SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 6, 2026

FEDERATED LEARNING-BASED ATRIAL FIBRILLATION DETECTION USING DEEP LEARNING AND GENERATIVE AI-BASED RISK EXPLANATION

Alina Sher[1], Zobia Zafar[2], Muhammad Toseef Javaid*[3], Muhammad Umar Javed[4], Umair Bin Yaseen[5]

[1,2, *3,4,5]*Department of Computer Science, University of South Asia, Lahore 54000, Pakistan*

DOI:https://doi.org/10.5281/zenodo.21161506

Keywords
Atrial Fibrillation, Federated Learning, CNN-BiLSTM, ECG Classification, FedAvg, SMOTE, Generative AI, Phi-3-mini-4k-instruct, Explainability, Privacy-Preserving Healthcare AI.

Article History
Received: 25 April 2026
Accepted: 04 June 2026
Published: 21 June 2026

Corresponding Author: *
Muhammad Toseef Javaid

Abstract
Cardiovascular diseases, which in short terms are called "CVD," are linked to a group of diseases related to the heart, brain, and many disorders of blood vessels. Atrial Fibrillation is also a common type of CVD which can cause heart disorders and brain strokes. We research the early detection of this dangerous heart disease so that we can save the lives of patients by spreading knowledge about this irregular heartbeat. Our research uses a technique which can protect the privacy of patients' data because there are also risks of data leakage, and we built a combined CNN and BiLSTM model to reduce this leakage of data. We use the Federated Learning technique in which we can solve the issue of privacy and data security of patients. In this technique, we also use 12-Lead ECG data for binary AF classification, and we use the Federated Averaging (FedAvg) method to train our decentralized model across three local nodes. But we have imbalanced data inside each node, and we solve this issue by using the Synthetic Minority Oversampling Technique (SMOTE). We set up an edge-based pipeline that runs a local Phi-3-mini-4k-instruct language model directly on the device to explain the potential risks of stroke, irregular heartbeat complications, and heart failure after the detection of Atrial Fibrillation. This Generative AI system strictly follows all privacy rules of General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA) to create medical reports and understandable summaries for doctors in hospitals, which is a good point of this system. When we trained our model for AF prediction, our federated model achieved 94% accuracy, which is greater as compared to the centralized model in which we achieved 91% accuracy. This comparison of accuracy shows that we can manage data privacy with AF prediction. In the end, our project proves that we can safely screen for AF in the real world.

1. INTRODUCTION

Heart diseases are spreading all over the world and Atrial Fibrillation is also a common and dangerous type of this disease. In this condition, the electrical signals do not work properly in the upper heart chambers and can lead to an irregular heartbeat. In this irregular heartbeat, blood does not flow properly and it is dangerous because it makes a person to suffer an ischemic stroke five times more. Approximately 37 million people are suffering from Atrial Fibrillation because as people grow older, diseases such as high blood pressure and diabetes also increase. So doctors need to find different methods to detect this

condition as soon as possible so that they can save the lives of people who are suffering from this disease.

The 12-lead ECG test is known as an accurate and best method to detect Atrial Fibrillation, but deep learning frameworks can analyze these ECGs automatically. On the other hand, models like CNN and RNN can detect heart problems with high accuracy similar to doctors. We know that patients' data in hospitals is very sensitive and privacy laws like HIPAA in the US and GDPR in Europe do not allow different hospitals to share this data for the detection of any type of disease. So hospitals cannot share data on a central database for AI training and this becomes a major problem for the development of any AI system for Atrial Fibrillation detection.

As discussed earlier, patient data in hospitals is sensitive so we use Federated Learning in our project in which each hospital trains a local model on its own data and shares the learning information (weights) instead of the data. In this way, we can maintain the privacy of patient data. We build a system in which we use a combined CNN and BiLSTM model to analyze heart signals. We use an imbalanced dataset so we apply SMOTE preprocessing for class equalization to solve the data imbalance issue and we also implement a Generative AI system for risk explanation after Atrial Fibrillation detection. In this research, we analyze the results of both the centralized model and the federated model.

1.1 Problem Statement

There are three main problems that we focus on in this research. First, in deep learning systems, hospitals need to share patients' data in a central place so they can train an AI model which can detect normal and Atrial Fibrillation conditions using ECG reports. But they cannot share this data due to privacy laws. So centralized AI training becomes difficult in the healthcare field. The second problem is the imbalanced data in ECGs which contains a large number of normal conditions compared to cases of Atrial Fibrillation. Therefore, the AI model does not learn properly and cannot detect cases of Atrial Fibrillation accurately. The third problem is that the majority of deep learning models work like a "black box" which can predict accurately but these models cannot explain the reasons for their results. So doctors do not rely completely on these AI systems and the use of these systems becomes difficult.

In this research, we try to solve all of these problems using a single framework. We use the Federated Learning technique so that the privacy of patients' data remains secure and to solve the data imbalance issue, we use Synthetic Minority Oversampling Technique (SMOTE). We also build a Generative AI system to explain the risks of the predicted condition which helps doctors understand the patient's result reports.

1.2 Research Objectives

The main objectives of this research are structured as follows:

- The hybrid deep learning model is designed by combining Convolutional and Bidirectional Long Short-Term Memory (CNN-BiLSTM) layers for the detection of normal heartbeat and Atrial Fibrillation conditions using a 12-lead ECG dataset.
- The Federated Learning system is developed in this research in which three local nodes/hospitals train their models separately on their own patient data and share learning weights by using the FedAvg method. In short, different hospitals train the AI model without sharing their patient data.
- The Synthetic Minority Oversampling Technique (SMOTE) is implemented at each local node to solve the data imbalance issue so that the model can learn properly from both normal and Atrial Fibrillation cases.
- The Generative AI explanation system is driven by a locally hosted Phi-3-mini-4k-instruct large language model. It not only predicts Atrial Fibrillation but also explains the results in a way that is helpful for doctors to understand patient risks.
- The results of the centralized and federated model are compared using different evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

2. Literature Review

2.1 Deep Learning for ECG-Based Arrhythmia Classification

In the last decade, deep learning has improved ECG analysis significantly. Rajpurkar et al. developed a 34-layer CNN model trained on more than 64,000 ECG recordings, which achieved performance similar to cardiologists in detecting different heart rhythm disorders. This study also introduced an important deep learning architecture that is widely used in ECG analysis today. However, the study showed that a very large amount of labeled ECG data is required to achieve high accuracy, which makes the development and deployment of such systems difficult in different hospitals.

