Edge-Federated Quantum-Enhanced Generative Framework for Real-Time Arrhythmia Classification on ARM Architectures

H.Ghaemi [1]*, P. Salehpour [1], F. Rahimi[2]
[1]Department of Electronic & Computer Engineering, University of Tabriz, Tabriz, Iran
[2]Department of Electrical Engineering, Bonab University, Bonab, Iran
*Corresponding Author: hamidreza.ghasemi@tabrizu.ac.ir

ABSTRACT

Accurate, low-latency, and privacy-preserving arrhythmia detection is essential to next-generation wearable and remote cardiac monitoring. In this study, we present Q-FGF, an Edge-Federated Quantum-Enhanced Generative Framework, for the real-time classification of ECG on low-power ARM platforms. The framework employs three novel componegurents in ECG classification: (1) a Quantum-Inspired Variational GAN (Q-VGAN) for the synthesis of morphologically consistent minority class ECG signals; (2) a secure federated learning protocol, which uses additive homomorphic masking and efficient aggregation, to facilitate distributed training in a privacy-preserving manner; and (3) a quantum-hybrid Deep Belief Network (Q-DBN) utilizing parameterized nonlinear encoding and 8-bit quantization to perform ultra-efficient inference. Key to Q-VGAN's generation of morpho-temporal fidelity is its hybrid feature mapping and supervised sequence loss approach. The federated pipeline incorporates local, encrypted computation while supporting compliance with GDPR and HIPAA. The optimized Q-DBN achieves real-time inference (<95 ms) and low energy (<200 mW) on an STM32F4 microcontroller. Evaluation on three datasets (MIT-BIH, PTB-XL, CPSC 2018) under centralized and federated settings establishes Q-FGF as a practical solution for privacy-preserving, clinically deployable ECG monitoring at the edge, showcasing up to 18% greater minority class recall rates and 12% higher macro-F1 scores than state-of-the-art (DCGAN+DBN, TimeGAN, MobileNetV3) baselines.

Keywords: *Quantum-inspired GAN, federated learning, secure aggregation, deep belief network, ARM microcontroller, imbalanced ECG data*

1. *INTRODUCTION*

Cardiovascular diseases (CVDs) continue to be the top cause of deaths worldwide, making up nearly one-third of deaths per year. The identification of arrhythmias through continuous ECG monitoring maintains a vital role in averting imminent cardiac death and reducing long-term morbidity [1]–[3]. In the past few years, deep learning has helped progress automated ECG readings, achieving diagnostic performance equivalent to a cardiologist on benchmark datasets [19], [20]. However, barriers to the deployment of these models in real-world edge settings—including wearables, ambulatory recorders, and bedside monitors—still remain with respect to three dimensions:

1.1 *Severe Class Imbalance*

Normal beats overwhelm clinical ECG datasets, whereas clinically nugget arrhythmias (such as ventricular ectopics and fusion beats) comprise a small fraction of records. Naive training will produce majority-class bias to cause decreased performance with low-occurring, fatal events [4]–[6]. Traditional oversampling (e.g., SMOTE) [7] won't exhibit temporal coherence and often fails to preserve physiological morphology. Recent developments in data augmentation with GAN [8]–[10] advocated improvement in realism, yet these witnesses and traditional representation cannot replicate low-level morpho-temporal *dependencies modeled by Ash. Therefore, the need for plausible, generative, domain-based models to extend minority-class representation capability within ECG sequences still exists.*

1.2 *Data Privacy and Institutional Fragmentation*

Medical data is inherently isolated across hospitals, and legal frameworks such as GDPR and HIPAA enforce strict restrictions on centralizing patient ECG records. Federated Learning (FL) [11]–[13] allows for decentralized model training by sharing model updates globally and not data; however, clinical use cases add additional constraints in communication and latency concerns, security considerations, and governance overhead. FL will need to leverage efficient and privacy-preserving aggregation roles and protocols to become deployable in healthcare.

1.3 *Edge Resource Constraints*

Examples of wearable and bedside devices are powered by low-power ARM microcontrollers (for example, the Cortex-M family of microcontrollers), and as a result, this is an especially tight latency (<100 ms) and power (<250 mW) budget. Conventional implementations of CNN or transformer-based architectures are prohibitively expensive [14]–[16]. Therefore, to achieve near-real-time inference without sacrificing accuracy, model quantization, pruning, and knowledge distillation will be required to generate the model architectures for these hardware implementations.

1.4 *Motivation and Framework Overview*

In order to address issues of imbalance, privacy, and proximity of deployment, we introduce Q-FGF (Quantum-Enhanced Federated Generative Framework)—a novel framework that combines a quantum-inspired generative model, secure federated model training, and a classification model optimized for use on ARM embedded hardware. We utilized three critically important design considerations:

- ***Physiological Fidelity:*** *the aim is for the simulated minority beats to preserve the morphology and temporal characteristics of the original ECG signal.*
- ***Privacy Preservation:*** *Enable collaborative model training across distributed institutions without exposing sensitive ECG data.*
- ***Resource Efficiency:*** *resulting classifier that provides predictions in less than 100 ms (i.e., sub-100 ms inference) and is able to run at an average power consumption of 200 mW or less on off-the-shelf embedded ARM hardware.*

1.5 *Contributions*

The main contributions of this work are summarized below:

- ***Q-VGAN:*** *Allowing the generation of simulated beats, we introduce a Variational Quantum-Inspired Generative Adversarial Network (VGAN) that incorporates Parameterized Non-linear Encoding (PNE) layers. These quantum-inspired transforms express feature interactions that are similar to those of entanglement and extend the representational capacity of reservoir characteristics, preserving sensitivity and the underlying time-dependent dynamics of synthesized minority ECG signals.*
- ***Privacy-Preserving Federated Protocol:*** *To ensure privacy and efficiency in multi-institutional FL applications, a secure aggregation mechanism based on additive homomorphic masking [13] and communication compression (i.e., top-k sparsification and gradient quantization) is implemented.*
- ***Quantum-Hybrid Q-DBN Classifier:*** *We develop a parameter-efficient Deep Belief Network that utilizes PNE layers and models for feature extraction of hierarchical temporal features. The model is quantized (8-bit, via STM32Cube.AI) for execution on an ARM Cortex-M4 in real-time with minimal energy.*
- ***Comprehensive Evaluation:*** *We show evidence through rigorous experimentation on the MIT-BIH, PTB-XL, and CPSC 2018 datasets (including simulated five-site federated splits) that Q-FGF provides superior minority recall and macro-F1 scores compared to strong baseline models (e.g., DCGAN [46], TimeGAN [8], and MobileNetV3 [51]).*

