Minimax Theory and its Applications

Volume 11 (2026), No. 4s, 1042-1054

TinyFed-XAI: Explainable Federated TinyML for Early Arrhythmia Detection on Wearable IoT Devices

Dr.Araddhana M A Deshmukh[1], Dr Sowmya M S[2], Dr. Tummala Purnima[3], Dr K Sridhar[4], R Mohemmed Yousuf[5], Dr.A.Devendran[6], K.Malarvizhi[7]

[1]My affiliation Dr.Araddhana Manisha Arvind Deshmukh, Professor, School of Computing, MIT Art, Design & Technology University, Loni Kalbhor, Pune, India, Email: Aaradhna.Deshmukh@mituniversity.edu.in
[2]Associate Professor, Department of CSE, Faculty of Engineering and Technology, Jain (Deemed to be University),Jain Global Campus, Jakkasandra, Bangalore South District, 562112, Email: mssowmya.sbmjce@gmail.com, Orcid id: 0000-0002-2254-1938
[3]Assistant Professor, Department of Computer science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram-522302, India. Email: tpurnima78@gmail.com, orcid id: 0000-0001-7051-4793
[4]Associate Professor, School of CSE, Department of CSE, Malla Reddy (MR) Deemed to be University, Telangana State - 500100 India, Email: sridhark529reddy@gmail.com, Orcid id: 0009-0007-6850-8731
[5]Assistant Professor, Department of Information Technology, Bannari Amman Institute of Technology, Erode, Sathyamangalam. Email: mohemmedyousuf@bitsathy.ac.in,Orcid id:0009-0003-8172-1829
[6]Associate Professor, Department of Data Science, Sri Ramachandra Institute of Higher Education and Research (SRIHER) (Deemed to be University) Chennai– 600 116,, Tamil Nadu, India. Email: Devendran.alagarsamy@gmail.com,Orcid id:0000-0001-7192-521X
[7]Assistant Professor, Computer Science and Engineering , Akshaya college of engineering and technology Email: malarcbe94@gmail.com,Orcid id: 0009-0004-5788-0413

Abstract
Wearable Internet of Things (IoT) devices are promising tools to help detect fatal cardiac arrhythmia, which is a common cause of sudden heart events. The challenge still exists when deploying deep learning models on resource-limited wearables without compromising the patient's data privacy. The designed framework called TinyFed-XAI, helps to establish the relationship between Federated Learning (FL), TinyML compression, and Explainable AI (XAI) is developed in this paper. Using this, early detection of arrhythmia in wearable devices can be possible. Utilizing the MIT-BIH along with PTB derived ECG heartbeat dataset to simulate a realistic non-IID deployment, we take the help of Dirichlet-based partitioning to carry out training on ten clients. We use a lightweight (7,013 parameters) 1D Convolutional Neural Network (CNN) trained through Federated Averaging (FedAvg). In order to tackle the big class imbalance problem in the arrhythmia data, we consider a global class-weighting strategy in combination with extended local training. By using this combination, we can boost the Macro F1-score (federated model) from 0.446 to 0.538, while keeping the classification accuracy at 85.3%. The post-training INT8 quantization compresses the trained model and reduces its size from 33.8 KB to 17.2 KB with a negligible loss of accuracy confirming it can be deployed on microcontroller-class wearable hardware. Ultimately, SHAP was applied to the federated and centralized models to explain classification decisions at the waveform level. This shows that both models pay attention to clinically meaningful parts of ECG such as the QRS complex, and the P-wave shape. The trade-offs between federated and centralized training are analysed. The federated model achieves higher accuracy on the majority classes while the centralized model performs better than the federated model on minority arrhythmias, with Macro F1 of 0.592, highlighting a long-standing challenge of federated learning under extreme non-IID and long-tailed class distribution. Our results show that wearable IoT arrhythmia detection can be privacy-preserving, computationally efficient and interpretable. Moreover, we identify rare-class representation as a key open challenge for future federated healthcare applications.

Keywords*:* Federated Learning, TinyML, Explainable AI (XAI), Arrhythmia Detection, Wearable IoT, Edge Intelligence, ECG Classification, Model Quantization, SHAP, Non-IID Data.

1. Introduction

The leading cause of human death worldwide is cardiovascular diseases. Death by Cardiac Arrhythmias is Huge After Heart Life-threatening conditions may arise from heart rhythm irregularities. It is essential to detect arrhythmias early in time and continuously as many life-threatening episodes like ventricular fibrillation can occur with no warning and require immediate treatment. The most common method of arrhythmia diagnosis is

through monitoring of the ECG. At present, this occurs in the clinic using Holter monitors or in hospital using hospital-grade ECG devices. These devices are only good for limited periods of monitoring and cannot be easily used for continuous monitoring for longer durations. There has been a growing use of wearable internet of things (IoT) devices like smartwatches and chest straps as well as patch-based ECG sensors for monitoring the heart in real-time outside of hospitals. In fact, it even has the capability now to detect abnormalities during everyday life.The potential of deploying deep learning-based arrhythmia detection models on wearable IoTs is fraught with three interrelated challenges.

Wearable devices possess very limited capacity due to small size. As a result, they will be unable to deploy deep learning architectures designed for inference on servers. Also, ECG data is sensitive personal health information. Training models on these data that are centralized from a high number of users can lead to serious privacy and compliance trouble under HIPAA, GDPR, and more. In addition, models for classifying arrhythmias are typically treated as a black box, providing minimal understanding of the physiological basis of their output. Interpretable systems can be important for healthcare applications when automated diagnosis is implemented in an existing context.

The existing research has studied these issues in isolation. TinyML techniques including model quantization and pruning can help compress neural networks for deployment on microcontroller-class devices. Nevertheless, much of this work was evaluated under centralized training assumptions that often ignore data privacy constraints.

Federated learning (FL) is gaining more and more popularity to train models without exchanging raw data over clients. Although FL is being utilized in healthcare tasks with increasing frequency, FL usage for arrhythmia detection typically makes IID assumption across clients. This assumption is not valid in the real wearable deployment of arrhythmia detection. Different patients will display different types of arrhythmias – which will be quite a few and highly imbalanced. In addition, few existing works leverage federated training and post-hoc explainability together. As a result, we do not fully understand how the nature of the learning context impacts interpretability under decentralized resource-constrained and non-IID settings.

To bridge this gap, we put forth the coalesced architecture TinyFed-XAI which exploits Federated Learning, TinyML model compression, and Explainable AI (XAI) for early arrhythmia on wearable IoT devices. In contrast to prior research, which examined various components in isolation, TinyFed-XAI is assessed in its whole, under realistic settings. The data in TinyFed-XAI is non-IID (non-Independent, and Identically Distributed) across simulated wearable clients and exhibits an extreme class imbalance that emulates the real-world, long-tailed prevalence of arrhythmia. It further employs a lightweight convolutional architecture deployable on microcontroller devices.

As illustrated in this section, they will be able to evaluate the contribution of decentralised training towards the clinical plausibility of the decisions made by a model that is already trained with federated and centralized variants.

