Towards Explainable Federated Learning in Healthcare: A Study on Heart Arrhythmia Detection

Sileshi Nibret Zeleke[1,*], Mario Bochicchio[1,2]

[1]Department of Computer Science, Università Degli Studi di Bari Aldo Moro, Via Edoardo Orabona, 4, 70125, Bari, Italy
[2]Digital Health National Lab, CINI - Consorzio Interuniversitario Nazionale per l'Informatica, Roma, Italy

Abstract

Advancements in artificial intelligence (AI) hold promise for revolutionizing healthcare, but challenges related to model explainability and data availability persist. Federated learning (FL) offers a privacy-preserving solution for health data usage by enabling decentralized model training without data sharing. In this study, we propose a novel framework that integrates explainable AI (XAI) with FL, specifically targeting heart arrhythmia detection. AI based heart arrhythmia detection usually relies heavily on the analysis of electrocardiogram (ECG) signals. We propose an attention based temporal convolutional network (TCN) for the analysis of time-series ECG data. The FL approach enables collaborative training across multiple local AI models without compromising local data or patients privacy. The attention mechanism is employed to enable the model to capture long-term dependencies across multiple time steps and weight different parts of the signal accordingly for improved classification performance. From the attention layer of the network we extract the attention weights for explanation. These weight values show important property of the model as well as input signal. To evaluate the proposed approach, we conducted experiments using the MIT-BIH arrhythmia dataset comprising 300 time steps per full beat cycle and 5 major arrhythmia classes. Our model achieved a balanced accuracy of 93.5%, F1-score of 98.7%, and 93.4% G-mean score in the presence of imbalanced data distributions. Notably, the explainability aspect of our model demonstrated consistent explanations comparable to expert benchmarks from related literature, enhancing trust and interpretability. Overall, this study underscores the potential of explainable FL with attention-based TCN architecture to enhance arrhythmia detection by prioritizing both privacy-preservation, performance and interpretability.

Keywords

XAI, PPFL, Attention weight, Heart arrhythmia, TCN, Healthcare

1. Introduction

A heart arrhythmia is an abnormal heartbeat, which can be caused by the heart beating too quickly (tachycardia), too slowly (bradycardia), or in another irregular manner [1]. The heart pumps blood effectively throughout the body when it beats in a regular, synchronized rhythm. However, this capacity to pump blood efficiently can be impacted by arrhythmias, which disturb the heart's rhythm. Arrhythmias can occur due to various reasons, including dysfunctions in the heart's electrical system, structural heart problems, electrolyte imbalances, and other factors. They can range from being harmless to life-threatening. Diagnosis of arrhythmia's

EXPLIMED: First workshop on explainable artificial intelligence for medical domain, 19-20 October 2024, Santiago de Compostela, Spain

*Corresponding author.

sileshi.zeleke@uniba.it (S. N. Zeleke); mario.bochicchio@uniba.it (M. Bochicchio)
0009-0006-8172-9646 (S. N. Zeleke); 0000-0002-9122-6317 (M. Bochicchio)

typically involves a combination of medical history assessment, physical examination, and diagnostic tests such as ECG. ECG is a one-dimensional time series signal that displays the electrical activity of the heart and aids in the diagnosis of cardiac disease as shown in Fig.1.

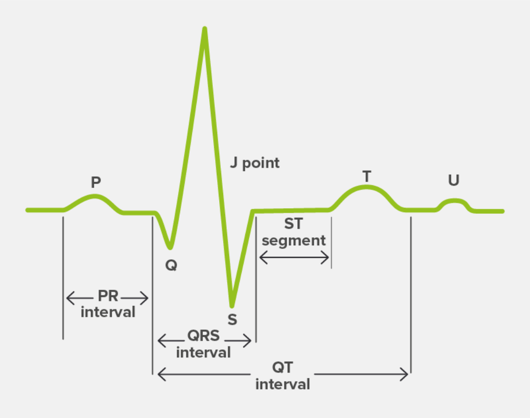


Figure 1: ECG signal key region labels of normal heart beat [2].

Classification of ECG signals primarily relies on the analysis of the PR interval, ST segment, P-wave, QRS-complex, and T-wave to establish the equipotential line and detect abnormal heart rhythms based on the position and amplitude of the waves. The growing availability of extensive medical data and the increase in computational power, AI has become more important in clinical practice for early disease detection, accurate diagnosis, predictions, and prognosis. However, training AI models requires centralized datasets, where a central server has access to the data of all patients. This can create privacy concerns for patients who may not want to share their personal information, and the regulatory requirements within the healthcare industry may also limit the sharing of sensitive information [3]. As a result, this leads to the existence of data silos, which will limit the use of traditional ML-based solutions. Moreover, privacy issues and data silos might limit the availability of data, which can hinder the adoption of large-scale healthcare systems and introduce biases into machine learning techniques. As a result, these difficulties may make the current health inequities worse [4]. FL in a medical system is a technique where learning is performed using shared model parameters instead of raw medical data. In FL a global model is typically deployed on a central server, while client models, which consist of subsets of local models stay on the data source location. Using their local data, the selected clients train their models locally before sending the update to the server. FL can be applied at a personalized level, where individual patients have local models, or at an institutional level, where hospitals with collections of patient data serve as local models. The latter approach is more applicable in studies because the data from these entities are more organized and suitable for training.

