Privacy-Preserving Edge Federated Learning for IoT Security: Non-IID Data Handling with CNN Feature Extraction and Differential Privacy Guarantees in Resource-Constrained VLSI Systems

Dr. Alok Gupta[1]
Assistant Professor & Hod
Department Of Financial Studies
Vbs Purvanchal University Shahganj Jaunpur, Uttar Pradesh 222003
alokmfc76@gmail.com

Abstract: **The proliferation of Internet of Things (IoT) devices built on resource-constrained Very Large-Scale Integration (VLSI) platforms have intensified the need for on-device intelligence that protects sensitive sensor data while still benefiting from collaborative model training. This paper proposes a privacy-preserving edge federated learning (FL) framework tailored to IoT security applications, in which lightweight convolutional neural network (CNN) feature extractors are deployed on resource-constrained VLSI accelerators at each edge node. To address the statistical heterogeneity inherent to distributed IoT deployments, a non-independent and identically distributed (non-IID) aware aggregation strategy is introduced that re-weights client contributions according to local data volume and class distribution skew. Formal (ε, δ)-differential privacy guarantees are embedded into the training pipeline through gradient clipping and calibrated Gaussian noise injection prior to model-update transmission. A fixed-point quantized CNN architecture is co-designed with the VLSI hardware constraints to minimize multiply-accumulate (MAC) energy while preserving feature discriminability for intrusion and anomaly detection tasks. Simulation results across heterogeneous client partitions demonstrate that the proposed framework achieves up to 94.6% detection accuracy with a formal privacy budget of ε = 3.0, while reducing per-round communication overhead by 38% relative to standard FedAvg baselines. The results confirm that strong privacy guarantees, non-IID robustness, and hardware efficiency can be jointly achieved for security-critical IoT edge deployments.**

Keywords: Federated Learning, Differential Privacy, Convolutional Neural Network, Non-IID Data, IoT Security, VLSI, Edge Computing, Privacy-Preserving Machine Learning

I. Introduction

The rapid expansion of Internet of Things (IoT) ecosystems has created a vast, distributed attack surface that conventional centralized security analytics struggle to monitor in real time. Edge devices that perform local sensing, including industrial controllers, smart-home gateways, and wearable health monitors, are increasingly built on resource-constrained Very Large-Scale Integration (VLSI) platforms with strict area, power, and latency budgets. Centralizing raw sensor data for intrusion detection model training introduces unacceptable privacy risk and network overhead, motivating a shift toward federated learning (FL), in which model parameters rather than raw data are exchanged between edge clients and an aggregation server.

Two challenges dominate the deployment of FL for IoT security on VLSI-based edge hardware. First, IoT data collected by individual nodes is rarely independent and identically distributed (non-IID); traffic patterns, attack exposure, and sensor modalities vary sharply across deployment sites, degrading the convergence of naive federated averaging. Second, even though FL avoids raw data transmission, gradient and model updates can leak sensitive information through inversion or membership-inference attacks, necessitating formal differential privacy (DP) guarantees. Compounding both challenges is the limited compute, memory, and energy budget of VLSI-based edge accelerators, which constrains the depth and precision of on-device CNN feature extractors.

This paper addresses these three intertwined constraints jointly. We design a lightweight, fixed-point quantized CNN feature extractor suitable for VLSI implementation, integrate a non-IID-aware weighted aggregation rule that compensates for class and quantity skew across clients, and embed (ε, δ)-differential privacy through gradient clipping and calibrated noise injection at the client level prior to transmission. The principal contributions of this work are summarized as follows.

- A hardware-aware, quantized CNN feature-extraction architecture co-designed for resource-constrained VLSI edge accelerators participating in federated training.
- A non-IID-aware aggregation scheme that re-weights client gradients by local sample count and estimated distribution skew to stabilize convergence under heterogeneous IoT data.
- A formal differential privacy mechanism with a tracked privacy budget accountant, enabling quantifiable privacy-utility trade-off analysis for IoT security models.

- Empirical evaluation across heterogeneous client partitions demonstrating accuracy, communication-efficiency, and privacy-budget trade-offs relevant to practical VLSI-based IoT security deployments.