Hannun et al. developed a 12-layer deep neural network using 91,000 ECG segments for heart rhythm classification. Their model achieved better F1-scores than six cardiologists in detecting Atrial Fibrillation. The study showed that AI can help doctors in clinical decision-making. However, like previous studies, this model was trained using centralized data collected from a single source. Therefore, it did not address the data privacy and sharing problems that exist between different hospitals under privacy laws such as HIPAA and GDPR.

2.2 Hybrid CNN-LSTM Architectures for Temporal ECG Modelling

Pure CNN models can understand only small and local patterns of ECG signals, such as the sharp movements of QRS complexes or the shape of P-waves. However, the limitation of CNN models is that they cannot model long-term sequences or continuous heart rhythms correctly. Atrial Fibrillation is not represented by only a single abnormal heartbeat; rather, it is characterized by irregularity throughout the heart cycle, especially in the R-R intervals that continue across multiple heartbeats. To understand these long-term patterns, recurrent networks are required; therefore, LSTM is used along with CNN.

Firstly, Yildirim worked on this architecture in which wavelet features and LSTM were combined. In this way, good accuracy was achieved, but the preprocessing was computationally expensive. On the other hand, Tan et al. used Bidirectional LSTM, which analyzes signals in both directions (forward and backward). By using this approach, AF detection becomes more accurate because the model understands the whole context of the signal. Petmezas et al. showed that model performance is affected when the dataset is imbalanced. They used class weighting, but their study also demonstrated that data imbalance is a serious issue.

Our study is based on these ideas, but we use the SMOTE technique to balance the dataset instead of using loss weighting. To evaluate performance, we do not rely only on accuracy; we also use strong metrics such as Recall, F1-score, and ROC-AUC.

2.3 Federated Learning in Healthcare: Principles and Empirical Evidence

The Federated Learning framework was first introduced by McMahan through the FedAvg algorithm. This method allows multiple clients to train their models locally and then send only model updates to a central server where a weighted average is calculated based on the size of each client's data. The main idea is that when data cannot be shared due to privacy rules, institutional policies, or technical limitations, federated learning enables collaborative training without moving raw patient data outside local hospitals. This approach was first tested on image and text data, but it was also suggested that it can be useful for medical applications.

Later, Sheller applied federated learning in a medical setting for brain tumor segmentation using MRI data from different hospitals. Their results showed that a federated model can achieve performance close to a centralized model while still keeping all patient data local. This study proved that federated learning can use data from different hospitals without sharing it, and performance depends on factors like number of training rounds and number of clients.

In ECG-based research, Bisna used a hybrid CNN-BiLSTM model with FedAvg and FedSGD on the PTB-XL dataset. Their study showed that federated learning can perform well in ECG classification without sharing sensitive patient data. They also used techniques like quantization

to make the model suitable for edge devices with limited resources. Similarly, Liu studied large-scale federated learning across multiple hospitals and found that non-IID data is a major challenge because each hospital has different data distribution. This can affect training performance, so methods like FedProx and SCAFFOLD are used to improve stability.

A recent study by Chorney and Ling compared different communication-efficient methods for Atrial Fibrillation detection. Their results showed that increasing communication between clients can reduce the performance gap between centralized and federated models, but it also increases communication cost. Overall, their study shows that there is always a trade-off between communication, convergence speed, and accuracy. In our work, we use the FedAvg method with controlled learning rate to keep training stable and achieve good performance in a federated environment.

2.4 Handling Class Imbalance in Clinical ECG Classification

Data is often unbalanced in Atrial Fibrillation datasets, and the number of normal heartbeat cases is much larger than the number of AF cases. Therefore, when standard neural networks are trained on this data, they learn to predict normal cases accurately and tend to ignore AF cases. As a result, the overall accuracy may be high, but the model cannot properly detect AF cases. To solve this problem, the SMOTE technique is used, which was introduced by Chawla et al. The purpose of SMOTE is to generate new synthetic AF samples by interpolating between existing AF samples. This method creates new variations instead of simply duplicating existing samples. As a result, the model is exposed to more diverse AF patterns and the risk of overfitting is reduced.

The use of SMOTE in Federated Learning is slightly different because data cannot be shared between hospitals. Therefore, each hospital applies SMOTE to its own local data. However, if a hospital has a small number of AF cases, the diversity of the synthetic data is also limited. If SMOTE is not used, the model may collapse and predict only normal heartbeats while failing to detect AF cases. By using SMOTE, the model gets the opportunity to learn both normal and AF classes. The global model also performs better after federated learning. Therefore, SMOTE is an important step in federated learning, especially when the dataset is highly imbalanced.

2.5 Explainable and Generative AI for Clinical Decision Support

The basic idea of federated learning was introduced by McMahan et al., in which the FedAvg algorithm is used. Multiple hospitals train their model locally and send updates to a central server. The server combines these updates through a weighted average, where the contribution of each hospital depends on its data size. Its main purpose is that when data sharing is not possible due to privacy restrictions, model training can still be performed without sharing raw data. After that, Sheller et al. proved the working of federated learning in the medical field and showed on an MRI brain tumor dataset that a decentralized model performs as well as a centralized model and performs better than a single hospital model. Then it proves that federated learning improves learning by using data from different hospitals without transferring data.

Bisna et al. used FedAvg and Fed SGD along with a CNN-BiLSTM model and achieved good results on the PTB-XL dataset. They also stated that quantization is useful to run models on edge devices. Liu et al. highlighted in a large-scale study that real-world hospital data is not of the same type (non-IID problem), which means that each hospital's data has a different distribution, making training unstable. Methods like FedProx and SCAFFOLD are proposed to solve this problem. Chorney and Ling compared different federated methods in AF detection. The result was that if the global model is updated more frequently, performance improves but network bandwidth usage increases. This shows that there is always a trade-off in federated learning and a balance must be maintained between communication cost, training stability, and model accuracy.