In addition to privacy-preservation, interpretability is also an important component for healthcare applications. XAI helps models to become more interpretable and transparent, enabling healthcare professionals to comprehend the underlying decision-making processes. This integration makes models more reliable while also making it easier to identify biases, mistakes, and odd trends in the data. Combining FL with XAI makes it possible to create more understandable, trustworthy, and privacy-preserving AI systems, all of which are necessary for the creation of dependable and morally sound applications. Given this ideas, this study is aiming to implement the combination of FL and XAI for heart arrhythmia disease detection.

In this work, we analyze the application of FL coupled with explainability to an arrhythmia detection model using ECG data. Additionally, we explore the advantages and challenges of using such an approach in scenarios where data is not independently and identically distributed (non-IID) among clients. We emphasize good practices for working with such data in a distributed setting.

2. XAI and FL in healthcare

Most of existing researches on healthcare domain are focused on either the interpretability of the AI models or the privacy-preserving approach of AI systems. However there are some studies like [5] and [6] applied XAI and FL techniques for domains like finance, automated vehicle networking, in 6G networks [7], and transportation [8], [9]. Another study post-hoc explainability technique based on integrated gradients has gained traction [10]. With the help of this interpretability, a user can determine whether a particular feature's value will influence a decision in a favorable or negative way.

However non-significant number of studies were focused on applying explainable FL for healthcare. Two key requirements for trustworthy AI are privacy preservation and explainability, especially when dealing with medical data from various health institutions that cannot be shared due to privacy concerns. To address this shortcoming, the study [11], leveraged FL as a paradigm for collaborative training without any disclosure of private data in the detection of Parkinson disease. The study [12], proposed cloud-based framework, where FL implementations of high-performance boosting classifiers to solve unmet mucosa-associated lymphoid tissue (MALT) lymphoma with 4805 patients with primary Sjogren's Syndrome and to provide interpretable risk factors. However the study does not provide the data distribution between the local model. To study the non-independent and identically distributed (non-IID) nature of ECG data, that may lead to non convergence of the FL-based algorithm, the study [13] optimized the FL-based algorithm using a sharing strategy for partial ECG data of each medical institution combined with elastic weight consolidation. However this sharing strategy makes each medical institution share ECG data to the central server which leads to violation of the idea of privacy-preserving. To protect knee Osteorthritis patient confidentiality and explain convolutional models using x-ray images the research [14], achieved interpretability of deep learning models trained on pre-trained models. Lin et. al [15], proposed class-level contribution explainable FL based on comparable prototypes collaboration for multi-site medical image classification.

While several studies such as [16],[17],[18], [19] and [20] have addressed aspects of the FL for heart arrhythmia detection, a comprehensive and explainable solution remains elusive. For instance, [17] explored privacy-preserving methods to train AI models over heterogeneous 12-lead sensor data collected from six heterogeneous source based on long-short-term memory models compared with central learning to conclude that FL is well-suited for cloud-edge architecture. However the study does not perform the explainability aspect of the models. Similarly, the study by [18] proposed weighted FL approach is proposed for ECG arrhythmia classification providing valuable insights into considers the heterogeneity of data distribution among multiple clients in FL settings. Despite these advancements, existing approaches often falter in achieving the overarching goal proposed in this study. While they may offer partial solutions or address specific components of the problem, they frequently lack the integrative framework or scalability required to fully realize the proposed solution. Moreover another study [20] aims to solve the data variation, label issue and privacy. Semi-supervised approach is proposed and achieved acceptable classification result. More related to our proposed methodology the study [21], proposed FL and interpretable heart arrhythmia classification using convolutional neural network (CNN). But the study failed to consider the data distribution and non-iid nature of real-world FL implementation. As such, there exists a significant gap between current research efforts and the comprehensive solution advocated in this work. To the best of our knowledge, either centralized or federated schemes are used in most suggested investigations on the application of attention-based AI models on time-series data related to heart arrhythmia to enhance the classification performance rather than exploring the interpretability advantage it have into consideration. Furthermore, the bulk of methods concentrate on the performance of the models, with only a small number of studies discussing their XAI aspect and realistic data distribution of privacy-preserving approach.