The paper features major contributions which are summarized as follows,

1. Our training pipeline for arrhythmia classification is based on federated averaging and implements realistic non-IID data distribution. The non-IID distribution is simulated across ten wearable client devices using Dirichlet based partitioning.
2. To and obtain effective and on-class abstraction of global approach was extended local stops. To a remedy compounded effect of lets class imbalance and data heterogeneity federation arrhythmia detection we proposed and shown with experiments and limitations particularly strongly rare arrhythmia type.
3. We utilize post-training INT8 quantization to compress the trained model, resulting in a significant decrease in model size while maintaining classification performance, and show it is feasible for deployment on microcontroller-class wearable hardware.
4. We perform SHAP-based analysis for explaining a model's prediction at ECG waveform level, and provide comparison of feature attribution patterns in the federated and centrally trained models.
5. We present an extensive empirical evaluation of federated versus centralized training in terms of accuracy, per-class performance, model compression and interpretability, laying out the practical trade-offs relevant for federated healthcare applications in practice.

The rest of the paper devised as follows; Section II reviews related work on Federated Learning, TinyML, explainable AI for Healthcare, and arrhythmia detection. The TinyFed-XAI framework is presented in Section 3 along with data distribution, model structure, federated training process, and explainability pipeline. Section 4 presents experimental results in conjunction with their comparative analysis. Section 5 states the implications, limitations and open challenges exposed by this study, particularly the enduring issue of learning extremely rare classes under federated non-IID settings. The paper concludes in section 6 and gives future directions.

2. Related Work

2.1 Federated Learning for ECG-Based Arrhythmia Detection

In recent years, researchers have extensively studied federated learning as a means to safeguard privacy while training ECG models in a decentralized manner. Zhang et al. [16] proposed a distributed arrhythmia detection algorithm using federated learning since the ECG data of patients are sensitive and private like other medical

data coupled with the necessity for hospitals and medical institutions to train models jointly without transferring the original patient data. This work gave a formal treatment to non-IID ECG in the FL literature earliest and one of the first. Following on from this work, subsequent studies have reported strong diagnostic performance when combining personalized federated learning and differential privacy [3]. In particular, the work compares a personalized federated learning model against CNN, ResNet and attention-based RNN baselines and it showed improved robustness under non-IID conditions. A few more recent works surveyed the trade-offs between privacy, security and model performance in a FL system [5]. An interesting open research direction is finding scalable and adaptive solutions for heterogeneous FL environments. Complementary empirical work [4] reveals that moderate non-IID conditions may only cause marginal degradation of accuracy for IID settings. However, extreme non-IID distributions may induce local model divergence, requiring multiple rounds of federation to stabilize. Most relevant to the present work, recent studies [11] show that ECG data collected across IoMT devices is not only non-IID, but also long-tailed, with normal sinus rhythms dominating while many arrhythmia subtypes occur sparsely a condition under which naive aggregation strategies are biased to the majority class and provide weak supervision for the minority or absent class at individual clients. Similar broader reviews of FL applications [9] considered the prediction of cardiovascular diseases have noted that bio signal-based federated learning continues to face persistent challenges of non-IID data distributions, limited labelled data, sensor heterogeneity, with client clustering, personalized training and generative augmentation all being proposed as partial mitigations.

2.2 TinyML for On-Device ECG Classification

Another branch of research has studied the model compression for arrhythmia detection targeted for microcontroller-class hardware. Recent viability studies [15] have showcased the implementation of quantized 1D-CNN models on extremely inexpensive microcontrollers like Arduino UNO, which includes ECG acquisition and preprocessing as well as real-time classification within a memory footprint of less than 24 KB. Other work [7] targeting five-class ECG rhythm classification on embedded devices has used a combination of handcrafted P-, R, and S-wave feature extraction along with compact CNN-DNN architectures, achieving high accuracy at the model size of 143 Kb. Methods leveraging autoencoders [8] have also been investigated for on-device anomaly detection. Another branch focused on repeated pessimizations, INT8 post-training, and quantization-aware optimization for microcontroller deployment to reduce memory footprint and enhance inference speed and efficiency. In general, quantization-aware training has been found to recover more accuracy loss than post-training quantization alone. The research on hierarchical federated learning [14] has been performed on sub-5KB CNN-LSTM models for ultra-resource-constrained wearables. This is largely due to the domain knowledge of ECG morphology to classify an abnormal cardiac signal using only a few hundred parameters. In totality, these works demonstrate that it is technically feasible to deploy arrhythmia classifiers on wearable-class hardware but most of these works assume centralized training and thus do not discuss any privacy issues or data-heterogeneity issues that would arise in a real-world multi-patient deployment.

2.3 Explainable AI for ECG Interpretation

As a result of a hesitance to deploy black-box arrhythmia classifiers in the clinic, interpretability has become an important consideration. A recent user-centred study [2] demonstrated that clinicians preferred saliency-map based explanations to counterfactual visualizations, citing clearer correspondence with established workflows for ECG interpretation. The study compared permutation importance, KernelSHAP, gradient-based methods and DeepLIFT. Study authors found that gradient-based and DeepLIFT methods generated explanations that were consistent with clinical reasoning while permutation and KernelSHAP approaches produced cluttered outputs. Other research [6] has deployed SHAP to detect ventricular ectopic beats, supraventricular ectopic beats and atrial fibrillation in LightGBM and 1D-CNN architectures, oversampling to address class imbalance. Larger reviews of methods of XAI for ECG classification [13] have documented a wide variety of interpretability methods, including Grad-CAM, Layer-wise Relevance Propagation, Integrated Gradients, and Deep Taylor Decomposition. Along with this, comparative studies [10] have shown that trustworthiness is highly variant across these methods. Recent models [1] that integrate SHAP into their multi-label ECG classification pipeline show that SHAP-based attribution yields model decisions that agree with physiologically relevant waveform segments from ECG leads that help clinically in the diagnosis while preserving diagnostic accuracy. Likewise, it was reported that comprehensive frameworks [13] that integrate oversampling-based class balancing with SHAP, LIME, and feature-importance analyses indicate that SHAP measures the marginal contribution of individual segments of the ECG signals towards a given prediction. Moreover, it provides local explanations for the predicted classes of arrhythmia classifications consistently with fairness.

2.4 Research Gap

Though these three areas of research: federated learning, TinyML compression and explainable AI have been used for ECG-based arrhythmia detection, they have not been leveraged together in a single end-to-end evaluated framework. Federated learning studies on ECG data that is not independent and identically distributed typically assess the model's accuracy [4]. They do not take into account interpretability for deployment downstream constraints. Much of the TinyML literature assumes a centralized training process [7]. However, for

wearable deployment across multiple patients, privacy considerations motivate the need for federated approaches. The studies of explainable artificial intelligence are essentially all on centrally trained models [6] and it is unknown how decentralized non-IID training would affect the clinical plausibility of model explanations. This is a relevant question given the fact that the long-tailed distribution of classes biases federated aggregation towards majority classes [11]. TinyFed-XAI is created with the aim to fill this void, getting jointly assessed in a single non-IID, class-imbalanced experimental pipeline for federated training, INT8 quantization for microcontroller-class deployment, and SHAP-based explainability and directly comparing the interpretability characteristics with a centrally-trained counterpart.

2. Methodology

The proposed TinyFed-XAI framework consists of five stages. Stage application includes,

(i) data acquisition and pre-processing
(ii) non-IID partitioning of dataset to replicate realistic wearable client distribution
(iii) federated model training using FedAvg while mitigating class-imbalance issue
(iv) quantizing the model after training for deployment on microcontroller-class device
(v) using XAI tools for explainability analysis.