Our study uses FedAvg with a careful learning strategy where the aim is that the model remains

stable and AF detection is accurate even if training is slow.

2.6 Research Gap and Positioning of the Present Work

There are three main problems in this research, and this study aims to solve all of them. First, the majority of AF detection models, such as those developed by Rajpurkar and Hannun, are trained on a central server using centralized datasets and are not tested in a federated learning environment. Therefore, their performance in real-world distributed healthcare environments is not clearly evaluated. Second, existing federated ECG classification studies, such as those by Bisna and Liu, mainly focus only on prediction accuracy. They do not provide proper explanations of model decisions for doctors, which limits their use in real clinical practice. Third, current explainable AI systems using large language models for cardiac applications either depend on cloud-based services, which can create privacy risks, or they are only tested as basic prototypes without real deployment in working systems.

In contrast, our proposed system addresses all these gaps together. It combines privacy-preserving federated learning for AF detection, uses SMOTE at each local client to handle data imbalance, and includes a fully working local generative AI explanation system that runs on-device and follows privacy regulations like HIPAA and GDPR. Table I shows the comparison of our work with existing studies.

Table 1: Literature Review Comparison

Reference	Method	Dataset	Acc.	AUC	Federated?	Explainability?
Rajpurkar *et al.* [3]	ResNet CNN	64K ECG	~84%	~0.97	No	No
Hannun *et al.* [4]	12-layer DNN	91K ECG	~85%	0.97	No	No
Tan *et al.* [7]	BiLSTM	PhysioNet	91.5%	0.88	No	No
Bisna *et al.* [9]	FL+CNNBiLSTM	PTB-XL	99.1%	N/R	Yes	No
Liu *et al.* [10]	FL + CNN	Multi-hospital	89.3%	0.83	Yes	No
Proposed (Central)	CNN-BiLSTM	12-lead ECG	91%	0.76	No	Implemented
Proposed (Federated)	FL+CNNBiLSTM	12-lead ECG	94%	0.63	Yes	Implemented

3. Proposed Methodology

3.1 System Architecture

The proposed system follows a four-stage pipeline: ECG data preprocessing, centralised CNN-BiLSTM model training as a performance baseline, federated learning across three distributed client nodes, and an operationalized Generative AI explanation pipeline that produces clinical narratives for individual predictions via a locally hosted Phi-3-mini-4k-instruct LLM. Figure 1 provides an overview of the system architecture.

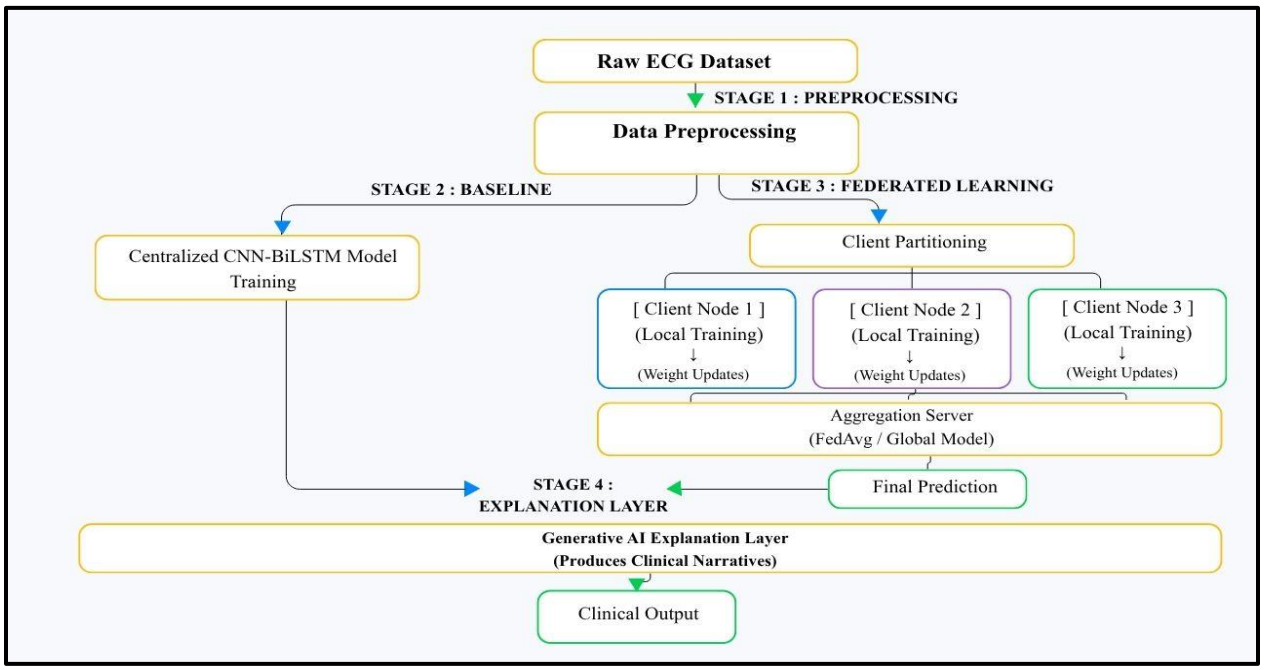

Figure 1: Workflow of four-stage system architecture.

3.2 Data Preprocessing

First, ECG data can have different value ranges because it may be collected from different machines or electrode setups. To solve this problem, we use Min-Max Normalization (Equation 1). This process converts all ECG values into a range between 0 and 1. As a result, all ECG signals become consistent and the model can learn the data more effectively. After normalization, the ECG data is reshaped into the required format so that it can be processed by the Conv1D layers.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min} + \epsilon}$$

After that, we address the class imbalance problem because the number of Atrial Fibrillation (AF) samples is much smaller than the number of normal heartbeat samples. To solve this issue, we use the Synthetic Minority Oversampling Technique (Equation 2). This technique creates new artificial AF samples and increase the quantity of these AF samples so that the model can train more effectively. As a result, the number of AF cases increases and the model can learn both classes more effectively. Each hospital applies SMOTE on its own data separately and does not share any data with other hospitals to maintain data privacy.

$$x_{new} = x_i + \lambda(x_{zi} - x_i), \quad \lambda \sim \text{Uniform}(0,1) \tag{2}$$

At last, we do not use the default threshold of 0.5 for model prediction and instead choose a suitable threshold with the help of validation data. This helps the model to detect AF cases more accurately and reduces the chances of missing AF samples.

