma, Z., Liu, X., Ma, S., Nepal, S., and Deng, R. (2022). Distributed intrusion detection system for IoT based on federated learning and edge computing. *Computers & Security*, 115:102622.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 54:1273–1282.

Mothukuri, V., Khare, P., Parizi, R. M., Pouriyeh, S., Dehghantanha, A., and Srivastava, G. (2022). Federated-learning-based anomaly detection for IoT security attacks. *IEEE Internet of Things Journal*, 9(4):2545–2554.

Neto, E. C. P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., and Ghorbani, A. A. (2023). Ciciot2023: A real-time dataset and benchmark for large-scale attacks in iot environment. *Sensors*, 23(13):5941.

Okey, O. D., Rodríguez, D. Z., Guimarães, F. G., and Kleinschmidt, J. H. (2025). Cafiks: Communication-aware federated ids with knowledge sharing for secure iot connectivity. *IEEE Access*, 13:140694–140708.

Prakash, P., Ding, J., Chen, R., Qin, X., Shu, M., Cui, Q., Guo, Y., and Pan, M. (2022). Iot device friendly and communication-efficient federated learning via joint model pruning and quantization. *IEEE Internet of Things Journal*, 9(15):13638–13650.

Rahman, S. A., Tout, H., Talhi, C., and Mourad, A. (2020). Internet of things intrusion detection: centralized, on-device, or federated learning? *IEEE Network*, 34(6):310–317.

Shen, M., Zhu, L., Du, X., and Guizani, M. (2024). Federated learning ensemble knowledge distillation for intrusion detection in heterogeneous IoT networks. *IEEE Transactions on Information Forensics and Security*, 19:3918–3931.

Statista Research Department (2025). Number of iot devices 2015-2025. `https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/`. [Accessed, April 25, 2025].

Talpini, J., Civiero, N., Sartori, F., and Savi, M. (2025). A federated approach to enhance calibration of distributed ml-based intrusion detection systems. In *In Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025)*, volume 2, pages 840–848.

Talpini, J., Sartori, F., and Savi, M. (2024). Enhancing trustworthiness in ml-based network intrusion detection with uncertainty quantification. *Journal of Reliable Intelligent Environments*, 10(4):501–520.