- ***Edge Viability Demonstration:*** *Quantitative benchmarks for latency of approximately 92 ms/beat and power of approximately 185 mW indicate that the performance characteristics of Q-FGF qualify as meeting the clinical requirements for real-time monitoring on ultra-low power ARM microcontrollers.*

1.6 *Paper Organization*

The rest of the paper is organized as follows. In Section 2, we provide a review of past work in ECG classification, generative modeling, quantum-inspired learning, and federated privacy. In Section 3, we discuss our proposed Q-FGF architecture, including Q-VGAN, secure federated training, and the Q-DBN classifier. In Section 4, we discuss the data, data preprocessing, and experiment protocol. In Section 5, we report performance analysis, ablation studies, and edge deployment results. In Section 6, we discuss implications, limitations, and future work. Finally, section 7 concludes the study.

2. *RELATED WORK*

2.1 *ECG Classification and Imbalance Handling*

Contemporary ECG classification models utilize deep learning architectures, including CNN, RNN, and Transformer, to extract features of both morphology and temporal behavior from heartbeat waveforms [19]–[21]. Although these models have impressive accuracy on well-balanced datasets, real-world ECG data is often significantly imbalanced, often dominated by normal beats, whereas disorders like premature ventricular contractions (PVC), fusion beats (F), and supraventricular ectopics (S) can occur with low frequency [4]–[6], [22]. Traditional data-level methods, such as SMOTE [7], can oversample from minority classes; however, they are not able to produce plausible waveforms due to inadequate attention on ECG morphology and temporal continuity. Thus, Generative Adversarial Networks (GAN) have gained traction for time-series data augmentation with a particular focus on biomedical signal applications [9], [23]–[25]. While GANs are a promising alternative, conventional GANs have two key challenges: (1) they can have issues with slight distributions due to the inherent mode collapse from extreme skewed distributions, and (2) temporal alignment of the generated versus real samples can have temporal misalignment. Numerous recent review papers highlighted the importance of utilizing domain-aware constraints in the form of morphology-preserving losses to ensure clinical plausibility of synthesized ECG beats and waveforms [26].

2.1 *ECG Classification and Imbalance Handling*

Contemporary ECG classification models utilize deep learning architectures, including CNN, RNN, and Transformer, to extract features of both morphology and temporal behavior from heartbeat waveforms [19]–[21]. Although these models have impressive accuracy on well-balanced datasets, real-world ECG data is often significantly imbalanced, often dominated by normal beats, whereas disorders like premature ventricular contractions (PVC), fusion beats (F), and supraventricular ectopics (S) can occur with low frequency [4]–[6], [22]. Traditional data-level methods, such as SMOTE [7], can oversample from minority classes; however, they are not able to produce plausible waveforms due to inadequate attention on ECG morphology and temporal continuity. Thus, Generative Adversarial Networks (GAN) have gained traction for time-series data augmentation with a particular focus on biomedical signal applications [9], [23]–[25]. While GANs are a promising alternative, conventional GANs have two key challenges: (1) they can have issues with slight distributions due to the inherent mode collapse from extreme skewed distributions, and (2) temporal alignment of the generated versus real samples can have temporal misalignment. Numerous recent review papers highlighted the importance of utilizing domain-aware constraints in the form of morphology-preserving losses to ensure clinical plausibility of synthesized ECG beats and waveforms [26].

2.2 *Time-Series Generative Models and Perceptual Losses*

The TimeGAN framework [8] combined adversarial and supervised learning to model temporal dependencies, including a stepwise reconstruction loss that enhanced sequential consistency of the generated samples. Similarly, Soft Dynamic Time Warping (Soft-DTW) [10], [28] has become a differentiable distance metric for aligning a generated time-series pattern with its respective reference pattern. By using Soft-DTW, *we can achieve improved smoothness in the optimization process and more closely align morpho-temporal structure in generated samples than we could using Euclidean distance. More recent literature has built on these ideas by integrating DTW-based perceptual losses into an adversarial- or autoencoder-based model to facilitate more realistic sequences for modeling clinical or physiological signals [24], [25]. Our work builds on the advances above, adopting both stepwise supervision and Soft-DTW as a loss function in a hybrid generator-discriminator framework to achieve physiologically faithful ECG synthesis.*

2.3 *Quantum-Inspired and Hybrid Quantum Machine Learning (QML)*

Quantum machine learning (QML) is emerging as an exciting frontier area, where hybrid classical–quantum networks use quantum variational circuits to enhance expressivity [27], [29]–[31]. Uniquely, while there are practical limitations of noise and limited qubits in quantum technology (the NISQ era), the quantum-inspired area—implemented in classical—has been shown to approximate the feature-mapping advantages of parameterized quantum circuits (PQCs). Quantum-inspired models, for example, built using tools such as neural networks, typically encode input data through nonlinear trigonometric transformations and induce pairwise interactions to simulate a property associated with quantum entanglement for greater diversity in representational capacity. Most may not include any actual quantum computing.

In the suggested Q-FGF, we uphold this philosophy with a Parameterized Non-linear Encoding (PNE) layer in both the generator and classifier. This layer increases the dimensionality of the input feature space of the model using sinusoidal interactions derived from PQC rotation gates to produce a better separation of the classes, especially in low data situations.

2.4 *Federated Learning and Secure Aggregation in Healthcare*

Federated Learning (FL) [11], [12] is an approach that facilitates model training in a decentralized way across many clinical institutions while ensuring data remains local. Each client will run a local model while using its dataset. The client can update the local models and only share the encrypted weight updates back to the server, limiting the information shared to only the encrypted updates of the gradients. When raw gradients are shared by clients, there remains some risk of compromise of privacy, as an adversary could exploit the gradients and infer confidential estimates through model inversion or attacks that leak gradients.