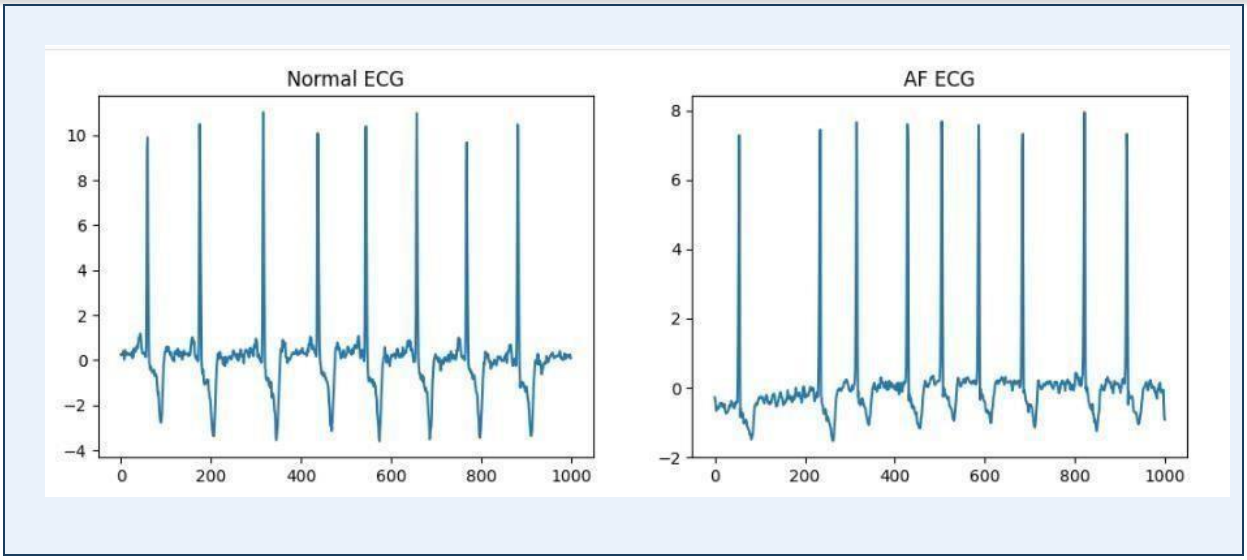

Figure 2: ECG signals of Normal heartbeat and Atrial Fibrillation.

3.3 CNN-BiLSTM Deep Learning Model

ECG signals are given to the model in a multi-channel 1D form which means that heart activity is analyzed as a sequence signal. First, Conv1D layers extract local features from the ECG such as P-waves, QRS complexes, and T-waves, which represent the basic patterns of heartbeat. After the Conv1D processing, a Max Pooling layer reduces the size of the data so that unnecessary information is removed while important features are preserved.

Then, a Bidirectional LSTM layer analyzes the ECG sequence in both forward and backward directions, so that irregular heart rhythm patterns, especially Atrial Fibrillation, can be better understood. At the final stage, the extracted features are passed through a Dense layer where the ReLU activation function is used, along with Dropout to reduce overfitting. Finally, the Sigmoid layer produces the output, which indicates whether the patient is suffering from Atrial Fibrillation or not. Figure 3 shows the complete stack of layers in the model.

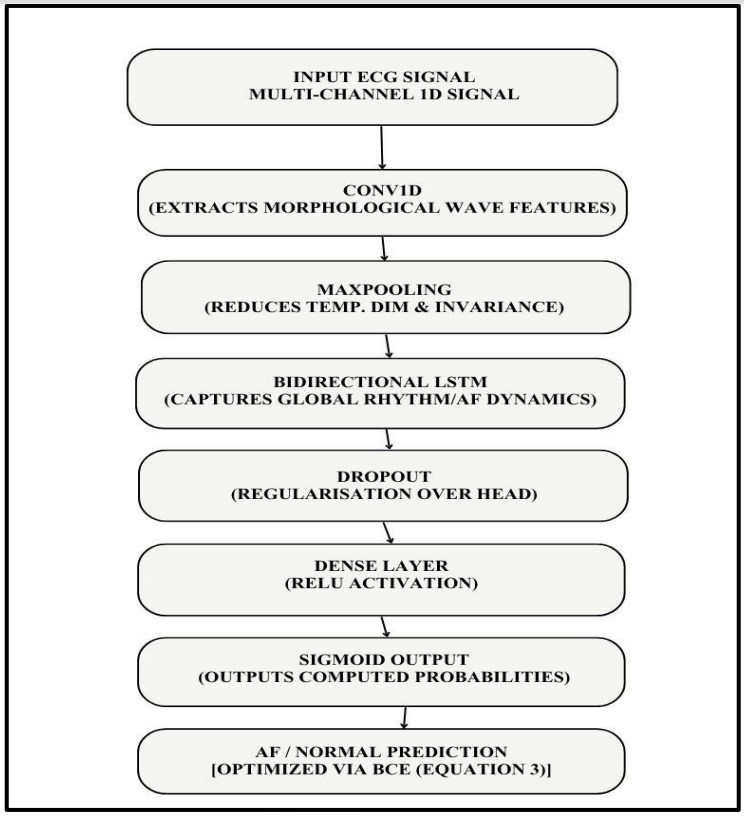

Figure 3: Structure of the CNN-BiLSTM Model

Binary Cross-Entropy (BCE) loss function and the Adam optimizer are used for model training. The main purpose of BCE is that when the model makes a wrong prediction, it applies a penalty. If the model makes a wrong prediction with high confidence, it receives a higher penalty. If the model is unsure while making a wrong prediction, the penalty is comparatively lower. Therefore, BCE is best suited for binary classification problems, especially when the dataset is imbalanced. The Adam optimizer makes the training process fast and efficient by continuously adjusting the learning process.