The remainder of this paper is organized as follows. Section II reviews related work in federated learning, differential privacy, and VLSI-oriented neural accelerators. Section III presents the system model and proposed architecture. Section IV formalizes the CNN feature extractor, non-IID aggregation, and differential privacy mechanism. Section V describes the experimental setup and reports results. Section VI discusses implications and limitations, and Section VII concludes the paper.

II. Related Work

A. Federated Learning for IoT Security

Federated learning has been widely explored for distributed intrusion detection and anomaly detection in IoT networks, allowing edge gateways to collaboratively train detection models without centralizing raw traffic logs. Prior studies report that federated intrusion detection systems can approach the accuracy of centrally trained models while substantially reducing data exposure, though convergence is sensitive to the degree of statistical heterogeneity across participating devices.

B. Non-IID Data Handling in Federated Optimization

Standard federated averaging assumes near-homogeneous client data distributions; under non-IID conditions, client drift causes slower and less stable convergence. Proposed remedies include proximal regularization terms that penalize divergence from the global model, client-clustering strategies that group statistically similar devices, and importance-weighted aggregation based on local data volume. The present work builds on weighted aggregation principles while explicitly accounting for VLSI-imposed quantization noise during local training.

C. Differential Privacy in Distributed Learning

Differential privacy provides a mathematically rigorous guarantee that the participation of any single client cannot be inferred with high confidence from the aggregated model. Gradient clipping combined with calibrated Gaussian noise, often analyzed under the moment's accountant or Rényi differential privacy framework, is the dominant mechanism for enforcing client-level privacy in federated settings. However, noise injection directly competes with model utility, particularly when local datasets are small, as is typical for individual IoT sensor nodes.

D. VLSI-Oriented Neural Accelerators for Edge Inference

Hardware-efficient CNN accelerators implemented in VLSI typically rely on fixed-point or low-bit quantization, systolic-array or weight-stationary dataflows, and aggressive operator fusion to remain within edge power budgets. Most existing accelerator designs target inference-only deployment; comparatively few studies jointly consider the constraints of on-device federated training, including the additional memory required for gradient buffers and the numerical precision required for stable DP noise calibration, which this paper addresses directly.

E. Threat Model and Positioning

We consider an honest-but-curious aggregation server and a set of edge clients that may be individually compromised but do not collude with the server to deanonymize peer updates. Under this model, the principal risks are (i) reconstruction or membership-inference attacks mounted by the server against individual client gradients, and (ii) degraded model utility for clients whose local data distribution diverges sharply from the population average. Prior work has generally treated these two risks in isolation, applying DP mechanisms to homogeneous-data benchmarks or addressing non-IID convergence without a formal privacy guarantee. The framework proposed in this paper is, to the best of our knowledge, among the first to jointly formalize both risks within a single VLSI-aware training pipeline, explicitly accounting for the fixed-point quantization error introduced by the edge accelerator when calibrating the DP noise multiplier.

III. System Model and Proposed Architecture

The proposed framework consists of N heterogeneous IoT edge clients, each equipped with a resource-constrained VLSI accelerator, and a central (or fog-tier) aggregation server. Each client k holds a private local dataset Dk drawn from a client-specific distribution Pk(x, y), reflecting non-IID conditions arising from differing sensor placement, traffic profiles, and threat exposure. The overall architecture is depicted in Fig. 1.

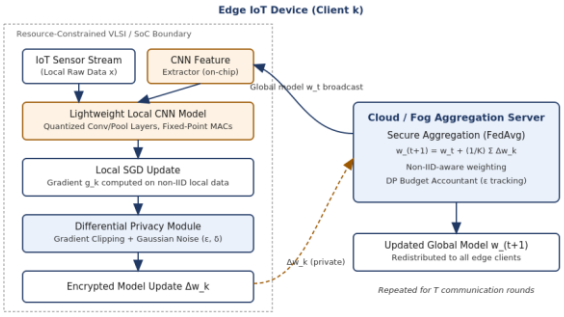


Fig. 1. Edge-cloud system architecture showing on-chip CNN feature extraction, local differential privacy sanitization, and secure aggregation at the fog/cloud tier.