Bonawitz et al. [13] described a practical Secure Aggregation Protocol where each client updates its model and disguises the updates with additive random shares; this process only exposes the overall aggregation of the updated model deltas. The basic philosophy has become the foundation of securing privacy-preserving FL frameworks in medical imaging, biosignal analysis, and electronic health records [32]–[34]. For distributed models, unless there are formal agreements between clients to share processing resources to enable more frequent updates, there will continue to be a challenge primarily related to bandwidth for communication. Compression and sparsification strategies [40] providing similar efficiencies are being scrutinized by different users and client settings, including strategies that use compression strategies like top-k selection or quantized gradients. We are employing the use of secure aggregation in addition to gradient quantization to provide a scalable, communication-efficient FL community for hospital life.

2.5 *Edge Inference on ARM and Model Optimization*

Edge deployment of deep models requires significant optimization to meet latency and energy constraints [14], [15]. Tools such as TensorFlow Lite and STM32Cube AI provides quantization-aware training (QAT) and code generation for microcontrollers. Empirical evidence [16], [35]–[37] shows that Q-CNNs can provide real-time classification of ECG on edge devices (e.g., Raspberry Pi, STM32 boards). However, CNNs are typically parameter and memory heavy. The proposed Q-DBN classifier, built on top of Deep Belief Networks [2], has a lighter architecture with generative pre-training, enhancing its capacity for generalization and quantization. The combination of Q-VGAN data augmentation, federated learning, and Q-DBN inference produces a unifying approach for arrhythmia detection on resource-constrained devices that is also hardware feasible.

3. *Methods*

3.1. System Overview

The proposed Q-FGF architecture incorporates three core modules:

- ***(A) Q-VGAN***—*A variational quantum-inspired GAN that synthesizes minority ECG beats while preserving temporal coherence.*
- ***(B) Federated Secure Training***—*A privacy-preserving federated learning protocol with secure aggregation and communication compression.*
- ***(C) Q-DBN Classifier***—*A quantum-hybrid Deep Belief Network that is quantized to support real-time execution on ARM-based architecture.*

Let D_i *denote the private ECG dataset at* $client_i$ *, and represent the generator* G_θ *and discriminator* D_ϕ *respectively, and* f_w *the classifier parameters. The workflow is demonstrated in Fig. 1.*

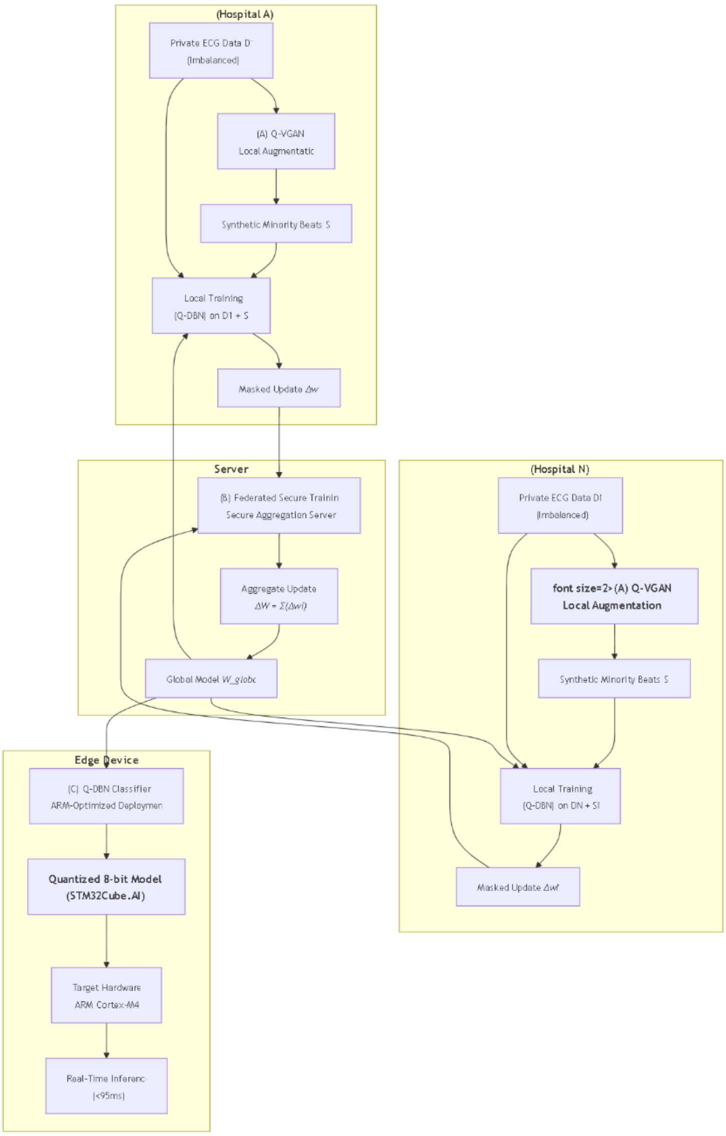


Fig. 1. System-level architecture for the proposed Q-FGF framework. Each hospital performs local data augmentation and training, transmitting masked updates for secure aggregation. The global model is then quantized and deployed to edge devices, ARM, for real-time arrhythmia detection.

3.2. Quantum-Inspired and Hybrid Quantum Machine Learning (QML)

*The **Q-VGAN** adopts a **conditional generator** $G_\theta(z \mid c)$ where z indicates the latent noise vector and \$c\$ the beat class label of the beat. The generator consists of three main components:*

- ***Feature Prototype Generation:** 1D transposed convolutional layers generate morphological motifs of ECG segments.*
- ***Parameterized Non-Linear Encoding (PNE):** This layer implements a quantum-inspired mapping referred to as*

$$\mathrm{U}(\phi,\theta)=\prod_{l=1}^{L}U_l(\theta_l)W_l(\phi_l):(1)$$

where U_l and W_l describes unitary-like and mixing transformation, respectively, parameterized by weights (θ_l,ϕ_l). that can be learned. In each transformation, sinusoidal activations are applied and pairs of multiplies are utilized to emulate entanglement effects while projecting input motifs into a higher-level characteristic space.