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)\right] \quad (3)$$

3.4 Federated Learning Framework

Our model uses a federated learning approach in which the FedAvg protocol is followed. First, the central server builds a CNN-BiLSTM model and sends its weights (learning values of the model) to all hospitals.

Each hospital trains the model for a few epochs on its local ECG data, which is balanced using SMOTE. After training, each hospital sends the updated weights back to the central server. Then, the central server combines all these weights using a weighted average, where the contribution of each hospital depends on its data size. A global model is created through this process, which is again sent to all clients for the next round of training. This cycle is repeated until the model is improved. This working process is shown in Figure 4.

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

(4)

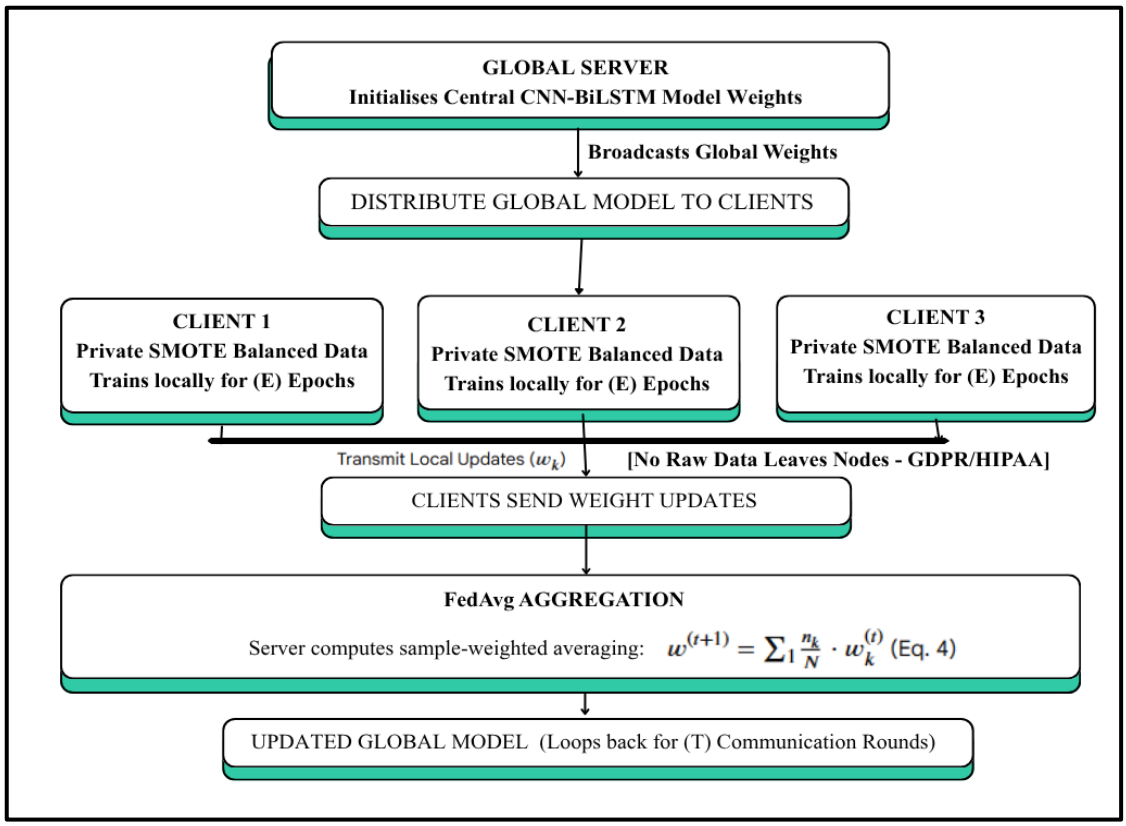

Figure 4: Training process for Federated Learning

This formula also means that the weight of each hospital/client is considered according to its data size. The hospital with large data has a higher impact than the hospital with low data. The main benefit of this approach is that raw ECG data is not shared. The patient data remains fully secure, and privacy laws (GDPR and HIPAA) are followed.

3.5 Generative AI-Based Risk Explanation — Fully Implemented Pipeline

The major issue of using deep learning models in hospitals is that AI just gives numerical results, such as the probability of AF being 85% or 90%. But doctors and nurses do not consider only this percentage; they also need an explanation of on which basis the AI gives these results. Doctors need a clear and structured explanation before writing a diagnosis for patients and referring them to a specialist. To solve this problem, we add a Generative AI explanation layer with a federated detection system of AF. This AI system not only predicts but also generates understandable reports about the prediction, which becomes helpful for doctors to understand and use these results.

This explanation system is locally hosted and runs only on a device in which the model operates in FP16 format. In this way, memory and resource

usage remains low, but the model can still generate summaries accurately. The main point is that this system completely runs on a local system. There is no need to send any ECG features, model results, or external API calls to any cloud service or third-party server. The benefit of this system architecture is that all the data is saved on one machine, and data privacy is secured, and healthcare data protection laws such as HIPAA and GDPR are followed.

To generate reliable medical reports, the model was configured so that it does not produce random outputs. Random sampling was disabled by setting *do_sample* = False and the temperature was fixed at 0.1. As a result, if the same input is provided multiple times, the system generates the same medical report every time. This is very important in healthcare because doctors require consistent and repeatable reports that can be reviewed and audited.

The explanation module uses extracted ECG features, the prediction confidence score, and the final classification result to generate medical summaries. In addition, the model is guided through a cardiologist-based prompt so that it uses professional medical terminology and focuses on important signs of Atrial Fibrillation such as missing P-waves, irregular RR intervals, and Fibrillatory (relating to fibrillation) baseline patterns. The explanation process is represented by Equation (5).

$$\mathcal{E} = LLM(F_{ECG}, \hat{p}, B_{clinical}) \quad (5)$$

This explanation module is not limited to the Phi-3 model and can also work with other instruction-tuned language models. We selected Phi-3-mini-4k-instruct because it is lightweight, follows instructions effectively, and runs efficiently on local hardware. The complete pipeline was verified using system logs, which showed that the model consistently generated accurate medical summaries based on standard Atrial Fibrillation symptoms. Examples of these generated reports are presented in Table 3.