3.1 Dataset

The MIT-BIH Arrhythmia Database and the PTB Diagnostic ECG Database provided the ECG Heartbeat Classification Dataset which is freely available. It is obtained from Kaggle. Each sample represents a beat which is segmented and resampled to 187 time-steps at 125 Hz In addition, the amplitude is scaled to [0, 1]. The whole dataset consists of 109446 samples of which 87554 are training samples and 21892 are test samples. The entire dataset is classified into five diagnostic classes as per AAMI EC57 Standard Normal(N) Supraventricular ectopic beat(S) Ventricular ectopic beat(V) Fusion beat(F) and Unknown/Paced beat(Q). The dataset is quite imbalanced resembling real-world arrhythmia distribution. From the training, we can see that Normal beat accounts for 82.8% (72,471 samples), while Fusion beat only 0.7% (641 samples). This class imbalance motivates our class-imbalance mitigation strategy which we will describe in Section 3.5.

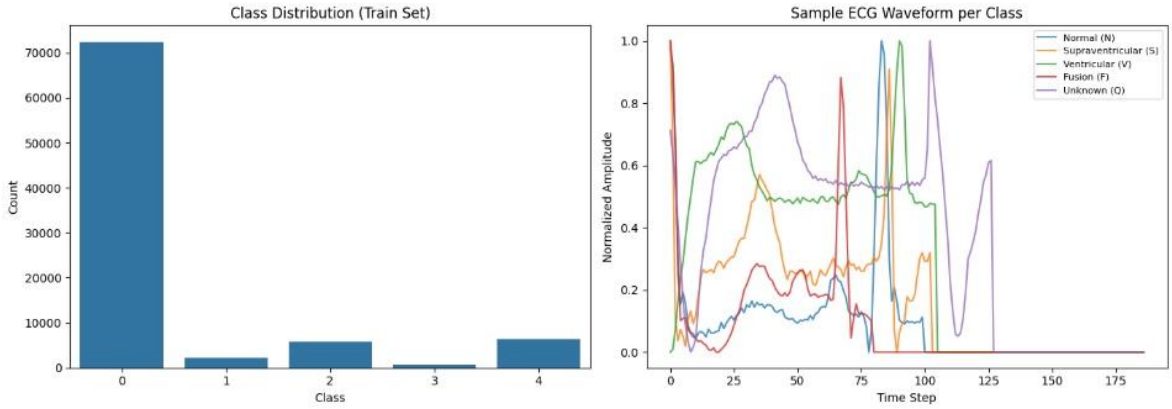

Figure 1. Class distribution of the training set and representative ECG waveforms per arrhythmia class

3.2 Non-IID Client Partitioning

In order to capture a realistic federated wearable IoT deployment, the training set has been partitioned across ten simulated client devices, which represent a patient/wearable sensor. Instead of presuming an independent and identically distributed (IID) split an assumption that is seldom valid in real-world multi-patient ECG monitoring a Dirichlet distribution (concentration parameter $\alpha = 0.5$) was used to control the degree of label heterogeneity across clients. This follows the standard practice in the federated learning literature. For each class c, we sampled a proportion vector from Dir(α) to determine the fractions of class c's samples assigned to each client. The intention was to create a purposely skewed, non-IID distribution to ensure that the class compositions of individual clients varied significantly. One of the clients received 78% of all samples belonging to the class Normal. Similarly, one or two clients were heavily populated with Unknown or Ventricular beats, and several clients received zero or near-zero samples of the rarest classes, namely Supraventricular, Fusion. The local data from each client is further divided into a training and validation subset. We take 85% of the samples for training and 15% for validation. This is performed using stratified sampling first, where we ensure that every class has a minimum of two samples in every client. We resort to random sampling otherwise, when the class has a singleton in a client.

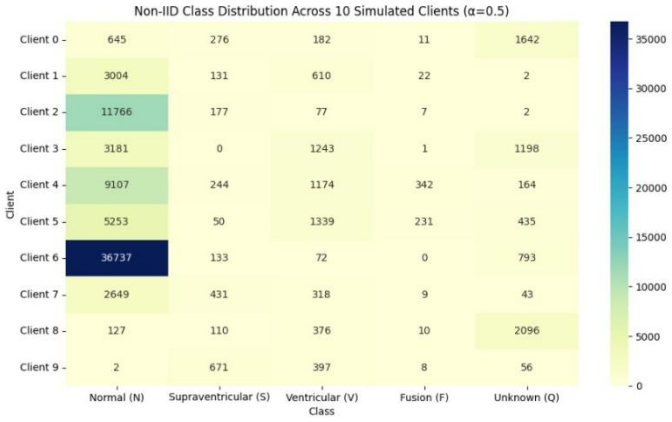

Figure 2. Non-IID class distribution across 10 simulated federated clients (Dirichlet partitioning, α = 0.5)

3.3 Model Architecture

A lightweight 1D-CNN architecture that meets TinyML constraints but is still powerful enough to classify five classes of heartbeat signal is designed in this work. The convolutional blocks of architecture have been designed with 16, 32, 32 filters respectively and filter size 5,5,3 respectively, with max-pooling after first two blocks. Finally, global average pooling, a fully connected layer of 32 units (ReLU + drop out, rate = 0.3 for regularization) and SoftMax output layer over five classes. The parameters of the model end up being 7013, resulting in the model having a floating-point memory footprint of ~27.4 KB. The 256 KB to 1 MB SRAM memory requirement is well within the capabilities of a typical microcontroller of wearable class, i.e., ARM Cortex M series.The Adam optimizer (learning rate = 0.001) with a sparse categorical cross-entropy loss was used to compile the model.

3.4 Federated Training via FedAvg

The model was trained using a Federated Averaging (FedAvg) procedure that was manually implemented in order to ensure full transparency. In every round of communication, the latest global model weights were passed along to each of the ten clients, all of which carried out local training for a fixed number of epochs on their own partitions before sending their updated weights back to a central aggregator. The server subsequently calculated a new global model by using all client weights as a weighted average with sample size as the weight.

$$w_{global}^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{N} w_k^{(t)}$$

where $w_k^{(t)}$ denotes the local model weights of client k after round t, n_k is the number of training samples held by client k, and $N = \sum_k n_k$ is the total sample count across all participating clients. This weighting scheme ensures that clients with larger local datasets have a more pronounced effect on the global model, as per the original FedAvg. Over the 15 rounds of communication for training, 2-4 local epochs (varied per experimental configuration, detailed in section 3.5) were performed by each client per round with a local batch size of 64. The global model was evaluated on the centrally held-out test set after every round to check convergence.

3.5 Class Imbalance Mitigation

The first experiments conducted with unweighted cross-entropy loss showed that both models, federated and centralized, achieved high overall accuracy (greater than 88%) while completely missing the minority arrhythmia classes. Notably, Supraventricular and Fusion beats achieved a recall of 0%, despite their clinical significance. To rectify this, class weights were calculated using the inverse-frequency ("balanced") heuristic.

$$w_c = \frac{N}{K \cdot n_c}$$

where N is the total number of training samples, K is the number of classes, and n_c is the number of samples belonging to class c. Importantly, the weights in this model were computed once from the global training distribution and were applied identically across all clients during local training. This process was different from using local class distributions or recomputing per client (which caused instability). An earlier variant that used per-client weight computation produced highly unstable, oscillating global model performance. The oscillating patterns arose since clients with extremely sparse local representation of a class computed disproportionately large local weights. When those weights were averaged across clients, they destructively interfered with convergence. We apply an additional fixed amplification factor of 2.5× on top of the balanced weight for the maximally underrepresented classes (Supraventricular and Fusion). The reason for this is that we want to counter-vary the dilution effect caused by FedAvg's weight-averaging on classes that are sparsely represented across clients. Also, increase the local training epochs per round (from 2 to 4) to allow local gradient signals for rare classes greater influence before aggregation. The centralized baseline model trained on the pooled data was

subject to the same global class weighting (without federated averaging). This enables a fair comparison of federated versus centralized learning under the same imbalance-mitigation conditions.