3. Proposed system architecture

Our proposed approach is composed of local and global model components as presented on Fig. 2. More detailed description of the components is presented in the following subsections. First we present the dataset we used then we continue how pre-processing was done. Next we present the TCN with self-attention system for model training and explanation. In the last section we present the FL process experimental settings, and evaluation metrics we used in the study.

3.1. Dataset

We used MIT-BIH dataset [22] consisting of 47 (25 men and 22 women) samples recorded for 30 minutes. It contains five major classes namely normal (N), supra-ventricular premature (S), Ventricular escape (V), fusion (F) and unclassified (Q). The normal class contains left/right bundle branch block(LBBB or RBBB), atrial escape, and nodal escape sub classes. However, the S class consists of atrial premature(APB), aberrant atrial premature, nodal premature and supra-ventricular premature. Moreover, V class contains two sub-classes premature ventricular contraction(PVC) and ventricular escape. The dataset is highly unbalanced between the classes and this is due to the nature of arrhythmia conditions representation in the real scenario as shown in Table 1.

3.1.1. ECG signal pre-processing

ECG signals are usually noisy due to factors such as electrode-skin interface, motion, electrical interference or baseline wonder. For non-stationary signal like ECG, discrete wavelet transform

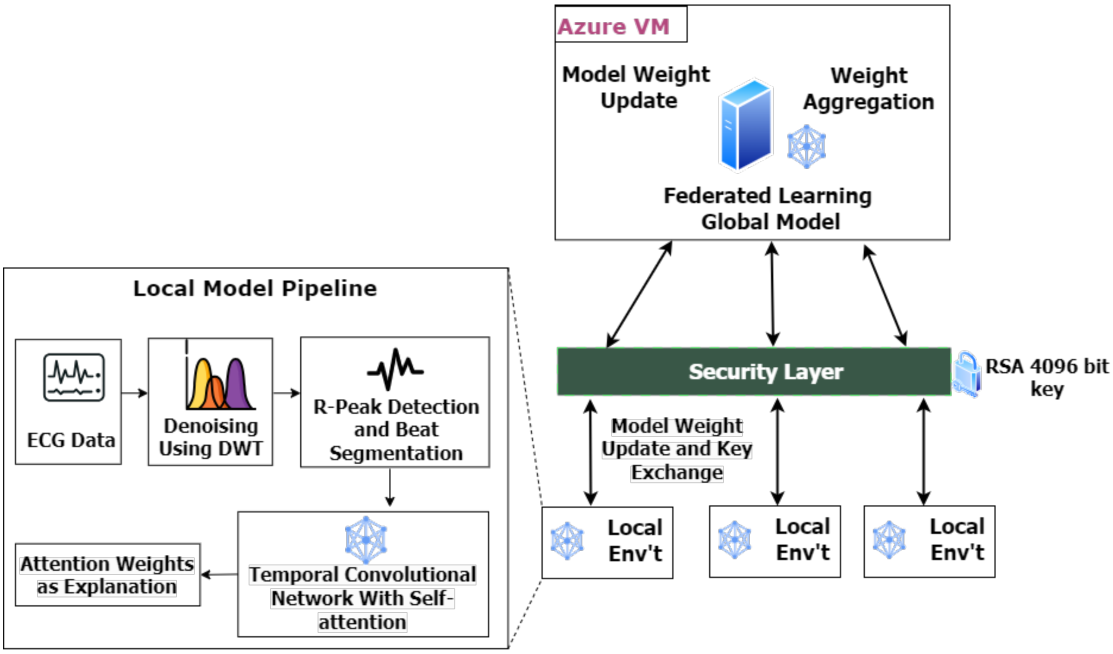

Azure VM
Model Weight Update
Weight Aggregation
Federated Learning Global Model
Security Layer
RSA 4096 bit key
Model Weight Update and Key Exchange
Local Env't
Local Env't
Local Env't
Local Model Pipeline
ECG Data
Denoising Using DWT
R-Peak Detection and Beat Segmentation
Temporal Convolutional Network With Self-attention
Attention Weights as Explanation

Figure 2: Proposed explainable FL system architecture

Table 1
Data distribution of arrhythmia samples of Mit-bih dataset into five major classes

Class description	Number of samples
Normal, left, right bundle branch block beats, atrial escape, and nodal escape beat (N)	90588
Supra-ventricular, nodal, aberrant atrial and atrial premature beat (S)	2779
Premature ventricular contraction and ventricular escape beat (V)	7236
Fusion of ventricular and normal (F)	803
Paced beat, fusion of paced and normal beat and unclassified(Q)	8039

is particularly useful. To reduce the noise we decomposed signal into nine levels of db5 wavelet coefficients for noise reduction as it was proven to be efficient in [23]. Then, R-peak detection and beat level segmentation was done. Since the dataset is already R-peak annotated, a window size of 300 samples were taken as input. According to the mean RR interval, 99 samples from the left side of the QRS mid-point, 201 samples after QRS mid-point, and the QRS mid-point itself were chosen as a segment, thereby one cardiac cycles are included.