Table 2: Training Hyper parameters for Centralized and Federated Learning Models

Hyper parameter	Centralized	Federated
Architecture	CNN-BiLSTM	CNN-BiLSTM
Optimizer	Adam	Adam
Loss Function	Binary Cross-Entropy	Binary Cross-Entropy
Learning Rate	0.001	0.00005 (local)
Batch Size	32	16
Epochs / Local Epochs	50	20 per round
Communication Rounds	N/A	8
Number of Clients	N/A	3
SMOTE	Applied globally	Applied per client
Threshold Optimization	Yes	Yes

Table 3: Model Predictions with Confidence Scores and Clinical Explanations

Predicted Class	Confidence Score ($\hat{p}$)	System-Generated Clinical Narrative
AF	0.84	The ECG shows signs of Atrial Fibrillation (AF). This is identified by an irregular heartbeat pattern, especially irregular R-R intervals, and the absence of clear P-waves. These features indicate abnormal electrical activity in the heart and suggest AF condition.
Normal	0.07	The ECG shows a normal heartbeat pattern. The presence of clear P-waves and regular spacing between heartbeats indicates a normal heart rhythm. There are no signs of irregular electrical activity or abnormal rhythm.

4. Experimental Setup

We performed all the experiments in Python 3.10 and used TensorFlow 2.12 and Keras for the development of the deep learning model. Custom Python scripts were designed to implement Federated Learning which simulates the FedAvg method. This whole system was deployed on a single computer, but the files of each client/hospital and the central server were managed separately so that the setup of a federated learning environment could be maintained. To solve the data imbalance issue, the SMOTE technique was used through the imbalanced-learn library. The dataset was divided in such a way that 80% was used for training, 10% for validation, and 10% for testing. Stratified sampling was used so that each part of the dataset had balanced classes. During training, testing data was not used for training and model selection and was used only for checking the performance of the model.

4.1 Hardware and Generative AI Inference Infrastructure.

An NVIDIA T4 GPU with 15 GB VRAM was used to run the model. At the same time, an Intel Core i7 (12th generation) processor and 32 GB RAM were also used to ensure smooth training. The training of the Federated CNN-BiLSTM model and the whole system was run on this same setup. For the Generative AI explanation, Phi-3-mini-4k-instruct was loaded in FP16 format which reduces memory usage. During testing, GPU memory usage was only 7.64 GB which is much lower than the 15 GB limit. This shows that the complete system, including the federated learning model and AI explanation system can run easily on a normal/mid-range GPU without memory crashes or performance loss.

4.2 Decoding Configuration.

The Generative AI system was integrated in such a way that it generates the same outputs and does not produce random answers. For this purpose, deterministic decoding was used to ensure correct and consistent output generation. To prevent the model from selecting random words, we set do_sample = False and used a temperature of 0.1 to produce fixed and stable outputs. As a result, if we give the same input repeatedly, the AI system generates the same medical report every time. This is important in the medical field where doctors need repeatable and consistent reports so that they can verify or audit the system.

5. Results and Evaluation

5.1 Performance Overview

When both models were evaluated, we observed a confusing difference: the federated model achieved better performance with 94% accuracy which is higher than the centralized model which achieved 91% accuracy. But when the ROC-AUC score was checked, the centralized model performed better with a score of 0.76 compared to 0.63 for the federated model. This difference was due to the difference in decision threshold for both models. We manually adjusted the decision threshold of the federated model and set a lower learning rate for this model. This technique reduces false alarms but it also reduces the overall classification performance of the model. On the other hand, the centralized model distinguished both classes more effectively. Finally, we compared

the results of both models and presented them in Table 4.

Table 4: Final Comparison of Model Performance on Test Data

Metric	Centralized	Federated	Clinical Interpretation
Accuracy	91%	94%	Overall correct classification rate
Precision	0.27	0.44	AF predictions confirmed as true AF
Recall (Sensitivity)	0.50	0.40	True AF cases detected
F1-Score	0.35	0.42	Harmonic mean of precision and recall
ROC-AUC	0.76	0.63	Global discrimination across all thresholds
Specificity	~93.2%	~97.4%	Normal cases correctly classified

5.2 Confusion Matrix Analysis

We used 200 samples for testing, which contained 190 normal cases and 10 cases of Atrial Fibrillation. The details of this data are given in Table IV. The centralized model detected 5 cases out of these 10 samples of Atrial Fibrillation, so the recall of this model was 50%. However, this model also predicted 13 normal heartbeat cases as Atrial Fibrillation patients. On the other hand, the federated model detected only 4 AF cases. Therefore, this model had a recall of 40%, but it showed only 5 false alarms, which means that it handled normal cases more accurately. As we know, the dataset contained only 10 AF cases, therefore if the model missed even 1 patient, the score decreased by 10%. This shows that large and balanced testing data is very important for future research so that the results can become more reliable.