3.6 TinyML Model Compression

In order to assess whether our model is deployable on wearable IoT hardware, our final federated model and the centralized model were converted to TFLite under three different quantization settings. These settings are:

(i) baseline without quantization (which we also converted using TFLite for standardized latency measurement),

(ii) dynamic-range quantization, in which case model weights are converted to 8-bit integers while activations remain in floating point, and

(iii) full integer (INT8) quantization, in which case both weights and activations are quantized using the statistics from a representative calibration dataset comprising of 200 training samples to compute appropriate quantization scale factors.

Each quantized model was evaluated on the complete held-out test set (21,892 samples) in terms of classification accuracy and macro F1-score. Additionally, we benchmarked the per-sample inference latency of the models on a random sample of 500 test samples using the TFLite Interpreter API.

3.7 Explainability Analysis

In order to interpret model predictions at the ECG waveform level, we applied SHAP (SHapley Additive exPlanations) to both the federated and centralized trained models with the Gradient Explainer variant, which is computationally efficient for differentiable neural network architectures (in our case CNNs). To establish the baseline expectation for SHAP value computation, the background reference set comprises 100 randomly sampled training instances. To obtain per-time-step SHAP attribution of the ECG signals, we selected the appropriately-classified test samples for each class (where applicable). These attributions are shown in the waveforms coloured in red and blue.Regions of the 187 time-step-long ECG segment are shown to have contributed positively and negatively towards the predicted class respectively. The global feature importance was also calculated by averaging the absolute SHAP values for all explained samples per class. This provides a global importance profile across the ECG time-step axis. This independent analysis of the federated and centralized models enables direct comparisons of how decentralized, non-IID training impacts the spatial distribution and clinical plausibility of model attributions.

3.8 Evaluation Metrics

The overall classification accuracy, the precision, recall, F1-score per class, and macro-averaged F1-score (unweighted mean of F1-scores per class) were used to assess model performance. The macro-averaged F1-score was used as the primary metric to compare model behaviour under class imbalance since the overall accuracy was dominated by the majority Normal class and thus doesn't reflect poor performance on clinically relevant minority arrhythmias. Confusion matrices were calculated to visualize class-wise misclassifications. As part of the evaluation for TinyML, in kilobytes, we also reported model size and inference latency averaged per sample in milliseconds to characterise deployment feasibility on resource-constrained wearable hardware.

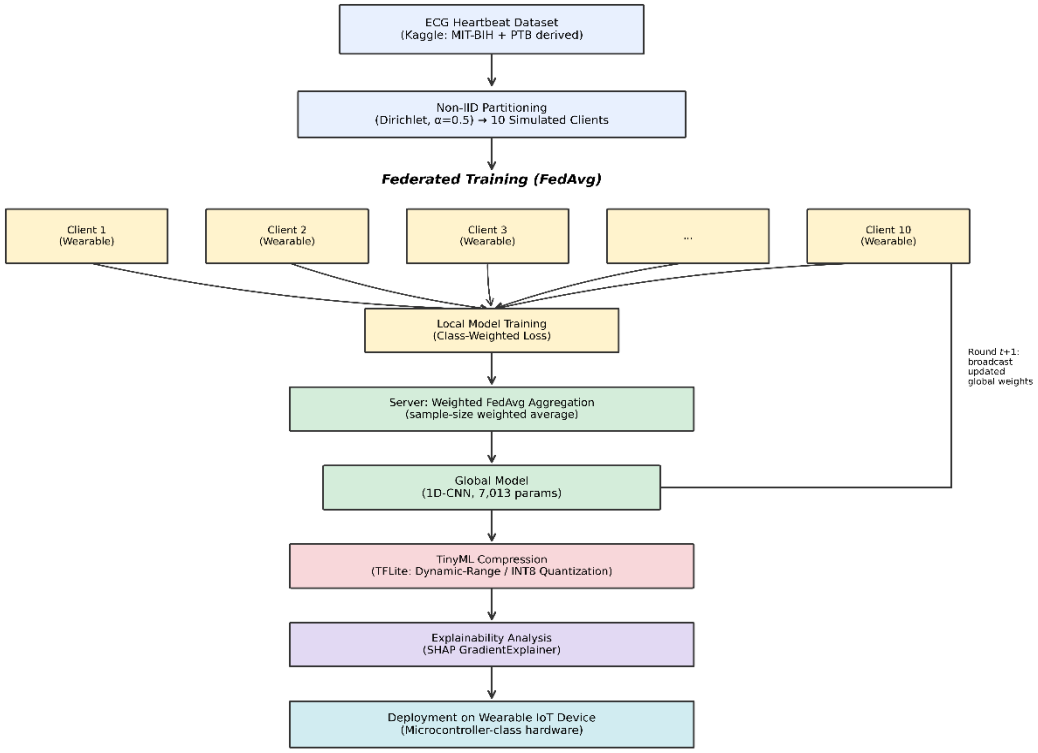

Figure 3. System architecture of the proposed TinyFed-XAI framework

3. Results

This section presents the experimental results obtained from the TinyFed-XAI pipeline, organized into four parts:

1. baseline federated versus centralized performance,
2. the effect of class-imbalance mitigation strategies,
3. TinyML compression outcomes, and
4. SHAP-based explainability findings.

4.1 Baseline Federated vs. Centralized Performance

The unweighted cross-entropy loss trained the federated model for 15 communication rounds. The model accuracy steadily converged from 82.8% to 91.1% on the test set with a monotonic decrease in test loss. Rather than report a standard deviation across random initialization, the authors compare against a centralized baseline trained on the pooled dataset for the same number of epochs. Here the federated model was tele-trained to a better final accuracy (91.1% vs. 88.98%) against a centralized model despite training under non-IID data partitioning across ten clients.