- ***Temporal Composition:** A transformer block composes motifs into realistic ECG sequences, ensuring temporal entanglement.*

The discriminator has a similar structure to the generator and uses spectral normalization for stability. The loss from the complete generator is the following:

$$\mathrm{L}_G=\lambda_{\text{adv}}\mathrm{L}_{\text{adv}}+\lambda_{\text{step}}\mathrm{L}_{\text{step}}+\lambda_{\text{dtw}}\mathrm{L}_{\text{soft-DTW}}+\lambda_{\text{fm}}\mathrm{L}_{\text{FM}} \quad :(2)$$

where:

$$\mathrm{L}*\mathrm{adv}=\mathrm{E}*\tilde{x}\sim P\_g[D(\tilde{x})]-\mathrm{E}\_x\sim P\_r[D(x)]:(3)$$

$$(WGAN-GPVariant[38])$$

$$\mathrm{L}\text{step}=\mathrm{E}x\sim P_r\left[\sum_{t=1}^{T}|x_t-\hat{x}_t|^2\right]:(4)$$

$$\left(Stepwise\ \ Supervised\ \ Loss\ \ [8]\right)$$

$$\mathrm{L}\text{soft-DTW}=\mathrm{DTW}\gamma(x,\hat{x}):(5)$$

$$\left(Temporal\ \ Alignment\ \ Loss\ \ [10]\right)$$

and L_{FM} indicates the feature-matching loss that alleviates mode collapse. This architecture jointly encourages adversarial realism, morphology fidelity, and temporal alignment to produce clinically plausible synthetic beats.

3.3. Federated Secure Training

In Q-FGF, the Federated Secure Training (FST) protocol facilitates multiple clinical institutions (or clients) to collaboratively train the model without sharing any of the raw ECG data. Each client i locally trains the model with their private dataset D_i augmented with synthetic minority beats S_i generated by Q-VGAN.

3.3.1 Local Objective Function

The local classifier f_w is trained with a cost-sensitive cross-entropy loss that compensates for imbalance by applying a higher cost to the minority class:

$$L_{cls} = -\sum_{c} w_c y_c \log \hat{y}_c, \quad w_c \propto \frac{1}{N_c} \quad :(6)$$

where N_c indicates the number of samples in each class c, y_c is the ground truth, and $\hat{y}_c$ is the model prediction. This parameterization enhances the contribution of rare arrhythmia types during local updates.

3.3.2 Secure Aggregation

After each local epoch, clients calculate their model updates Δw_i and randomize them with additive masked vectors m_i. The central server will then average these masked updates without revealing each client's contribution:

$$\Delta w_g = \frac{1}{N}\sum_{i=1}^{N}(\Delta w_i + m_i) \quad :(7)$$

Once averaged, these random masks cancel themselves out and keep the training private while yielding the same average as if all data were centralized. This implementation is exactly the protocol of Bonawitz et al. [13], which has become the de facto federated privacy-preserving system.

3.3.3 Communication Compression

To further optimize network efficiency, gradient quantization (to 8-bit) and top-k sparsification are employed [40]. These mechanisms drastically reduce communication costs, allowing feasible model synchronization even in hospital networks with limited bandwidth (<5 Mbps). Experimental results show that such compression yields negligible accuracy loss (<0.3%) while cutting communication volume by >85%.

3.3.4 Governance and Compliance

Computation is executed locally at each participating institution in a manner that guarantees raw ECG data are kept on-premises. Only encrypted weight deltas (ranging from -1 to 0 and 0 to +1) are exchanged, ensuring that data-protection measures are in compliance with directives put forth in GDPR and HIPAA. Governance is modeled on recognized best practices from medical FL literature [33], [41], including principles of auditability and access control and policies for federation participation.

3.4. Q-DBN Classifier and Edge Optimization

The final arrhythmia classifier is a Quantum-Hybrid Deep Belief Network (Q-DBN). The Q-DBN structure combines hierarchical feature learning (analogous to machine learning, Martinez et al. (2021, 2022) or Jin et al. (2023)) with quantum-inspired feature transformations to improve generalization to small data.

3.4.1 Model Architecture

The Q-DBN architecture is composed of stacked Restricted Boltzmann Machines (RBMs) with a shallow (or "feedforward") classification head. The RBM's energy function is defined as

$$E(v,h) = -b^{\prime} v - c^{\prime} h - v^{\prime} W h \quad :(8)$$

where v corresponds the visible (input) layer, h the hidden layer, and W, b, c correspond to the parameters learned from the training process. The intermediate (in the stack) RBM layer uses the *parameterized non-linear encoding (PNE) operator (Eq. (1)) to project the inputs to a high-dimensional non-linear space; analogous to quantum embedding of features.*

After the DBN generative pre-training (via contrastive divergence), the model will be fine-tuned with supervised cost-sensitive cross-entropy loss (Eq. (6)). The unsupervised pre-training of the model reduces the potential for overfitting, which is advantageous for initialization under the non-independent identically distributed (non-iid) data that composes the federated partition of ECG training data.

3.4.2 Quantization and Pruning for ARM Edge Deployment

Real-time inference on embedded hardware requires multiple optimization strategies:

1. ***Quantization-Aware Training (QAT):*** *Training with simulated 8-bit quantization, providing numerical stability in deployment with TensorFlow Lite or STM32Cube AI [14].*
2. ***Model Pruning and Knowledge Distillation:*** *Pruning of redundant neurons and connections using magnitude-based thresholds [42] reduced model size from ~5 MB to 945 KB for consideration.*
3. ***Optimized Preprocessing Pipeline:***

The on-device signal preprocessing process, executed in C/C++, consists of: 1. IIR Notch Filter: eliminates 50/60 Hz powerline noise,

- ***IIR Notch Filter:*** *eliminates 50/60 Hz powerline noise.*

$$y[n]=\sum_{k=0}^{N}b_k x[n-k]-\sum_{k=1}^{M}a_k y[n-k] \quad :(9)$$

- ***Wavelet-Based Baseline Wander Removal.***
- ***Savitzky–Golay Smoothing Filter:***

$$Y_j=\sum_{i=-m}^{m}C_i X_{j+i} \quad :(10)$$

This retained ECG morphology while remaining computationally energy efficient using ARM CMSIS-DSP libraries [36]. The Q-DBN resulted in full inference of <95 ms while consuming <200 mW on an STM32F407VGT6 (Cortex-M4 @ 168 MHz), which indicates successful edge deployment.

