Proxy-Only Fog–Cloud Bidirectional Distillation for Privacy-Preserving, Deployable IoMT ECG Diagnosis under Cross-Site Heterogeneity

Tiancheng Cao[*1,2], Zixuan Shu[*1], Wei Soon Ng[1], Tong Zou[1], Chen Shen[1], Hen-Wei Huang[1,3]
[*]Equal contribution [1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore [2]Department of Emergency Medicine, Brigham and Women's Hospital, Boston [3]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

Correspondence to: Chen Shen, shenchen@ntu.edu.sg

1. Introduction

Wearable ECG monitoring is generating large-scale longitudinal data, yet robust cross-site modelling remains challenging under privacy regulation, heterogeneous patient populations, and tight edge compute and energy constraints [1]. Clinically important arrhythmias are often long-tailed and site-dependent, so simple cross-site aggregation can be brittle [2,3]. Together, these constraints demand AI-for-healthcare methods that are trustworthy by design and feasible for real-time IoMT deployment [4-6].
Because patient data cannot be centralized, cross-site learning often relies on federated optimization. However, exchanging model weights or gradients can be communication-heavy and may expose update-based leakage, especially under non-IID and label-shifted conditions [7]. Parameter averaging can also be unstable when sites differ substantially in prevalence and acquisition protocols [8]. We therefore adopt a different framework which the cloud never receives patient-level signals, labels, or model parameters.
We propose a fog–cloud bidirectional distillation framework shown in Fig. 1 that coordinates multiple hospital fog nodes and a central cloud teacher via proxy-only logits exchange. This design stabilizes collaboration under heterogeneity while keeping communication and governance overhead low. We further demonstrate an edge-oriented quantized deployment pathway for low-latency inference, closing the loop from privacy-preserving learning to deployable IoMT operation.

2. Method

2.1 Setting and communication boundary

We study cross-site ECG diagnosis with $\mathcal{K}$ hospital sites. Each site trains a local student model on private ECG data, while a cloud teacher coordinates learning. To satisfy governance constraints, the cloud never receives patient-level signals, labels, or model parameters. Coordination is performed only on a small public proxy dataset $\mathcal{D}_p$, where sites upload proxy-evaluated logits and the cloud returns proxy soft targets.

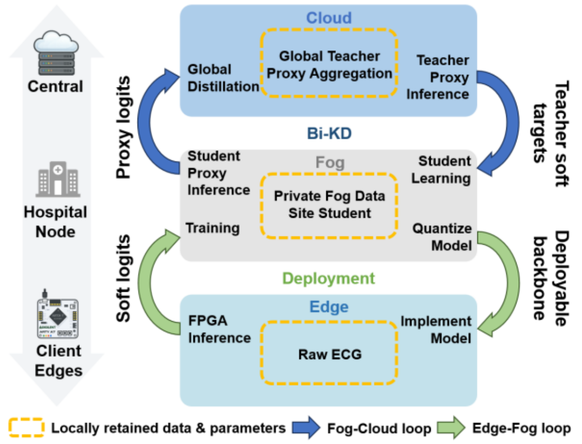

Fig. 1. The overview of the proposed framework where proxy-only logits up, teacher soft targets down, with Edge-QAT enabling deployable real-time FPGA inference.

This boundary makes the collaboration easy to audit in practice and reduces communication compared with parameter exchange.

2.2 Fog–Cloud Bidirectional Knowledge Distillation (FCBiKD)

Our training proceeds in rounds and follows a simple bidirectional loop (Algorithm 1). Bottom-up, each fog node updates its student on private data and then evaluates the updated student on the proxy dataset $\mathcal{D}_p$, uploading only the resulting logits to the cloud. The cloud aggregates logits from all sites to form consensus soft targets on $\mathcal{D}_p$ and uses them to update a cloud teacher model. We use temperature-scaled soft targets on $\mathcal{D}_p$ to smooth cross-site supervision and avoid overfitting to any single site's bias. Top-down, the cloud teacher produces global soft targets on the proxy set and broadcasts them back to all sites. Each fog node then updates its student by combining standard supervised learning on private data with a distillation regularizer that matches the teacher's proxy soft targets. Because the proxy set is fixed and public, the exchanged signal has a consistent reference across rounds, which improves convergence stability compared with averaging parameters under severe label shift. This design transfers cross-site knowledge through proxy predictions rather than parameter exchange, making the collaboration lightweight and stable under heterogeneity.