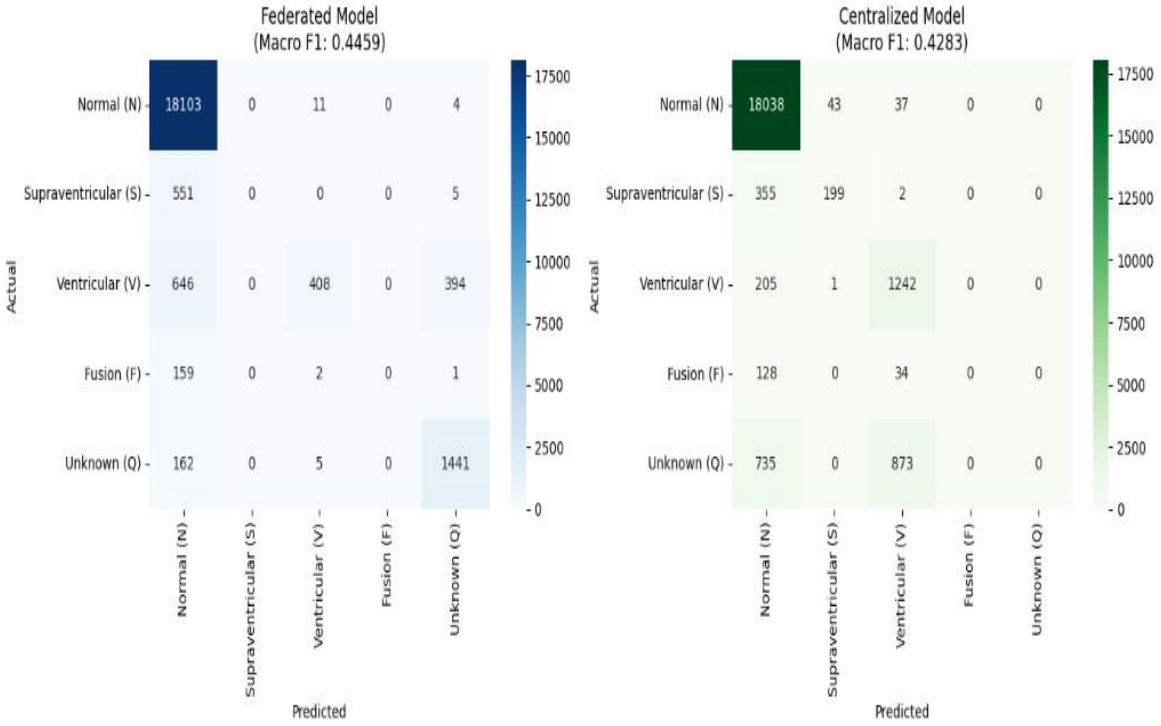

Figure 4. Confusion matrices for the unweighted federated (left) and centralized (right) models, illustrating majority-class bias under severe class imbalance

Table 1. Baseline (Unweighted) Federated vs. Centralized Performance

Model	Test Accuracy	Test Loss	Macro F1	Weighted F1
Federated (FedAvg)	0.9114	0.3321	0.4459	0.8841
Centralized	0.8898	1.4318	0.4283	0.8523

Still, a closer inspection of per-class performance shows that both models achieve such accuracy due to strong performance on majority Normal class and almost a complete failure on minority arrhythmia types. The federated model achieved 0% recall on the Supraventricular and Fusion classes. The centralized model had a slightly better Supraventricular recall of 35.8% but also failed entirely on Fusion (0% recall), along with failure on Unknown (0% recall). This confirms that the raw accuracy is a misleading measure in the context of the severe class imbalance of the dataset (Section 3.1), and that the macro F1-score 0.446 for federated and 0.428 for centralized better represents the models' inability to detect clinically relevant minority arrhythmias at this stage.

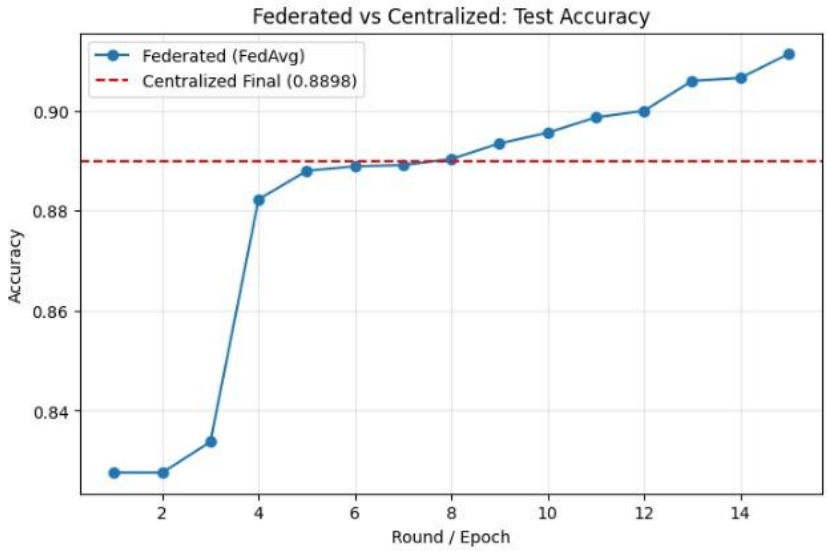

Figure 5. Federated (FedAvg) test accuracy across communication rounds compared to the final centralized baseline accuracy

4.2 Effect of Class-Imbalance Mitigation

Table 2. Effect of Class-Weighting Strategies on Federated and Centralized Models

Model Variant	Test Accuracy	Macro F1	Weighted F1
Federated (unweighted)	0.9114	0.4459	0.8841
Federated (per-client weighted)	0.5589	0.4053	0.6422
Federated (global weighted)	0.9102	0.4814	0.8914
Federated (global weighted + amplified + extended local epochs)	0.8528	0.5376	0.8724
Centralized (unweighted)	0.8898	0.4283	0.8523
Centralized (global weighted)	0.7957	0.5924	0.8406

When we apply balanced inverse-frequency class weights independently per client, our training becomes brittle, showing test accuracy as low as 39% and as high as 63% across rounds and a decreasing macro F1 (0.405 vs. 0.446 baseline unweighted). Clients whose representations of a given class are very sparse are computed to have disproportionately large local class weights, leading to locally over-corrected updates which interfere destructively during weight averaging. To solve the mentioned instability, we recompute class weights once from global training distribution and then apply weights uniformly across all clients. This leads to smooth convergence (Table 2) and macro F1 score of 0.481 (without lowering the accuracy at 91.0%).

To further boost recall on the most significantly underrepresented classes a final amplification factor of 2.5× was used for the Supraventricular and Fusion class weights, as well as an increase in local epochs per round (2→4). The highest macro F1 was obtained using one of the federated variants (0.538). Supraventricular recall was recovered (59.0% rather than 0%) but at a cost of lower overall accuracy of 85.3% with higher round-to-round variance. Fusion recall did not exceed 0% for any of the federated configurations that were tested and will be further examined in Section 5.

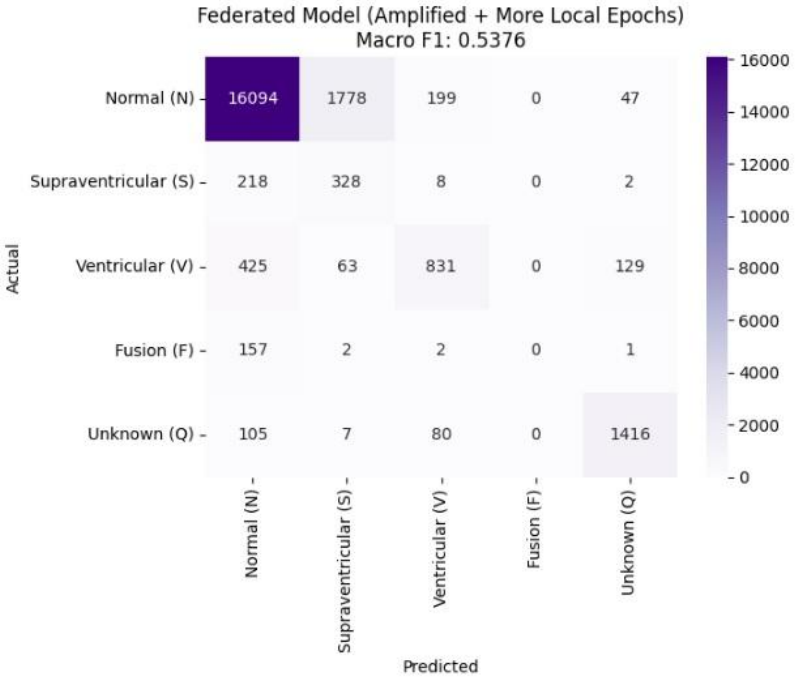

Figure 6. Confusion matrix for the federated model after global class-weight amplification and extended local training (best federated configuration, Macro F1 = 0.538)