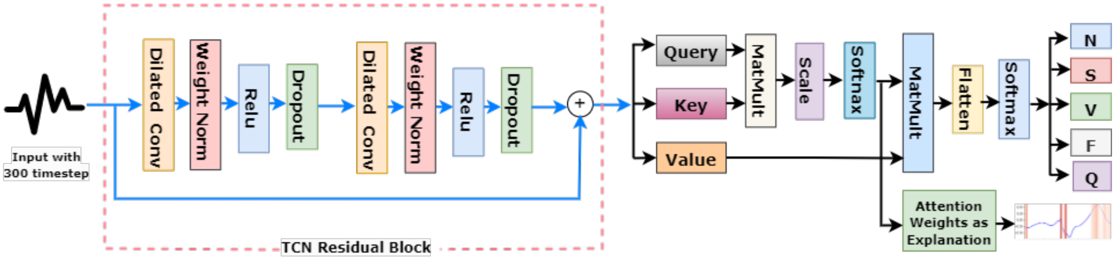


Figure 3: Attention based TCN architecture

3.2. Model architecture

For time-series data like ECG, models such as long short-term memory or recurrent neural networks could be applied. However, such models are suffered from gradient stability and signal irregularities. Specifically for FL applications gradient stability is more important for collaborative learning and weight sharing. Given this, we propose attention based TCN model which contains two important components: TCN network and self-attention mechanism as shown in Fig. 3.

3.2.1. TCN network

The TCN network is a time-series specialized deep learning architecture designed based on the foundation of CNN to capture the temporal dependencies and to preserve the ordering of the input sequence. The main components of TCN are the causal, and dilated convolution. In a causal convolution, the output at time step t is convoluted only with the input elements from the current and previous time steps. This ensures that the network's predictions at any given time step are dependent only on the past and present inputs, preserving the temporal ordering of the data and preventing information leakage from future time steps [24].

Dilated convolutions systematically increasing the receptive field of a convolutional layer without increasing the number of parameters or the amount of computation. In TCN, the dilation factor is exponentially increased with the depth of the network, allowing the model to efficiently capture hierarchical patterns across multiple time scales. The residual block is composed of a two dilated convolutional layers, accompanied by a non-linear activation function, ReLU. Weight normalization layer after convolution helps the model to stabilize the learning process and to mitigate vanishing gradient issue. To enhance regularization within the network, weight normalization and dropout layers are incorporated into each layer. The input of the residual block is then directly summed with its output, facilitating the unimpeded transfer of information across layers. This design mitigates the issue of information loss that can arise from excessively deep networks [25].

3.2.2. Attention mechanism

The attention mechanism has the ability to prioritize important features by assigning higher learning weights to them. The core idea of concatenating with the TCN residual block is to enhance the training performance by emphasizing the more informative temporal segment of ECG signal in addition to the improved feature representation. The attention mechanism we used is self-attention, meaning that the query and key-value pairs are derived from the outputs of the preceding TCN-based residual block. The self attention matrix as presented in [26] is calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}) \cdot V \tag{1}$$

Where Q,K and V represented the the query, key, and value vectors derived from the input X using weight matrices. Moreover, the d_k is the dimension of input data.

3.2.3. Explainability

Attention weights provide interpretability by assigning importance scores to input sequence, indicating their contribution to the model's predictions. Inspired by human visual attention, attention mechanisms in deep learning models learn these weights through additional layers or components. Attention scores are computed using a dot product or similarity function between a query vector and a set of key vectors, derived from the input features as shown in Eq. 2. The dot product measures the alignment between the query and key vectors, and the resulting attention scores are normalized using a softmax function.

$$\text{Attention weights: softmax}(\frac{QK^T}{\sqrt{d_k}}) \tag{2}$$

The attention scores create a weighted sum of input features, called the context vector, which is fed into subsequent layers for the final prediction. In our model, the attention weights are visualized to explain the model's focus on specific regions of the ECG signal. To visualize the influential regions of an ECG signal we follow the following steps:-

- After training the attention based TCN model on ECG signal to classify the arrhythmia, the attention mechanism will learn to assign attention weights w_i to different parts of the ECG signal.
- To obtain w_i corresponding to each time step or input feature, input ECG signal passes through the model.
- Then w_i are normalized using a softmax function to ensure that the weights sum up to one. This will facilitate the interpretation of the weights as a probability distribution over the input features.

$$\text{softmax}(w_i) = \frac{e^{w_i}}{\sum_{j=1}^{300} e^{w_j}} \tag{3}$$

- Convert the 1D attention weights and the corresponding ECG signal into a 2D matrix by repeating the attention weights along one axis and the ECG signal along the other axis to create a matrix with dimensions (300, 2), where one column represents the attention weights, and the other column represents the ECG signal. Then heatmap is generated to have the ECG signal on one axis, the attention weights on the other axis, and the colors in the heatmap will represent the intensity of the attention weights at each time step.