Algorithm 1 Algorithm Framework of FCBiKD

Initial: nodes $\mathcal{K} = \{k \mid k = 1, 2, \ldots, K\}$, teacher model parameter $w_s^{0,0}$, student model parameter $w_{\mathcal{K}}^0 = \{w_k^0 \mid k = 1, 2, \ldots, K\}$.
Input: training round R, teacher per-train epoch E, public proxy dataset $\mathcal{D}_p$, private dataset $\mathcal{D}_{\mathcal{K}} = \{\mathcal{D}_k \mid k = 1, 2, \ldots, K\}$, balance factor λ, temperature T, student training learning rate $\eta_{\mathcal{K}}^f = \{\eta_k^f \mid k = 1, 2, \ldots, K\}$, teacher training learning rate η_s^f. student distillation learning rate $\eta_{\mathcal{K}}^h = \{\eta_k^h \mid k = 1, 2, \ldots, K\}$, teacher distillation learning rate η_s^h.
Output: trained teacher model parameter $w_s^{E,R}$, trained student model parameter $w_{\mathcal{K}}^R = \{w_k^R \mid k = 1, 2, \ldots, K\}$.
for $e \leftarrow 1$ to E **do**
    $w_s^{e+1,0} \leftarrow$ pre-train teacher model.
end for
for $r \leftarrow 1$ to R **do**
    for all clients in parallel **do**
        $w_k^{r+1/2} \leftarrow$ train student model,
        $Z_k^r \leftarrow$ inference on $\mathcal{D}_p$,
        Send Z_k^r to the server.
    end for
    $Z^r \leftarrow$ aggregate the uploaded logits,
    $Q^r \leftarrow$ calculate the soft target base on Eqn. (6),
    $w_s^{E,r+1} \leftarrow$ distillate teacher model using Q^r,
    $\tilde{Q}^r \leftarrow$ generate the soft label on $\mathcal{D}_p$,
    Send $\tilde{Q}^r$ to the clients.
    for all clients in parallel **do**
        $w_k^{r+1} \leftarrow$ distill student model on $\tilde{Q}^r$.
    end for
end for

2.3 Edge-oriented deployment pathway

After distillation, each site performs edge-oriented quantization-aware training to produce an integer-friendly student model for deployment. The resulting lightweight PPF backbone [9] is mapped to an Artix-7 FPGA, enabling deterministic low-latency inference under strict power and memory constraints. We report latency and energy efficiency to validate that the proxy-only learning protocol translates into practical real-time IoMT deployment.

3. Results

We evaluate cross-site ECG diagnosis on MIT-BIH dataset across 3 simulated hospital sites with non-IID data and long-tailed arrhythmia classes. Our proxy-only bidirectional distillation yields high and stable performance, achieving 98.1% particularly under label shift. This makes collaboration lightweight, since per-round communication scales with proxy size rather than model size, while maintaining a simple and auditable governance boundary. For deployment, we apply edge-oriented quantization-aware training and map the resulting lightweight PPF backbone to an Artix-7 FPGA for deterministic edge inference. The deployed system, as shown in TABLE. I, achieves 1.98 ms latency and 0.169 mJ per inference with a 27.3K-parameter model under streaming inputs, supporting real-time IoMT monitoring under tight power and memory budgets. Overall, the results demonstrate an end-to-end path from privacy-aligned cross-site learning to deployable AI-for-healthcare operation.

TABLE I
FPGA Implementation Details

Development Board		Arty A7 100T
NN Implemented		Parallel PoolFormer
Parallel PoolFormer	Precision	W8/A16
	LUTs	21661
	FFs	31440
	BRAMs	62
Inference Latency (ms)		1.98
Inference Accuracy (%)		98.1%
Inference Power (W)	Dev. Board	1.88
	FPGA core	0.222
Energy/Inference (mJ)	Dev. Board	1.43
	FPGA core	0.169

Acknowledgments

This research is supported by Nanyang Assistant Professorship Start-up Grant and the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Futures program.

References

[1] K. Bayoumy *et al.*, “Smart wearable devices in cardiovascular care: where we are and how to move forward,” *Nat. Rev. Cardiol.*, 2021.
[2] T. Cao et al, “Cybersecure End-to-End FPGA-Accelerated ECG Monitoring for Precision Diagnosis With Personalized CWT and Adversarial Defense,” IEEE J. Biomed. Health Inform., 2025.
[3] H. Habibzadeh *et al.*, “A survey of healthcare Internet-of-Things (HIoT): A clinical perspective,” *IEEE Internet Things J.*, 2020.
[4] T. Cao et al, “FPGA-Based Real-Time ECG Classification System Using Quantized Inception-ResNeXt Neural Network and CWT Approximation,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, 2026.
[5] A. Nechi *et al.*, “FPGA-based deep learning inference accelerators: Where are we standing?,” *ACM Trans. Reconfigurable Technol. Syst.*, 2023.
[6] T. Cao *et al.*, “LiMO: A Lightweight MambaOut system for end-to-end IoMT ECG Diagnosis with Fully Configurable Quantization Co-Design on MCU,” *Proc. IEEE EMBC*, 2025.
[7] I. Dayan *et al.*, “Federated learning for predicting clinical outcomes in patients with COVID-19,” *Nat. Med.*, 2021.
[8] C. Wu, F. Wu, L. Lyu, Y. Huang, and X. Xie, “Communication-efficient federated learning via knowledge distillation,” *Nat. Commun.*, 2022.
[9] T. Cao *et al.*, “DWT-PoolFormer: Discrete Wavelet Transform-Based Quantized Parallel PoolFormer Network Implemented in FPGA for Wearable ECG Monitoring,” *Proc. IEEE BioCAS*, 2024.