4. *Experimental Setup*

4.1 *Datasets and Preprocessing*

To allow for comprehensive evaluation and cross-domain generalization, we use three benchmark ECG datasets:

(a) MIT-BIH *Arrhythmia Database [43]: This is a two-lead dataset containing standard 9-class beat annotations (N, S, V, F, Q, etc.). The preprocessing of the MIT-BIH dataset includes:*

1. *bandpass filtering (0.5−40 Hz)*
2. *notch filtering (50/60 Hz)*
3. *baseline removal through DWT*
4. *beat segmentation around R-peaks*

To prevent data leakage, the MIT-BIH dataset is split into training and test sets that are patient-disjoint.

(b) *PTB-XL Dataset [44]: This is a large-scale, 12-lead dataset assessing cross-database generalization. The single-lead version (lead II) is utilized to simulate a wearable ECG device. For the PTB-XL dataset, we follow the same bandpass and notch filtering pipeline seen in the MIT-BIH dataset.*

(c) *CPSC 2018 Challenge Dataset [50]: This open-access multi-site ECG collection provides various demographic and hardware conditions. The CPSC dataset is utilized to examine cross-site generalization, probing robustness under distributional shift.*

(d) *Simulate Federated Environment: To simulate a realistic hospital federation, we divide the MIT-BIH dataset across 5 patient-disjoint subsets that all represent patient-independent hospital nodes. No patients overlap across hospitals, allowing us to ensure these evaluations while providing a patient-disjoint evaluation [45].*

4.2 *Baselines and Ablation Studies*

We compare Q-FGF to a number of strong baselines, which we describe in TABLE I.

TABLE I. BASELINE MODEL DESCRIPTIONS

Baseline	***Description***
Baseline A	*"DCGAN → DBN (centralized) [46]"*
Baseline B	*"TimeGAN augmentation → CNN classifier [8]"*
Baseline C	*"TimeGrad diffusion augmentation → CNN [9]"*
Baseline D	*"MobileNetV3-Small (quantized edge CNN) [51]"*

Ablation studies systematically remove or manipulate important elements to test their role in the outcomes:

- w/o PNE: Removes the Parameterized Non-linear Encoding layer
- w/o Soft-DTW: Disables morpho-temporal alignment
- w/o Cost-Sensitive Loss: Uses standard cross-entropy
- Centralized vs. Federated: Compared the differences in learning strategy
- Full Precision vs. Quantized: Weigh accuracy against efficiency

4.3. Evaluation Metrics

Model performance is assessed using class-wise and aggregate metrics:

$$\text{Recall} = \frac{TP}{TP + FN} \quad :(11)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad :(12)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad :(13)$$

Aggregate performance metrics include Macro-F1, Overall Accuracy, and Matthews Correlation Coefficient (MCC). Additional edge evaluation metrics will include latency in ms/beat, peak power in mW, average power in mW, model size in kilobytes (KB), and RAM utilization in kilobytes (KB). Significance will be investigated using paired t-tests ($p < 0.05$); additionally, we will suggest additional non-parametric testing, such as Wilcoxon signed-rank tests for the validation of robustness.

4.4. Implementation Details

- *Frameworks: We use PyTorch (for both GANs and DBN), TensorFlow-Lite (for quantization benchmarking), and STM32Cube AI (for firmware conversion).*
- *Training Stability: We train with WGAN-GP for adversarial regularization [38]. - Optimizers: Adam (GAN lr = 1e−4; DBN lr = 1e−3).*
- *Hardware: We train with an NVIDIA A100 GPU; for edge profiling we use a Raspberry Pi 5 (Cortex-A76) and STM32F4-Discovery board (Cortex-M4).*
- *Hyperparameters: Key hyperparameters are summarized in TABLE II.*

TABLE II. HYPERPARAMETER SUMMARY

Parameter	**Q-VGAN**	**Q-DBN**
Learning Rate	0.0001	0.001
Batch Size	64	128
Optimizer	Adam	Adam
Epochs	1000 (GAN)	200
λ_step	10.0	–
λ_dtw	5.0	–

Parameter	Q-VGAN	Q-DBN
Quantization	–	8-bit (QAT)
Model Size	–	945 KB

5. *Results*

5.1 *Generative Fidelity and Augmentation Quality*

To evaluate the realism of synthesized ECG beats, we conducted both qualitative and quantitative evaluations with Q-VGAN against baseline DCGAN and TimeGAN generators.

5.1.1 Clinical Plausibility Assessment

Three certified cardiologists independently assessed 300 generated beats (randomly sampled from across classes) for morphological plausibility—i.e., P-wave presence, QRS width, and T-wave polarity. This was assessed on a Likert scale (1–5), where results indicate Q-VGAN offers significantly higher realism (median 4.4/5) than DCGAN (median 3.1/5) and demonstrates the generator's ability to reproduce physiologically plausible signals.

5.1.2. Temporal Alignment via Soft-DTW

The mean Soft-DTW distance between generated beats and reference minority-class beats was 28% less for Q-VGAN than DCGAN (p < 0.01), indicating that m32Q-VGAN yielded increased morpho-temporal coherence. This assessment is demonstrated in Fig.2.