Table 2
Data distribution between clients.

Client	N	S	V	F	Q
Client 1	30113	866	2396	256	2626
Client 2	29757	879	2341	244	2677
Client 3	30719	1034	2499	303	2736

- Analyze the heatmap to identify the region R of the ECG signal that the model is focusing on, as indicated by w_i significantly higher than a threshold τ:

$$R = \{i | w_i > \tau,\ i \in \{1, \ldots, 300\}\} \tag{4}$$

3.3. FL process

The FL process is fundamentally designed as a server-client approach to realize a distributed privacy-preserving learning. The two main components are global and local models. The global model initiates the computation round by transmitting initial weight parameters to the local models, which subsequently perform training on their respective local datasets as shown in Fig. 2. Upon completion of each round, the global model aggregates the updated parameters on the central server. We employ Flower framework [27], which is a comprehensive FL framework that can be used to conduct large-scale FL experiments and consider highly heterogeneous FL device environments. To reduce communication overhead between the global and local models, we employ the Federated Averaging (FedAvg) algorithm, which enables multiple local updates before aggregation, thereby minimizing communication costs while maintaining model performance.

Since we are using a dataset collected in one center we need to split the dataset among the local systems as realistic as possible. For this purpose we adopted the splitting algorithm proposed by [28]. The algorithm partitions a dataset into client-specific subsets based on a set of non-proportional number of data points. By iterating through the clients and labels, randomly sample number of data points for each label. The samples are removed from the original dataset and finally the partitioned dataset is returned as a list of client-specific subsets. Since the data distribution between the simulated clients are non identical in which each class have different count of samples and labele in their local dataset. The distribution is shown in Table 2. The local dataset is split into 70% training set, 10% validation and 20% testing set. To secure the communication between the client and global model we implemented one of the most widely used asymmetric cryptosystem which is Rivest–Shamir–Adleman with 4096 bit key for secure weight transmission.

3.3.1. Experimental settings

We used a desktop computer with Intel Core i7-4770 CPU, 6GB RtX1080 GPU, 16GB RAM, and Windows 10 operating system. The global model is deployed on Microsoft Azure virtual machine and the local client were initiated on a local terminal. As it is common when using dilated convolutions, we increase dilation factor d exponentially with the depth of the network. A total of four cascaded encoding blocks were incorporated, taking into account the size of the model.

Given the distributed nature of FL, traditional hyperparameter optimization techniques like grid search or random search may not be feasible due to the high communication costs. Given this, we selected the hyperparameters empirically through iterative experimentation. Since we employed the same TCN architecture as the backbone, as described in the study [25], we used their hyperparameters as initial values for our empirical experiments. The two other critical parameters, namely num_filters, which represents the number of one-dimensional dilated filters in each dilated convolution were set to 64, and kernel_size, which denotes the kernel size of each filter set to 30. Moreover, we utilized Adam optimizer with exponentially decay learning rate with 0.001 initial rate. For global model update we experimented 5, 10, and 15 communication rounds. Since the each local model is trained with small number of local data and to protect the models from over-fitting we experiment using 10, 20, 30 and 40 epoch size with 32 batch size. Moreover, to effectively utilize disproportional distribution of the data between classes we used focal loss as it was proven to be effective loss functions by [2].

3.3.2. Evaluation metrics

Various evaluation metrics that take into account both class imbalance and the model's classification performance we considered. Balanced accuracy, F1-score and G-mean was selected as these metrics provide a more reliable measure of performance in the presence of class imbalance [29]. Balanced accuracy is the average of recall and specificity which is average per-class accuracy. However, F1-score balances the trade-off between precision and recall by taking the harmonic mean of precision and recall for each class. Furthermore, The G-mean is the geometric mean of sensitivity and specificity. For multi-class, it's the geometric mean of the sensitivity for each class.

4. Result and discussion

In this section the model performance is analyzed based on evaluation matrices. The influence of communication rounds between local and global model were discussed in addition to the explaianbility of the global and local model. Finally, the study is compared to the state-of-the-art (SOTA) studies in terms of classification performance and computational complexity.