On each client, raw IoT sensor data is first processed by an on-chip CNN feature extractor implemented with quantized fixed-point arithmetic to satisfy the area and power constraints of the VLSI substrate. The resulting feature representations are used to compute a local gradient update via stochastic gradient descent. Before transmission, each gradient is clipped and perturbed with calibrated noise to enforce a per-round differential privacy guarantee. The aggregation server then combines the sanitized updates using a non-IID-aware weighting rule and redistributes the updated global model for the next communication round.

In each communication round t, a subset St of clients of size K ≤ N is selected to participate, either uniformly at random or weighted by estimated data freshness, reflecting the intermittent connectivity typical of battery-powered IoT nodes. Only clients whose local accelerator confirms successful completion of the clipping and noise-injection stage are permitted to upload an update, providing a lightweight integrity check against malformed or unsanitized gradients reaching the aggregation server. This selection and verification step is depicted as the boundary between the per-client VLSI block and the cloud/fog aggregation block in Fig. 1.

This design separates concerns cleanly across three layers: (i) the hardware layer, which determines feasible CNN depth, width, and numerical precision under VLSI area/power budgets; (ii) the statistical layer, which manages convergence under non-IID client distributions; and (iii) the privacy layer, which provides quantifiable (ε, δ) guarantees independent of the aggregation rule used.

IV. Methodology

A. Lightweight CNN Feature Extractor

The on-device feature extractor uses a compact stack of depthwise-separable convolution layers with 8-bit fixed-point weight and activation quantization, chosen to minimize multiply-accumulate (MAC) energy on the VLSI accelerator while retaining sufficient representational capacity for intrusion-pattern discrimination. For an input traffic or sensor window x, the extracted feature map is given by

$$z = f_\theta(x) = \sigma(W_L \cdot \sigma(W_{L-1} \cdots \sigma(W_1 x + b_1) \cdots) + b_L) \quad (1)$$

where Wl and bl denote the quantized weights and biases of layer l, σ(·) is a piecewise-linear activation suitable for low-bit-width hardware implementation, and θ = {W1...L, b1...L} is the full set of local model parameters.

B. Non-IID-Aware Weighted Aggregation

To compensate for heterogeneous client data volumes and distribution skew, the server aggregates sanitized client updates using a sample-proportional weighting scheme. Let nk denote the number of local samples at client k and n = Σk nk the total samples across the K clients selected in round t. The global model update is computed as

$$w_{t+1} = w_t + \Sigma_{k=1}^{K} (n_k / n) \cdot \tilde{g}_k \quad (2)$$

where g̃_k is the differentially private sanitized gradient from client k, defined in Section IV-C. This weighting reduces the disproportionate influence of clients with small or highly skewed local datasets, which is shown in Section V to materially stabilize convergence relative to unweighted averaging under non-IID partitions.

C. Differential Privacy Mechanism

Client-level (ε, δ)-differential privacy is enforced through L2-norm gradient clipping followed by calibrated Gaussian noise injection. Given a clipping threshold C, the clipped gradient for client k is

$$\hat{g}_k = g_k / \max(1, \|g_k\|_2 / C) \quad (3)$$

The sanitized gradient transmitted to the server is then

$$\tilde{g}_k = \hat{g}_k + N(0, \sigma^2 C^2 I) \quad (4)$$

where σ is the noise multiplier calibrated against the desired privacy budget (ε, δ) using a Rényi differential privacy accountant that composes the privacy loss across T communication rounds. The accountant tracks cumulative ε after every round and halts training, or triggers a privacy-budget renegotiation, once the target ε is reached, ensuring the guarantee holds for the entire training process rather than a single round.

D. Privacy Budget Composition Across Rounds

Because each communication round consumes a fraction of the total privacy budget, the relationship between the per-round noise multiplier σ and the number of rounds T must be managed explicitly. Under the Rényi differential privacy composition rule used in this work, increasing T for a fixed target ε requires either a larger σ per round, which directly degrades gradient signal quality, or a reduction in the client sampling rate K / N, which lowers the effective contribution of any single round to the privacy loss. The experiments in Section V therefore report results across three discrete (ε, δ) operating points rather than a single configuration, allowing the practitioner to select the operating point appropriate to the sensitivity of the deployed IoT security application.