The centralized model, trained on the fully pooled (non-partitioned) dataset with the same global class weights, achieved the highest overall macro F1 of the study (0.592), with substantial cross-class recall improvements across all minority classes notably Fusion recall rising from 0% to 93.2% and Ventricular from 57.4% (best federated result) to 90.9%. This trade-off resulted in a drop in overall accuracy (79.6%), which confirmed our expectations of the trade-off in overall accuracy for balanced sensitivity from minority classes.

Table 3. Per-Class Recall Comparison, Best Federated vs. Centralized Weighted Models

Class	Federated (amplified) Recall	Centralized (weighted) Recall
Normal (N)	88.8%	77.5%
Supraventricular (S)	59.0%	67.8%
Ventricular (V)	57.4%	90.9%
Fusion (F)	0.0%	93.2%
Unknown (Q)	88.1%	95.0%

The findings suggest a clear trend; centralized training with access to the pooled dataset is much better at capturing ultra-rare class patterns (particularly Fusion) than federated training with the non-IID partition used in this study, even with aggressive class-weighting.

4.3 TinyML Compression Results

Table 4. TinyML Quantization Results (Full Test Set, N = 21,892)

Model	Quantization	Size (KB)	Test Accuracy	Macro F1	Avg. Latency (ms)
Federated (amplified)	None (float32)	33.79	0.8528	0.5376	0.0174
Federated (amplified)	Dynamic-range	15.52	0.8547	0.5386	0.0206
Federated (amplified)	Full INT8	17.22	0.8664	0.5451	0.0207
Centralized (weighted)	None (float32)	33.76	0.7957	0.5924	0.0193
Centralized (weighted)	Dynamic-range	15.49	0.7959	0.5944	0.0222
Centralized (weighted)	Full INT8	17.22	0.7997	0.5959	0.0215

Both models underwent significant size reduction as a result of quantization. Notably, the dynamic-range quantization process diminished the model size to around 54%, (33.8 KB → 15.5 KB). Additionally, full INT8 quantization brought about a decrease of 49% in model size (33.8 KB → 17.2 KB) compared to the baseline model before quantization. Interestingly, quantization did not impair the performance of either model; specifically, both accuracy and macro F1 improved marginally under INT8 quantization compared to the float32 baseline (federated: +1.36 points accuracy, +0.75 points macro F1; centralized: +0.40 points accuracy, +0.35 points macro F1), which seems like a mild regularization-like effect due to the noise-averaging properties of quantization on this small model.

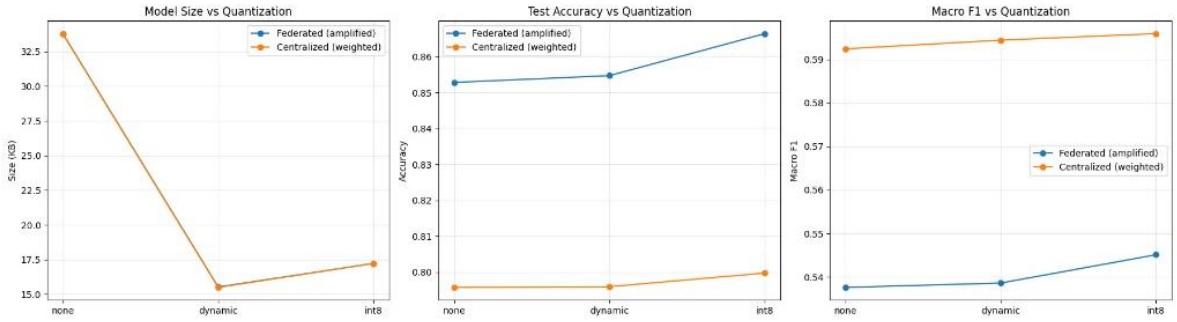

Figure 7. Effect of post-training quantization (dynamic-range and INT8) on model size, test accuracy, and macro F1-score for both federated and centralized models

The per-sample inference latency for all configurations was always below 0.03 ms, which is several orders of magnitude faster than the requirement for real-time continuous wearable ECG monitoring (a new heartbeat approximately every 0.6–1.0 seconds at typical resting heart rates). The sizes of quantized final models, between 15.5 and 17.2 KB, are within memory budget of common wearable-class microcontrollers, indicating the practical deployability of the TinyFed-XAI model.

4.4 Explainability Findings

The SHAP-based analysis was carried out for both models using the appropriate test samples (Section 3.7). As would be expected clinically, both federated and centralized models gave high attribution to physiologically meaningful waveform regions: for Normal beats, both models focused attribution on the QRS complex peak; for Supraventricular beats, both models attended to abnormal P-wave-region morphology and a secondary peak, consistent with known electrophysiological markers of supraventricular arrhythmias; and for Unknown/paced beats, both models converged on the terminal upstroke region of the waveform.

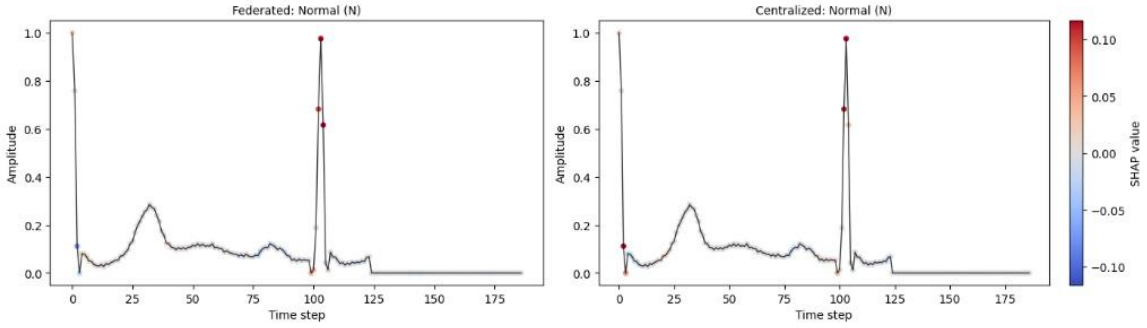

Figure 8. SHAP attribution overlays for a correctly classified Normal (N) beat - federated (left) vs. centralized (right) model

For Ventricular beats, the two models showed clearly differing attribution patterns. In particular, the explanation of the federated model focused on the waveform's onset, while that of the centralized model was more widely spread over the entire QRS-adjacent complex. These dissimilarities align with the understanding that the two models leaked signals from distinct training configurations. Furthermore, these configurations differ in their structure, with one being non-IID. Also, there might be weight-averaged federated updates. The other model might employ fully-pooled centralized gradients. Researchers must interpret these findings carefully. They must remember that the samples of correctly-classified example samples of this class from the two models necessarily differ from one another.