4.1. Experimental results and analysis

The performance of attention-based TCN in FL demonstrated promising results in accurately classifying ECG signals, with a balanced accuracy of 93.5% and F1-score of 98.7% when trained for 10 communication rounds in data as shown in Table 3. The model was able to effectively capture the temporal dependencies in the ECG signals and focus on the critical regions of the input signal, as evidenced by the attention weights calculated from the attention mechanism. The proposed approach predominantly concentrated on the QRS segment, T wave, and in cases where the P wave segment was absent, it primarily targeted the portion of the P wave that was present within the segment. Moreover, the experimental studies show minority classes like fusion and supra-ventricular arrhythmia were detected acceptable accuracy against there relatively small number of samples in the training set.

In terms of communication rounds between the local model and global aggregator we experimented 3 scenarios (i.e. 5, 10 and 15) rounds. At the end of each round the locally fitted model for specified epoch sends the weights for global model. As depicted in Table 3, the balanced weight shows increase as number of round increase. However, this is not always the case for example the model have a lower accuracy in a model trained for 10 rounds for 20 epoch. Furthermore, due to the small size of the dataset for such deep learning model, we found that the model tends to overfit when we train it for more than 50 epoch in any of the update rounds. Generally, the model perform better when we train the model for 10 rounds and 10 epoch, in this round the 93.5%, 98.7%, and 93.4% for balanced accuracy, F1-score and g-mean respectively. In addition on this setting all the classes perform relatively better. Moreover we used non-FL central training based on same model architecture and hyper-parameter as a baseline. The results showed that the FL approach improved the balanced accuracy by 3.9% with respect to the central learning.

As we increase model update rounds, the model performance increase a bit as shown in Table 3. however, this performance enhancement comes with the cost of communication and computation overhead as more weight needs to be transmitted between the client devices and central server in addition to the time complexity. For example to finish the training of 15 rounds with 40 epoch parameter takes around 22 hours of training. Early stopping mechanisms are commonly used to prevent overfitting and efficient computational usage, however in the context of FL early stopping could not be feasible due to heterogeneity and non-applicability of global validation mechanism.

4.2. Explainability

To achieve explainability in an arrhythmia detection model, it is imperative to thoroughly understand the morphological components of the ECG signal. The amplitude and duration of components are critical features in ECG analysis, as they significantly contribute to the interpretation and evaluation of cardiac function [30]. As depicted in Fig. 4, the red segments indicate the regions of the heartbeat that hold greater significance in the process of predicting a specific class. This indicates attention weights calculated from the attention mechanism reveals that the classifier focuses on the critical regions of the input signal. The findings derived from the XAI module can subsequently aid clinical practitioners in diagnosing underlying health conditions.

Moreover, the proposed approach predominantly concentrates on the QRS segment, T wave and in cases where the P wave segment is absent, it primarily targets the portion of the P wave that is present within the segment as shown in Fig. 4 (e). When we consider individual arrhythmia conditions our model identified specific part to of the signal to to distinguish. For example PVC is abnormal cardiac rhythms, characterized by a widened an oddly shaped QRS complex accompanied by a T wave with an opposing direction with no preceding P wave [31]. As illustrated in Fig. 4(c), the model accentuates the widened QRS complex and the T wave, while also emphasizing the isoelectric line after the T wave that signifies probability of the

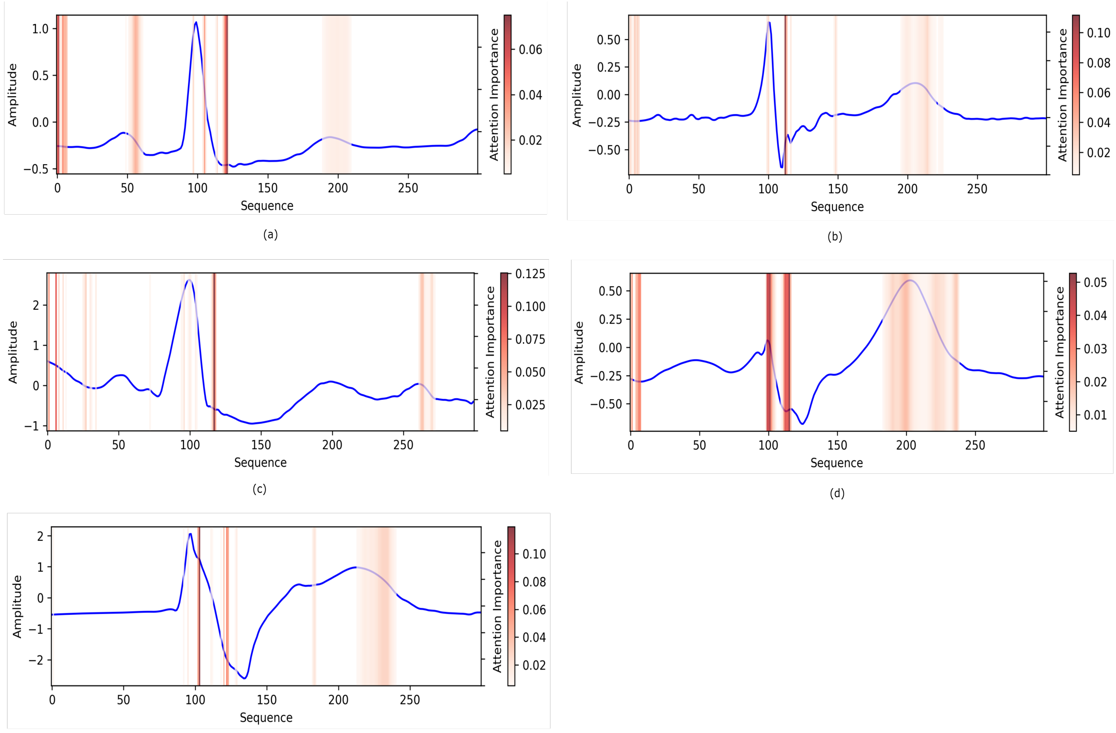
Table 3
Classification performance of proposed approach in different scenarios, central learning (CL) as baseline, training epoch in each local training, balanced accuracy, F1-score, G-mean and accuracy of each arrhythmia class