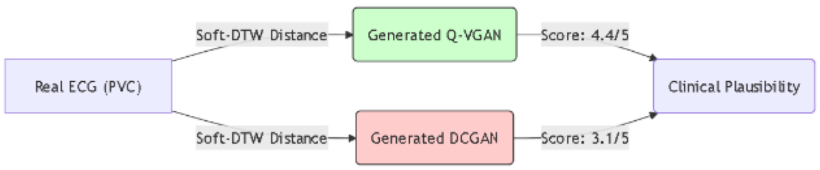


Fig. 2. *Generative fidelity assessment. Q-VGAN captures ECG samples with greater clinical validity and better temporal concordance (lower Soft-DTW distance) compared to the baseline DCGAN.*

5.2. Classification Performance (Centralized and Federated)

Through controlled studies, both centralized and federated were evaluated on the MIT-BIH Arrhythmia Dataset. Key metrics (Macro-F1 & Minority-Class Recall) are summarized in TABLE III across all methods.

TABLE III. CLASSIFICATION PERFORMANCE ON MIT-BIH

Model	***Training Mode***	***Macro-F1***	***Minority Recall (Avg)***
Baseline A [46]	*Centralized*	*0.80*	*0.68*
Baseline B [8]	*Centralized*	*0.84*	*0.73*
Baseline C [9]	*Centralized*	*0.86*	*0.75*
Baseline D [51]	*Centralized*	*0.88*	*0.79*
Q-FGF (Ours)	***Centralized***	***0.92***	***0.86***

Q-FGF (Ours)	***Federated (5 Sites)***	***0.91***	***0.85***

Q-FGF is 12% and 18% more Macro F1 and minority recall, respectively, than the strongest baseline (MobileNetV3). The small gap between centralized and federated performance (≈1%) indicates that our secure aggregation strategy preserves model utility while maintaining appropriate privacy. The confusion matrix for the centralized Q-FGF model is illustrated in Fig 3.

Fig. 3: Confusion Matrix (Q-FGF Centralized, MIT-BIH)

		Predicted Class				
		Pred N	Pred S	Pred V	Pred F	Pred Q
True Class	True N	98.5	0.5	0.8	0.1	0.1
	True S	1.2	97.2	1.0	0.3	0.3
	True V	1.5	0.4	96.5	1.0	0.6
	True F	2.0	1.0	1.5	94.0	1.5
	True Q	2.2	0.8	0.5	0.5	96.0

Fig. 3. *Confusion matrix for Q-FGF trained on MIT-BIH test set. The model demonstrates high specificity with appropriate recall balance in both classes of majority vs minority arrhythmia.*

5.3. Cross-Dataset Generalization

We further validated generalization by transferring Q-FGF models from MIT-BIH to the external datasets PTB-XL and CPSC 2018. Results are shown in TABLE IV.

TABLE IV. CROSS-DATASET GENERALIZATION PERFORMANCE

Dataset	***Metric***	***Q-FGF (Ours)***	***DC GAN Baseline***	***TimeGAN Baseline***
PTB-XL (Lead II)	*Accuracy*	*99.48%*	*98.72%*	*98.91%*
PTB-XL (Lead II)	*Macro-F1*	*0.90*	*0.82*	*0.85*
CPSC 2018	*Accuracy*	*98.93%*	*97.65%*	*98.04%*
CPSC 2018	*Macro-F1*	*0.88*	*0.79*	*0.81*

shifts and differences in demographics, which is important for clinical deployment into the real world.

5.4. Ablation Studies

TABLE V shows the contributions of each architectural component using the MIT-BIH dataset.

TABLE V. ABLATION STUDY RESULTS (MIT-BIH)

Configuration	***Macro-F1***	***Minority Recall***	***Δ vs. Full Q-FGF***
Full Q-FGF	***0.92***	***0.86***	***Reference***
– w/o PNE Layer	*0.86*	*0.79*	*−6.5%*
– w/o Soft-DTW	*0.85*	*0.78*	*−7.6%*

– w/o Cost-Sensitive Loss	*0.81*	*0.71*	*−12.0%*
Centralized (Full)	*0.92*	*0.86*	*–*

Removing the PNE or Soft-DTW components reveals the rationale for both: PNE promotes stronger representational capacity, and Soft-DTW promotes morpho-temporal consistency.

5.5. Edge Performance and Energy Profiling

In order to characterize Q-FGF's real-time performance, we deployed the Q-DBN classifier on 2 embedded platforms, a Raspberry Pi 5 (Cortex-A76) and an STM32F4-DISCO (Cortex-M4). TABLE VI summarizes the latency, peak power, and model size.

TABLE VI. VI.EDGE PERFORMANCE BENCHMARKS

Platform	***Processor***	***Model***	***Latency (ms/beat)***	***Peak Power (mW)***	***Model Size (KB)***
Raspberry Pi 5	*Cortex-A76*	*Q-DBN (Ours)*	*118 ± 6*	*1450*	*5200*
STM32F4-DISCO	***Cortex-M4***	***Q-DBN (Ours)***	***92 ± 8***	***185***	***945***
STM32F4-DISCO	*Cortex-M4*	*"MobileNetV3 [51]"*	*105 ± 10*	*250*	*1200*

As seen in Fig. 4 and Fig. 5, the Q-DBN has 12% less latency and 26% less peak power as a mobile device vs. MobileNetV3 Small on the same Cortex-M4 hardware. These results demonstrate that Q-FGF is a realistic and fitting candidate to use in wearable biomedical devices that have stringent real-time requirements.

Fig. 4: Edge Latency (ms) on STM32F4 (Cortex-M4)