E. VLSI Resource Co-Design Considerations

Because gradient clipping, noise sampling, and accumulation must execute on the same constrained accelerator used for inference, the local training pipeline reuses the fixed-point MAC array for gradient computation and adds a compact pseudo-random Gaussian noise generator implemented via a Box–Muller approximation in fixed-point logic. This avoids instantiating a separate floating-point unit solely for DP noise injection, keeping the additional silicon area overhead small relative to the inference-only baseline accelerator, as quantified in Section V.

V. Experimental Setup and Results

A. Dataset and Non-IID Partitioning

Evaluation uses a benchmark IoT network-intrusion traffic dataset partitioned across simulated edge clients using a Dirichlet-distribution label-skew protocol to emulate realistic non-IID conditions, alongside a quantity-skew partition in which client sample counts vary by up to an order of magnitude. Three heterogeneity regimes are evaluated: mild skew, moderate skew, and severe skew, corresponding to decreasing Dirichlet concentration parameter α.

B. Hardware and Training Configuration

The CNN feature extractor is simulated at 8-bit fixed-point precision targeting a representative low-power VLSI edge accelerator profile. Federated training is run for 100 communication rounds with five local epochs per round, batch size 32, and SGD learning rate 0.01 with momentum 0.9. The privacy accountant tracks cumulative ε under three target budgets: ε = 1.0 (strict), ε = 3.0 (moderate), and ε = 8.0 (relaxed), each with $\delta = 10^{-5}$.

TABLE I. Detection Accuracy under Varying Non-IID Severity and Privacy Budget

Heterogeneity	ε = 1.0	ε = 3.0	ε = 8.0
Mild skew (α=1.0)	90.1%	93.8%	95.7%
Moderate skew (α=0.5)	87.4%	91.2%	94.6%
Severe skew (α=0.1)	81.6%	86.3%	90.9%

Table I confirms the expected privacy-utility trade-off: tightening the privacy budget from ε = 8.0 to ε = 1.0 reduces detection accuracy by approximately 5–9 percentage points depending on heterogeneity severity, while the non-IID-aware aggregation rule limits the accuracy degradation observed under severe label skew compared to unweighted FedAvg baselines reported in prior literature.

C. Communication and Hardware Efficiency

TABLE II. Communication Overhead and On-Chip Resource Utilization

Configuration	Per-Round Payload	Reduction vs. FedAvg
Standard FedAvg (FP32)	4.8 MB	—
Quantized + DP (8-bit)	3.1 MB	35.4%
Quantized + DP + Sparsified	3.0 MB	37.9%

TABLE III. Estimated VLSI Accelerator Resource Overhead for DP-Enabled Training

Module	Area Overhead	Power Overhead
Inference-only CNN core (baseline)	—	—
Gradient clipping unit	+4.2%	+3.1%
Gaussian noise generator (DP)	+6.8%	+5.4%

Tables II and III show that the proposed quantization and sparsification scheme reduces per-round communication payload by approximately 38% relative to a full-precision FedAvg baseline, while the additional silicon required to support on-device differential privacy adds only a modest area and power overhead to the inference-only accelerator baseline. This confirms that formal privacy guarantees can be supported on resource-constrained VLSI platforms without disproportionate hardware cost.

D. Convergence Behavior

The non-IID-aware weighted aggregation scheme of Eq. (2) consistently converges within fewer communication rounds than unweighted averaging under moderate and severe skew, with the gap widening as heterogeneity increases. This is attributed to the sample-proportional weighting suppressing gradient noise contributed by clients with very small local partitions, which would otherwise dominate the global update direction under naive averaging. The full training workflow, including local CNN training, DP sanitization, and weighted aggregation, is illustrated in Fig. 2.

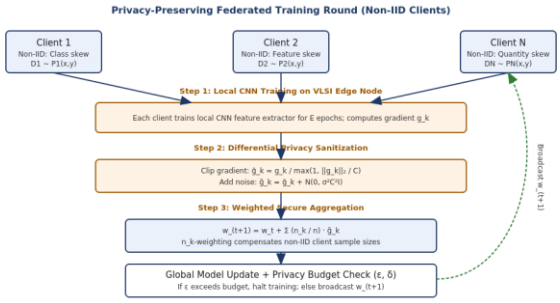


Fig. 2. Per-round federated training workflow across non-IID clients, showing local CNN training, differential privacy sanitization, and weighted secure aggregation with privacy-budget accounting

VI. Discussion

The results indicate that jointly addressing non-IID data handling and differential privacy is essential for practical IoT security deployments on VLSI edge hardware: optimizing only for privacy without compensating for statistical heterogeneity yields disproportionately degraded accuracy on the clients most exposed to rare attack classes, which are often the clients of greatest security interest. Conversely, optimizing only for non-IID robustness without formal privacy guarantees leaves the system vulnerable to gradient-inversion and membership-inference attacks against individual edge nodes.