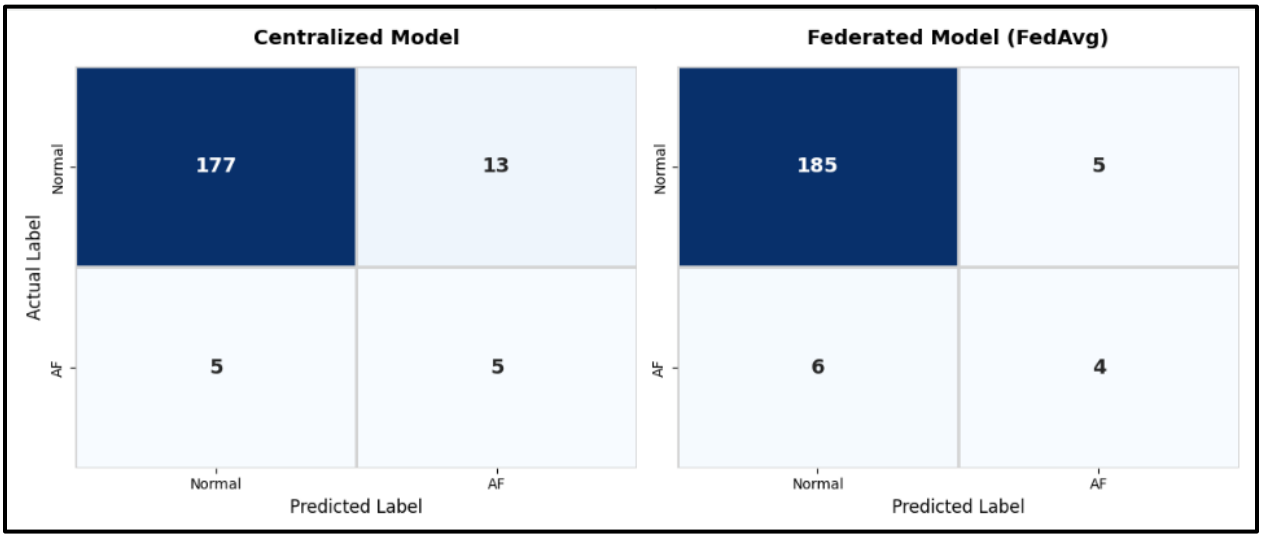

Figure 5: Confusion Matrix Analysis

5.3 ROC Curve Analysis

We used the ROC curve to check how accurately our system predicts normal heartbeat and Atrial Fibrillation signals. Basically, this metric is used to compare correct predictions with false alarms. If a model works perfectly, it achieves a score of 1.0; otherwise, random guessing gives a score of 0.5. Our centralized model achieved a 0.76 ROC-AUC score, which shows that it effectively distinguishes between normal and AF signals. On the other hand, the federated model's score of 0.63 shows that it is better than random guessing but not as strong as the centralized model. Therefore, the model becomes confused when classifying heartbeat signals that are unclear or fall between normal and AF categories. The main reason for this lower performance is that each hospital has a small number of AF cases. When AF samples are limited, the model cannot learn rare and unusual patterns properly.

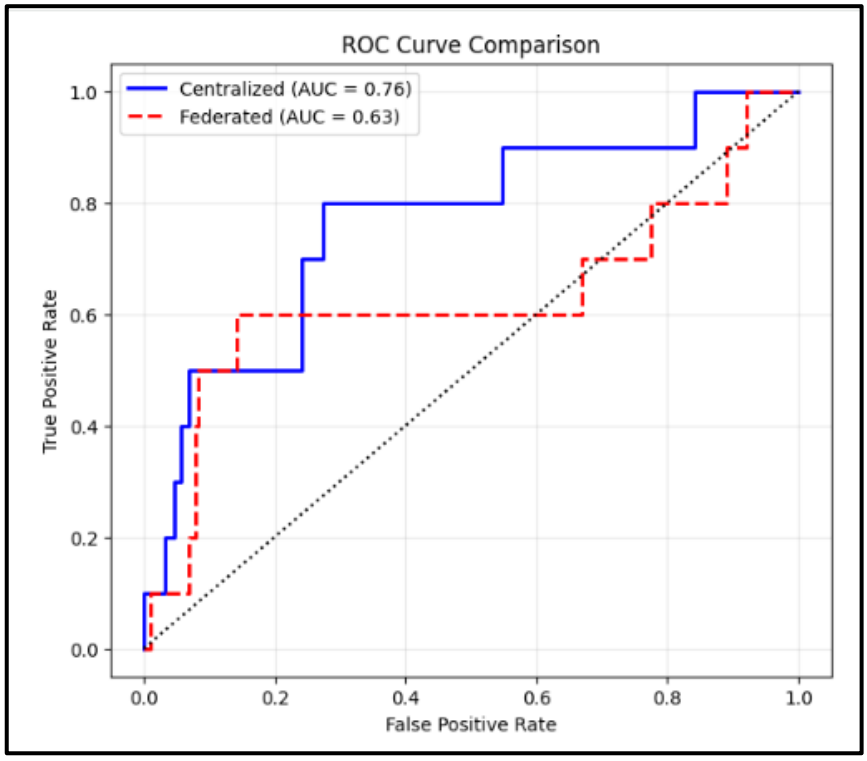

Figure 6: ROC curves by comparing the performance of both models.

6. Limitations

In this research, we have a few limitations, which are as follows:

6.1 Limited AF Data in Testing

We tried our best to arrange a 12-lead ECG dataset from a cardiologist but we failed. Then we found a dataset from the internet which contains only 10 Atrial Fibrillation cases so we used these few cases for model testing. Therefore, these results cannot be 100% perfect for each type of patient. But our Generative AI system is not affected by this dataset size and gives accurate medical summaries in each case.

6.2 Testing on Single Machine

We implemented the entire system on a single machine so we did not face real-world problems such as slow network issues, differences in machine behavior and delays caused by a real distributed environment. The main limitation is that we did not test these real-world problems in this research.

6.3 No Guarantee of Strong Data Privacy

The system implemented in this research can secure data privacy through local training but mathematical differential privacy is not added. Therefore, an attacker may guess information from the model weights.

6.4 Detection of Only Two Cases

The trained model detects only normal heartbeat and Atrial Fibrillation but there are many other heart diseases in hospitals which can cause the death of a patient. However, this system is trained only for the detection of Atrial Fibrillation.

6.5 No Real Hospital Testing for Explanation System

The Generative AI explanation system generates clinical reports for risk explanation but it has not been practically tested with doctors in real-world hospitals. Therefore, there is no certainty that it will be useful in clinical practice.

7. Conclusion

This research introduces a new federated learning system that is specifically developed for the detection of Atrial Fibrillation. The main benefit of this research is that it addresses major challenges in the use of AI in hospitals such as data privacy, unbalanced data, and lack of trust from doctors. First, we built a basic centralized CNN-BiLSTM model which was used as a baseline. This model achieved 91% accuracy and a 0.76 ROC-AUC score to provide a reference for comparison. Then, we tested the federated model, which achieved 94% accuracy, which is higher than the centralized model. This shows that different hospitals can develop an AI system without sharing patient data and patient privacy as well as ECG data remains secure in this process.