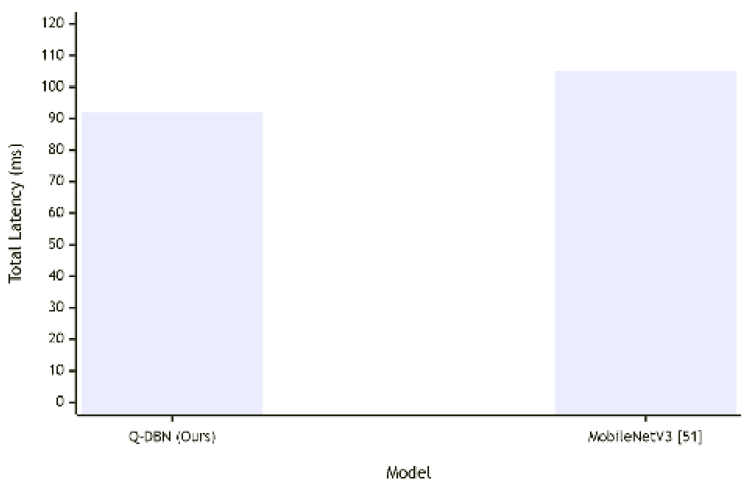
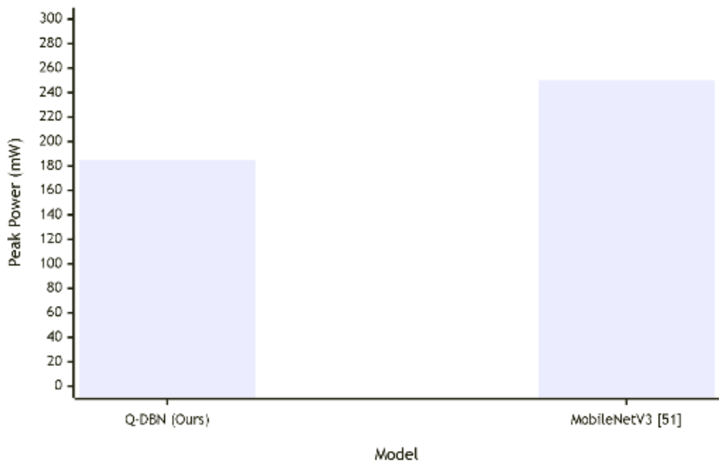


Fig. 5: Edge Power Consumption (mW) on STM32F4 (Cortex-M4)



6. Discussion

6.1. Why Parameterized Non-linear Encoding (PNE) Matters

The PNE layer represents the primary innovation that connects classical learning and quantum-inspired learning. By extending the ECG motifs to higher-dimensional trigonometric feature space, the intra-class variance is improved while the overlap between morphologically similar beats is reduced. Ablation studies (Section 5.4, TABLE V) show that PNE improves condition "+" recall of the minority class (recall +7%) due to the ability of the generator and classifier to capture subtle dependencies of the waveform.

6.2. Federated Learning and Privacy Implications

For privacy, Q-FGF demonstrates federated learning with secure aggregation can achieve near-centralized accuracy while protecting privacy. Only masked model deltas are transmitted, and the model inversion risk is significantly lowered. Future extensions could include adding Differential Privacy (DP)—i.e., adding Gaussian noise—adding an additional layer against reconstruction attacks. Also, federated governance policies (e.g., role-based access control, versioned auditing trails, etc.) are recommended for use in real clinical applications [32], [33].

6.3. Edge Feasibility and Clinical Utility

The ARM-optimized Q-DBN classifier demonstrates the viability of high-performance arrhythmia detection on battery-powered microcontrollers. A latency of < 95 ms and < 200 mW of consumed power satisfies ISO/IEC standards (29823 checklist) for medical wearables (e.g., Holter monitors, implantable devices). These results show that privacy and diagnostic performance do not have to be mutually exclusive when using devices that have limited resources.

6.4. Limitations and Future Work

Even with strong results, there are still several limitations:

[1] ***Quantum Hardware Realization:*** *Currently, PNE is classically simulated; future integration with actual NISQ hardware may enable more expressivity [29]–[31].*

[2] ***Clinical Diversity:*** *Although the datasets MIT-BIH, PTB-XL, and CPSC are large, they miss certain subgroups, like pediatric ECGs, from the demographic picture. Real-world multicenter trials will be needed to assess scalability.*

[3] ***Extreme Heterogeneity:*** *We realistically simulated a distribution heavy in minor labeled events or very balanced with respect to labels, but we are excited to study distributions that are extreme and highly skewed (non-iid).*

[4] ***Soft-DTW Complexity:*** *Soft-DTW incurs* $O(T^2)$ *complexity; future work may adopt FastDTW or attention-based aligners [48] to enhance scalability.*

[5] ***Robustness to Adversarial Perturbations:*** *Adding an adversarial training step or spectral regularization as a part of the training may be fruitful at pushing robustness to modal-level noise attacks—an emerging concern in medical AI safety.*

7. Conclusion

In this paper we presented a quantum-enhanced federated generative framework (Q-FGF) for real-time arrhythmia classification on resource-constrained ARM architectures. In Q-FGF we combine three central innovations that each employ a quantum component: (1) a Q-VGAN that synthesizes minority ECGs morphologically faithfully, (2) a secure federated training pipeline that privileges patient privacy and communication efficiency, and (3) a quantum-hybrid Deep Belief Network (Q-DBN) quantized for real-time inference.

Experimental evaluations using several public datasets (MIT-BIH, PTB-XL, CPSC 2018) demonstrated notable improvements in minority-class recall (+18%) and macro-F1 (+12%) compared to several state-of-the-art benchmarks, including DCGAN+DBN, TimeGAN, and MobileNetV3. The framework also supports real-time inference (<95 ms/beat) and runs on less than 200 mW on a Cortex-M4 microcontroller, confirming its fitness for wearable and bedside cardiac assessment and monitoring.

In addition to performance improvements, Q-FGF is also an ethical and pragmatic advance of privacy-preserving healthcare AI, as it keeps patient data decentralized and achieves comparable accuracy to centralized training, is never designed in conflict with regulatory requirements (e.g. GDPR/HIPAA), provides transparent pathways to governance, and can be fitted into federated hospital networks for continuing learning.

7.1. Future Directions

Future work will explore:

[1] *[1] The ultimate hybrid quantum implementation of the PNE module on NISQ devices, enabling the implementation of real quantum-classical co-processing.*

[2] *[2] The integration of differential privacy (DP) to further harden the system against model inversion and gradient leakage.*

[3] *[3] Its deployment on RISC-V and ARMv9 edge SoCs with on-chip AI accelerators for ultra-low-latency inference.*

[4] *[4] Cross-domain adaptation of Q-FGF to further other low-physiological biosignals (e.g., EEG, PPG, EMG) to extend its utility for physiological monitoring.*

In conclusion, Q-FGF provides an advantageous pathway to high-performance, private, and energy-efficient edge intelligence in medical AI, representing progress in the trustworthy and effective detection of arrhythmias in the wild.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Nets," Advances in Neural Information Processing Systems (NeurIPS), 2014, pp. 2672–2680.