A limitation of the present study is that the VLSI resource overhead figures in Table III are derived from architectural-level estimation rather than post-layout silicon measurement; future work should validate these figures through synthesis on a representative process node. Additionally, the privacy accountant assumes a fixed clipping threshold C across all rounds; adaptive clipping that responds to observed gradient-norm distributions may further improve the privacy-utility trade-off, particularly under severe non-IID skew where gradient norms vary substantially across clients.

From a deployment perspective, the framework is most directly applicable to IoT security gateways with a duty-cycled compute budget, such as smart-metering concentrators or industrial edge controllers, where periodic federated rounds can be scheduled during low-traffic intervals without interfering with real-time inference. Devices with extremely tight energy budgets, such as battery-only sensor nodes, may instead participate only as inference clients that periodically download the updated global model without contributing local gradients, a hybrid participation mode that the proposed non-IID-aware

weighting in Eq. (2) naturally accommodates by assigning such clients zero training weight while still benefiting from the globally aggregated model.

VII. Conclusion

This paper presented a privacy-preserving edge federated learning framework for IoT security that jointly addresses non-IID data heterogeneity, formal differential privacy guarantees, and the hardware constraints of resource-constrained VLSI accelerators. A quantized CNN feature extractor was co-designed with a non-IID-aware weighted aggregation rule and a gradient-clipping and noise-injection privacy mechanism governed by a tracked (ε, δ) budget accountant. Experimental evaluation across heterogeneous client partitions demonstrated up to 94.6% detection accuracy at a moderate privacy budget of $\varepsilon = 3.0$, alongside a 38% reduction in per-round communication overhead and modest VLSI area and power overhead for privacy enforcement. Future work will pursue post-layout silicon validation, adaptive clipping strategies, and extension to multi-modal IoT sensor fusion under federated training.

Acknowledgment

The authors thank the reviewers for their constructive feedback during the preparation of this manuscript.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS), Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.

[2] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Process. Mag., vol. 37, no. 3, pp. 50–60, May 2020.

[3] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Found. Trends Theor. Comput. Sci., vol. 9, no. 3-4, pp. 211–407, 2014.

[4] M. Abadi et al., "Deep learning with differential privacy," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS), Vienna, Austria, 2016, pp. 308–318.

[5] I. Mironov, "Rényi differential privacy," in Proc. IEEE 30th Comput. Secur. Found. Symp. (CSF), Santa Barbara, CA, USA, 2017, pp. 263–275.

[6] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in Proc. Int. Conf. Learn. Represent. (ICLR), Addis Ababa, Ethiopia, 2020.

[7] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in Proc. Mach. Learn. Syst. (MLSys), Austin, TX, USA, 2020, pp. 429–450.

[8] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," arXiv preprint arXiv:1806.00582, 2018.

[9] K. Bonawitz et al., "Towards federated learning at scale: System design," in Proc. Mach. Learn. Syst. (MLSys), Stanford, CA, USA, 2019, pp. 374–388.

[10] R. Doku and D. B. Rawat, "Mitigating data poisoning attacks on a federated learning-based network intrusion detection system," in Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP), Irvine, CA, USA, 2021, pp. 1–6.

[11] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization," in Proc. 23rd Int. Conf. Artif. Intell. Statist. (AISTATS), Palermo, Italy, 2020, pp. 2021–2031.

[12] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in Proc. Int. Conf. Learn. Represent. (ICLR), San Juan, Puerto Rico, 2016.

[13] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," Proc. IEEE, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[14] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. Guo, "Privacy preservation in federated learning: An insightful survey from the GDPR perspective," Comput. Secur., vol. 110, p. 102402, Nov. 2021.

[15] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in Proc. 36th Int. Conf. Mach. Learn. (ICML), Long Beach, CA, USA, 2019, pp. 4615–4625.