Update Rounds	Epochs	Balanced Accuracy	F1-score	G-mean	N	S	V	F	Q
5	10	0.916	0.986	0.913	0.995	0.822	0.950	0.826	0.990
	20	0.916	0.986	0.912	0.995	0.810	0.962	0.826	0.988
	30	0.914	0.984	0.911	0.993	0.816	0.954	0.826	0.987
	40	0.919	0.982	0.908	0.992	0.821	0.959	0.824	0.984
10	**10**	**0.935**	**0.987**	**0.934**	**0.994**	**0.871**	**0.950**	**0.870**	**0.994**
	20	0.913	0.986	0.909	0.994	0.798	0.958	0.826	0.992
	30	0.926	0.983	0.924	0.989	0.847	0.960	0.848	0.988
	40	0.922	0.985	0.921	0.992	0.822	0.958	0.848	0.990
15	10	0.924	0.986	0.921	0.994	0.804	0.964	0.870	0.990
	20	0.923	0.984	0.920	0.990	0.871	0.960	0.804	0.990
	30	0.913	0.984	0.909	0.992	0.814	0.956	0.801	0.983
	40	0.913	0.987	0.909	0.995	0.840	0.958	0.783	0.990
CL	39	0.896	0.985	0.889	0.992	0.866	0.961	0.794	0.989

sample to be premature beat.

Suptra-ventricular encompass a narrow QRS complex, often accompanied by undetectable or negative P waves [31]. Fig 4(b) confirms that the model have more attention weights to recognize samples associated with absent P waves, as well as highlighting specific samples within the QRS complex and more importantly focusing on the S wave to detect its narrowness. Therefore, the model's findings align with the diagnostic criteria for supra-ventricular arrhythmia, demonstrating its potential for accurately identifying such signals. Normal rhythm is characterized by a regular rhythm, an upright P wave preceding each QRS complex, and a consistent PR interval [31]. Our model also highlight these regions of the ECG as shown in Fig.4 (a) focusing on the left half of the QRS complex and the T wave segment.

Furthermore, pacemaker rhythms under Q class are mostly based on pacing site or cardiac anatomy. However, usually characterized by the presence of vertical spikes or artifacts of brief duration. The pacing artifact is usually followed by a paced QRS complex, which represents the depolarization of the ventricles in response to the pacing stimulus. These spikes, as illustrated in Fig. 4 (e), indicating that these samples hold greater significance in the classification of a specific rhythm as a paced rhythm. As depicted in Fig. 4(d) the correctly classified ventricular sample reveals that the model focused the wideness of the QRS complex, longer activation time of the R-peak in addition to the abnormal beat characteristics. This result is consistent with human expert baseline characterization by [31].



(a)

(b)

(c)

(d)

(e)

Figure 4: Attention weight heatmap of correctly classified samples using attention based TCN model, (a): correctly classified N class;(b):correctly classified S class ;(c):correctly classified F class ;(d):correctly classified V class;(e):correctly classified Q class.