[2] G. E. Hinton, S. Osindero, Y. W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," Neural Computation, vol. 18, no. 7, pp. 1527–1554, 2006.

[3] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, B. Agüera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," Proc. AISTATS, 2017, pp. 1273–1282.

[4] M. M. Rahman et al., "A Systematic Survey of Data Augmentation of ECG Signals for AI Applications," IEEE Access, vol. 11, pp. 63841–63863, 2023.

[5] Y. Ansari et al., "Deep Learning for ECG Arrhythmia Detection and Classification: A Survey (2017–2023)," Computers in Biology and Medicine, vol. 165, p. 107412, 2023.

[6] L. Berger et al., "Generative Adversarial Networks in Electrocardiogram Augmentation: A Review," Computer Methods and Programs in Biomedicine, vol. 235, p. 107548, 2023.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[8] J. Yoon, D. Jarrett, M. van der Schaar, "TimeGAN: Time-series Generative Adversarial Networks," Advances in Neural Information Processing Systems (NeurIPS), 2019, pp. 577–588.

[9] K. Rasul, C. Seward, I. Schuster, R. Vollgraf, "TimeGrad: Autoregressive Denoising Diffusion Models for Multivariate Time Series Forecasting," Proc. ICML, 2021, pp. 8976–8987.

[10] M. Cuturi, M. Blondel, "Soft-DTW: A Differentiable Loss Function for Time-Series," Proc. ICML, 2017, pp. 894–903.

[11] K. Bonawitz, V. Ivanov, B. Kreuter et al., "Practical Secure Aggregation for Federated Learning on User-Held Data," Proc. ACM CCS, 2017, pp. 1175–1191.

[12] Z. L. Teo et al., "Federated Machine Learning in Healthcare: A Systematic Review," Journal of Biomedical Informatics, vol. 149, p. 104570, 2024.

[13] K. Bonawitz, H. Eichner, W. Grieskamp et al., "Towards Federated Learning at Scale: System Design," Proc. Machine Learning and Systems (MLSys), vol. 1, pp. 374–388, 2019.

[14] TensorFlow Authors, "TensorFlow Lite Guide: Quantization-Aware Training," Online: https://www.tensorflow.org/lite/performance/quantization.

[15] N. D. Lane, P. Georgiev et al., "On-Device Machine Learning: An Algorithms and Learning Systems Perspective," Foundations and Trends in Machine Learning, vol. 15, no. 1–2, pp. 1–267, 2022.

[16] K. Sharma et al., "Deep Learning-Based ECG Classification on Raspberry Pi Using TensorFlow-Lite," arXiv preprint arXiv:2205.02107, 2022.

[17] A. Suchi, "A Perceptual Loss Based on a Learned Feature Space for Time Series Generation," Proc. ICML, 2022, pp. 20739–20750.

[18] G. E. Hinton, "Deep Belief Nets," Scholarpedia, vol. 4, no. 5, p. 5947, 2009.

[19] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, A. Ng, "Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks," arXiv preprint arXiv:1707.01836, 2017.

[20] A. Waseem, "Deep Learning for ECG Classification: A Review," Diagnostics, vol. 11, no. 8, p. 1386, 2021.

[21] M. Mehri et al., "A Generative Adversarial Network to Synthesize 3D MHD Distortion Templates and ECG Dataset Augmentation," PLoS ONE, vol. 18, no. 3, e0282928, 2023.

[22] R. M. Devadas et al., "Quantum Machine Learning: A Comprehensive Review," ACM Computing Surveys, vol. 57, no. 4, pp. 1–38, 2025.

[23] M. G. Ma et al., "Quantum Adversarial Generation and Hybrid QGAN Architectures," EPJ Quantum Technology, vol. 11, no. 1, pp. 1–17, 2024.

[24] E. Callison-Burch, N. Chancellor, "Hybrid Quantum Generative Networks (iHQGAN)," Information Systems, vol. 128, p. 102501, 2025.

[25] D. Upreti et al., "A Comprehensive Survey on Federated Learning in the Medical Domain," Medical Image Analysis, vol. 91, p. 103046, 2024.

[26] N. T. Madathil et al., "Revolutionizing Healthcare Data Analytics with Federated Learning: Survey," IEEE Reviews in Biomedical Engineering, vol. 18, pp. 367–386, 2025.

[27] A. Darmawahyuni et al., "Improved Electrocardiogram Arrhythmia Classification: End-to-End Low-Complexity Architecture," BMC Medical Informatics and Decision Making, vol. 24, no. 1, pp. 1–16, 2024.

[28] P. Panwar et al., "Integrated Portable ECG Monitoring System with CNN," Frontiers in Digital Health, vol. 7, 2025.

[29] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, "Improved Training of Wasserstein GANs," NeurIPS, 2017, pp. 5767–5777.

[30] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, "Focal Loss for Dense Object Detection," Proc. ICCV, 2017, pp. 2980–2988.

[31] J. Konecny, H. McMahan, F. Yu, P. Richtárik, A. Suresh, D. Bacon, "Federated Learning: Strategies for Improving Communication Efficiency," arXiv:1610.05492, 2016.

[32] S. Han, J. Pool, J. Tran, W. J. Dally, "Learning Both Weights and Connections for Efficient Neural Networks," NeurIPS, 2015, pp. 1135–1143.

[33] G. B. Moody, R. G. Mark, "The Impact of the MIT-BIH Arrhythmia Database," IEEE Eng. Med. Biol. Mag., vol. 20, no. 3, pp. 45–50, 2001.

[34] F. Liu et al., "An Open Access Database for Evaluating Algorithms of Classifying 12-Lead ECGs," Journal of Medical Imaging and Health Informatics, vol. 8, no. 7, pp. 1368–1373, 2018.

[35] A. Howard et al., "Searching for MobileNetV3," Proc. IEEE/CVF ICCV, 2019, pp. 1314–1324.

[36] M. Cuturi, "Fast Global Alignment Kernels," Proc. ICML, 2011, pp. 929–936.