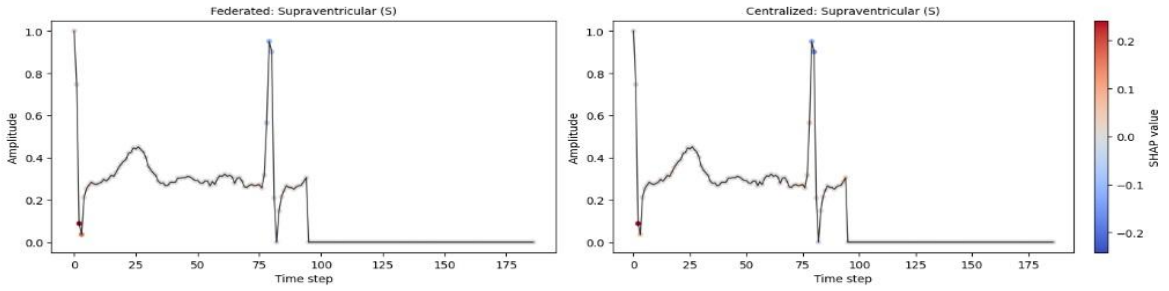

Figure 9. SHAP attribution overlays for a correctly classified Supraventricular (S) beat - federated (left) vs. centralized (right) model

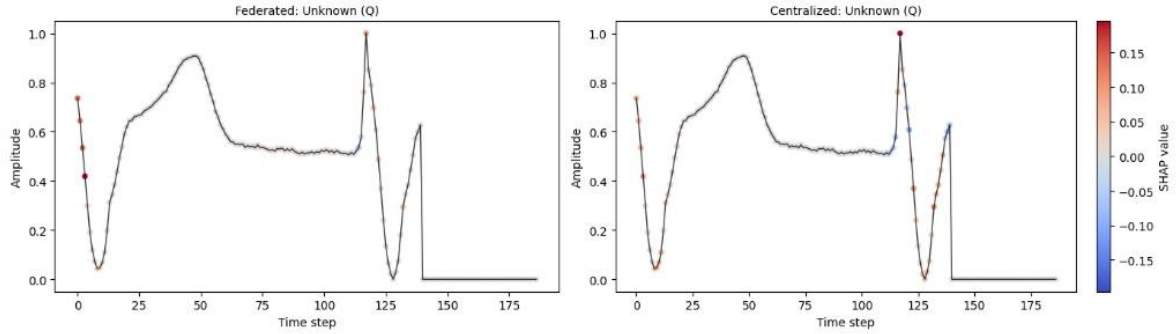

Figure 10. SHAP attribution overlays for a correctly classified Unknown/paced (Q) beat — federated (left) vs. centralized (right) model

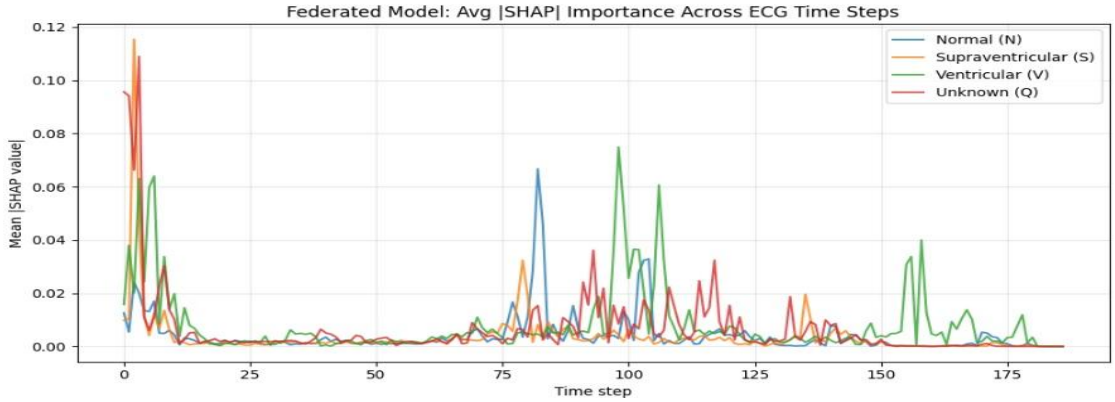

Figure 11. Global SHAP feature importance across ECG time-steps for the federated model, by class

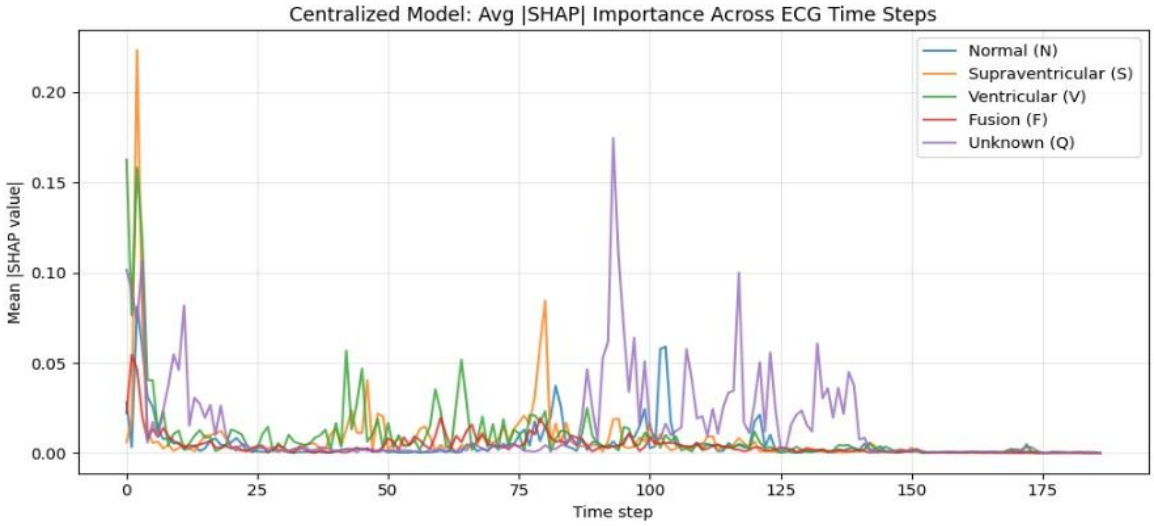

Figure 12. Global SHAP feature importance across ECG time-steps for the centralized model, by class

4. Discussion

The findings from this research offer insight into the possibilities and practical limitations of leveraging federated learning, TinyML, and explainability for the wearable detection of arrhythmia.

Federated Learning is effective yet it performs poorly in real-world settings. Throughout every experiment, the federated model either matched or surpassed the raw accuracy of the centralized model, indicating that FedAvg can successfully train an arrhythmia detector without centralizing sensitive patient ECG data. Despite the distributed nature of the federated approach, we witnessed suboptimal training of the global model. With the rise of AI, there is pressure on federated learning (FL) to democratize the deployment of high-performance systems. Our implementation does not fail here; rather, it is a known structural limitation of FedAvg. When classes are rare and unevenly distributed across clients, what happens is that some clients provide little or no learning signal for it, and averaging dilutes the little signal that is available. Our findings in Section 4.2 reveal the direct evidence. Although aggressive class weighting and more localized training narrowed the gap considerably, it could not be closed completely.