We did not hide that the performance of the federated model, which protects privacy, is slightly lower than the centralized model, and we honestly mentioned this because it shows that there is some cost to maintaining the protection of private data. However, it also means that the system can be improved in the future by using better algorithms and increasing the model training time. We also added an AI explanation tool which runs locally using Phi-3-mini-4k-instruct and works smoothly on an NVIDIA T4 GPU. This tool generates simple medical summaries which doctors can easily understand and verify. In the end, this research provides a strong starting point for building medical AI systems that follow legal privacy rules for the safety of patients' data.

All of these results prove that federated learning for medical AI is a useful and ethical method, especially when patients' privacy is important. At present, privacy laws are becoming very strict. Therefore, it is becoming difficult and expensive to collect or share ECG data from hospitals. Because of this, federated learning systems like ours will become very important in the future. In our system, federated training and local Generative AI explanations were used, so this research provides a strong base for the development of safe and responsible medical AI systems in real-world hospitals.

8. Future Work

We can make our system stronger in the future because at present, the system secures data through local training but hackers can still try to guess information from the model weights. To solve this problem, differential privacy can be integrated into the research, in which random noise is added so that patient information remains completely safe and private.

The dataset is not balanced in real-world hospitals and some disease cases are very small while normal cases are large in number. Therefore, advanced federated learning methods such as FedProx, SCAFFOLD, and FedNova can be used in the future. These advanced methods can train the model more accurately even with mixed or unbalanced data.

We use a dataset in which the samples of Atrial Fibrillation are small so the detection of rare conditions is difficult. We can give the model more training time and the learning speed can be adjusted automatically in the future. Then the AI system can detect rare medical cases more accurately.

Our model detects only two conditions right now as we discussed before but it can be expanded in

the future for the detection of multiple other dangerous heart diseases.

We implemented this system only in a research environment, not in real-world hospitals. However, this system should be practically tested in future projects to find out whether it works accurately and fast.

It is very important to get feedback from doctors and nurses so that we can test whether doctors trust this system and whether the AI system can help in treatment decisions. This feedback can help to approve this system for clinical use.

References

Alreshidi, M. & Alsaffar, M., "FedCL: An Atrial Fibrillation Prediction System Using ECG Signals Employing Federated Learning Mechanism," ResearchGate preprint, 2024. https://www.researchgate.net/publication/383879227

D. R. Santos et al., "Feasibility Analysis of Federated Neural Networks for Explainable Detection of Atrial Fibrillation," arXiv preprint, 2024. https://arxiv.org/abs/2410.19781

W. Chorney and S. H. Ling, "Federated Learning Strategies for Atrial Fibrillation Detection," Journal of Experimental and Theoretical Analyses, 2025. https://www.mdpi.com/2813-4648/3/3/23

S. Khurshid et al., "ECG-Based Deep Learning and Clinical Risk Factors to Predict Atrial Fibrillation," Circulation, 2022. https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.121.057480

G. Petmezas et al., "Automated Atrial Fibrillation Detection using a Hybrid CNN-LSTM Network on Imbalanced ECG Datasets," Biomedical Signal Processing and Control, 2021. https://www.sciencedirect.com/science/article/abs/pii/S1746809420303323

J. Lei et al., "A deep learning method for beat-level risk analysis and interpretation of atrial fibrillation patients during sinus rhythm," Biomedical Signal Processing and Control, 2024. https://www.sciencedirect.com/science/article/pii/S1746809424010863

M. Alreshidi and M. Alsaffar, "FedCL: An Atrial Fibrillation Prediction System Using ECG Signals Employing Federated Learning Mechanism," ResearchGate preprint, 2024. https://www.researchgate.net/publication/383879227

S. Siddiqui et al., "The Role of Generative Artificial Intelligence and Large Language Models in Atrial Fibrillation Care," Cardiology in Review, 2025. https://pubmed.ncbi.nlm.nih.gov/40930127/

E. Zvuloni et al., "End-to-end risk prediction of atrial fibrillation from the 12-Lead ECG by deep neural networks," Journal of Electrocardiology, 2023. https://pubmed.ncbi.nlm.nih.gov/37774529/

S. Khurshid et al., "ECG-Based Deep Learning and Clinical Risk Factors to Predict Atrial Fibrillation," Circulation, 2022. https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.121.057480

D. R. Santos et al., "Feasibility Analysis of Federated Neural Networks for Explainable Detection of Atrial Fibrillation," IEEE HealthCom, 2024. https://arxiv.org/abs/2410.19781

N. Gadde et al., "Automated Detection of Atrial Fibrillation Using Deep Convolutional Neural Networks: Advancing Cardiac Diagnostics through AI-Driven ECG Analysis," *ResearchGate*, vol. 12, pp. 102–109, 2024. researchgate.net

Y. Jia et al., "An End-to-end Deep Learning Scheme for Atrial Fibrillation Detection," Computing in Cardiology, 2020. https://www.cinc.org/archives/2020/pdf/CinC2020-106.pdf

S. Andersen et al., "A Deep Learning Approach for Real-Time Detection of Atrial Fibrillation," Expert Systems with Applications, 2019. https://www.sciencedirect.com/science/article/abs/pii/S0957417418305190

G. Petmezas et al., "Automated Atrial Fibrillation Detection using a Hybrid CNN-LSTM Network," Biomedical Signal Processing and Control, 2021. https://www.sciencedirect.com/science/article/abs/pii/S1746809420303323

J. Konečný et al., "Federated Learning: Strategies for Improving Communication Efficiency," arXiv preprint, 2016. https://arxiv.org/abs/1610.05492

P. Kairouz et al., "Advances and Open Problems in Federated Learning," Foundations and Trends in Machine Learning, 2021. https://arxiv.org/abs/1912.04977

K. Singhal et al., "Large Language Models Encode Clinical Knowledge," Nature, 2023. https://www.nature.com/articles/s41586-023-06291-2