The clearest instance of this limit is the Fusion class. A mere 641 training samples were available from ten clients, with a number of clients possessing zero or one example of a Fusion beat. A model cannot learn anything it has not seen in a training batch, regardless of loss-weighting. The model that had access to the complete pooled dataset could still learn Fusion quite well (93% recall). The federated model, which operates on fragmented views per client, could not (0% recall) with any configuration we tested.

A good and honest finding is that class-weighting fixes are not enough for extremely rare, unevenly distributed classes in federated healthcare and other settings. Future researchers should try synthetic minority oversampling, client-side data augmentation, knowledge-distillation-based aggregation, or whatever else might work to help fix this issue.

The cost of compression was close to nothing. It is a nice and practically useful discovery that quantization of the model to INT8 for microcontroller deployment does not hurt performance it helped very slightly. The tiny size (15–17 KB) and near-instant inference time evidence that the "TinyML" part of our framework can not only be done in theory but can actually be deployed on real wearable hardware today.

Explainability is a matter of degree. SHAP analysis shows that both models (federated and centralized) largely focused on the same clinically meaningful parts of the ECG waveform (e.g. the QRS complex for Normal beats), which is reassuring – indicating that federated training doesn't cause models to reason in a clinically implausible way. The reasons given by the federated model were regularly weaker and less broadly applicable than those of the centralized model. Simply put, a federated model could be less accurate on rare classes but could also be less confidently interpretable about what it does classify correctly. The new angle which is worth investigating is that federated learning which protects data privacy might carry a hidden interpretability cost.

5. Conclusion & Future Work

The present paper, TinyFed-XAI, proposes a framework based on Federated Learning, TinyML compression and Explainable AI to detect arrhythmia in wearable IoT. Through 10 wearable clients, the non-IID simulation was performed and a compact 1D-CNN with 7013 parameters was trained via FedAvg. It increased the macro F1-score from 0.446 to 0.538 with the help of global class-weighting and prolonged local training. The model was compressed to 15-17 KB by INT8 quantization without loss of accuracy, indicating feasibility for microcontroller-class hardware. The SHAP analysis suggests that both the federated model and the centralized model focus on clinically relevant ECG regions which may help with clinical interpretability.

Inspite of this limitation, we found that federated learning is unable to detect extremely rare Fusion beats with any weighting scheme, while centralized training can at least identify this class. This presents a genuine, unresolved trade-off between FL's privacy benefits and its fairness towards minority classes, something with immediate clinical implications since rare arrhythmias are often the most hazardous.

Our next steps are set on two axes: (1) In-client rare class synthetic minority (e.g. SMOTE and GANs based ECG) augmentation to give a learning signal to data-sparse clients, and (2) real hardware validation through deploying the quantized model on microcontroller platforms (e.g. ESP32, Arduino Nano 33 BLE) to validate performance under true field conditions.

The collective objective of these actions is to lessen whatever performance gap remains and bring federated, on-device, explainable cardiac monitoring more to clinical readiness in the real world.

References

1. Al-Bairmani, M. T., Yazdchi, M., & Nasimi, F. (2026). A joint CNN-Bi-LSTM-transformer architecture with SHAP explanations for multi-label arrhythmia detection from 12-lead ECGs. *Scientific Reports*, *16*(1), 11123.

2. Beck, J., & John, A. (2025). Explainable AI (XAI) for Arrhythmia detection from electrocardiograms. *arXiv preprint arXiv:2508.17294*.

3. Bokhari, S. M., Sohaib, S., & Shafi, M. (2025). Fusion of personalized federated learning (PFL) with differential privacy (DP) learning for diagnosis of arrhythmia disease. *Plos one*, *20*(7), e0327108.

4. Cantoro, D., Shumba, A. T., Semeraro, G., Cotardo, M., Rollo, D., Montanaro, T., ... &Patrono, L. (2025). Modeling and analyzing non-IID data in federated learning based ECG arrhythmia detection scenarios. In *Ital-IA-TW 2025 Thematic Workshops at Ital-IA 2025*. CEUR Workshop Proceedings (CEUR-WS. org).
5. Jimenez-Gutierrez, D. M., Falkouskaya, Y., Hernandez-Ramos, J. L., Anagnostopoulos, A., Chatzigiannakis, I., &Vitaletti, A. (2026). On the security and privacy of federated learning: A survey with attacks, defenses, frameworks, applications, and future directions. *Information Fusion*, 104155.
6. Kilic, M. E., Tufekcioglu, O. A., Yilancioglu, Y. R., Turan, O. E., & Ozcan, E. E. (2025). Explainable AI for specific arrhythmia detection: SHAP-based insights from multi-lead ECG data. *Journal of Medical and Biological Engineering*, *45*(2), 314-324.
7. Luong, D., Quang, B. N., Nguyen, L., & Ngoc, M. N. (2024, October). TinyML-based system for cost-effective ECG rhythm classification on embedded device. In *2024 International Conference on Advanced Technologies for Communications (ATC)* (pp. 209-214). IEEE.
8. Polachan, K. (2026). ArrythML: An Autoencoder-Based TinyML Approach for On-Device Arrhythmia Detection on Resource-Constrained Embedded Systems. *arXiv preprint arXiv:2606.02256*.
9. Ryu, H., Lee, M., Kim, S. H., Kim, J. H., & Yang, H. J. (2025, November). Federated learning for cardiovascular disease prediction: A comparative review of biosignal-and EHR-based approaches. In *Healthcare* (Vol. 13, No. 21, p. 2811). MDPI.
10. Showrav, T. T., Lincoln, S. I., & Hasan, M. K. (2025). EXGnet: a single-lead explainable-AI guided multiresolution network with train-only quantitative features for trustworthy ECG arrhythmia classification. *arXiv preprint arXiv:2506.12404*.
11. Shu, Z., Cao, T., & Huang, H. W. (2026). BiFedKD: Bidirectional Federated Knowledge Distillation Framework for Non-IID and Long-Tailed ECG Monitoring. *arXiv preprint arXiv:2605.14886*.
12. Snousi, H. M., & Aleej, F. A. (2025). Machine learning-augmented partial differential equation solvers for high-fidelity engineering design optimization. Journal of Applied Mathematical Models in Engineering, 34-40.
13. Talukder, M. A., Talaat, A. S., Muna, N. J., Alazab, A., Kazi, M., & Das, U. K. (2025). An explainable deep learning framework for trustworthy arrhythmia detection from ECG signals. *Scientific Reports*, *15*(1), 39496.
14. Wu, H. (2026). Towards Family-Grouped Hierarchical Federated Learning on Sub-5KB Models: A Feasibility Study of Privacy-Preserving ECG Monitoring for Ultra-Resource-Constrained Wearables. *arXiv preprint arXiv:2605.18862*.
15. Zambrano-de la Torre, M., Guzman-Alfaro, S., Acuña-Correa, A., Soto-Murillo, M. A., Guzmán-Fernández, M., Robles-Ortiz, R., ... & Cuevas-Rodriguez, E. O. (2026). Real-Time Cardiac Arrhythmia Classification Using TinyML on Ultra-Low-Cost Microcontrollers: A Feasibility Study for Resource-Constrained Environments. *Bioengineering*, *13*(5), 532.
16. Zhang, M., Wang, Y., & Luo, T. (2020, December). Federated learning for arrhythmia detection of non-IID ECG. In *2020 IEEE 6th international conference on computer and communications (ICCC)* (pp. 1176-1180). IEEE.