4.3. Comparison with related work

We compared our result with the state-of-the art methods. Most of the recent studies are using accuracy as their main evaluation metrics, we calculate the accuracy for such comparison. However, accuracy is an unreliable metric for imbalanced datasets, as it can yield misleadingly high values by predominantly predicting the majority class, disregarding the minority classes' significance, and hindering fair model comparisons [29]. In the investigation by [2], the application of TCN with self-attention in a central learning approach yielded an accuracy of 87.81% using the imbalanced MIT-BIH dataset. In contrast, our FL approach demonstrated superior performance, attaining an accuracy of 98.4% with the global model. This comparison underscores the benefits of employing FL, which not only enhances privacy preservation but also bolsters model performance. Another study by [32] proposed weighted FL approach to achieve accuracy of 93% on local model. However, our study showed superior performance achieving a higher balanced accuracy which is 93.5% and weighted recall of 99.5%, demonstrating temporal neural networks coupled with attention mechanism are promising approach for time-series signals like ECG.

5. Conclusion

The rapid advancements in the AI and practicality of personal bio-metric sensors have led to a substantial augmentation in their utilization, consequently resulting in the accumulation of vast amount of data. However, the access of these data is limited due to data protection regulations and privacy concerns. In this context FL plays crucial role to train efficient models from data silos with out having compromising privacy. This study presented robust deep learning models for time-series ECG data that can determines morphological characteristics of waveform. The proposed model utilizes a self-attention layer that calculates and analyzes data, determines the dependence between segments and hidden features, and classifies them. Furthermore this study presents a valuable contribution on the explainability of time-series deep learning models using attention mechanisms, in which the attention weights are used as post-hoc explanation. These results are consistent with human expert explanations in the studies. The findings can subsequently aid clinical practitioners in diagnosing underlying health conditions. Overall, the attention-based TCN model in FL provides a promising approach for accurate and explainable ECG signal classification.

Moreover, integration of FL, TCN and attention mechanisms not only protect the privacy of the data owners and provides post-hoc explanation but also to enhance the classification performance. However, the time and memory complexity needs further investigation. As a limitation of this study, since the study is based on single lead ECG data a 12-lead ECG would have different characteristics. Incorporating human experts to validate the explanations generated by the model is an important element for future work. This will help to build trust in the model's predictions and increase the acceptability. Additionally, investigating the the communication overhead in real-world FL setup could be a future research direction.

Acknowledgments

This study is supported by the European Union – Next Generation EU, in the context of the National Recovery and Resilience Plan Investment Partenariato Esteso PE8 ”Conseguenze e sfide dell'invecchiamento”, Project Age-It (Ageing Well in an Ageing Society).

References

[1] V. Amar, K. H. Juet, E.-M. Ahmed, D. Scott, Cardiology in a Heartbeat, Scion Publishing Ltd, 2022.

[2] Pezoulas, V. C., et al, An explainable and trustworthy ai framework for federated learning: a case study in rare autoimmune diseases, in: IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology. IEEE,, 2023.

[3] M. Paul, M. A. F. Leandros Maglaras, I. Almomani, Digitization of healthcare sector: A study on privacy and security concerns, ICT Express 9 (2023).

[4] P. T, M. H, A. R, Artificial intelligence and algorithmic bias: implications for health systems, J Glob Health (2019). doi:10.7189/jogh.09.020318.

[5] G. Wang, Interpret federated learning with shapley values, CoRR abs/1905.04519 (2019). URL: http://arxiv.org/abs/1905.04519. arXiv:1905.04519.

[6] T. Awosika, R. M. Shukla, B. Pranggono, Transparency and privacy: The role of explainable ai and federated learning in financial fraud detection, 2023. arXiv:2312.13334.

[7] R. A., D. P., M. F., S. D., F. M.C., N. G., S. G., V. A., M. D., R. D., et al., Federated learning of explainable ai models in 6g systems: Towards secure and automated vehicle networking, Information 13 (2022). doi:10.3390/info13080395.

[8] J. Fiosina, Explainable federated learning for taxi travel time prediction, in: International Conference on Vehicle Technology and Intelligent Transport Systems, 2021.

[9] J. Fiosina, Interpretable privacy-preserving collaborative deep learning for taxi trip duration forecasting. in: Klein, c., jarke, m., helfert, m., berns, k., gusikhin, o. (eds) smart cities, green technologies, and intelligent transport systems. vehits smartgreens 2021. communications in computer and information science, 2022. doi:10.1007/978-3-031-17098-0_20.

[10] D. Janzing, L. Minorics, P. Bloebaum, Feature relevance quantification in explainable ai: A causal problem, in: S. Chiappa, R. Calandra (Eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, volume 108 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 2907–2916.

[11] B. J.L.C., D. P., F. Marcelloni, A. Renda, R. F., Federated learning of explainable artificial intelligence models for predicting parkinson's disease progression in: Longo, l. (eds) explainable artificial intelligence, in: Communications in Computer and Information Science, volume 1901, 2023. doi:10.1007/978-3-031-44064-9_34.

[12] R. A., H. M.S., M. G., et al., Federated learning-based ai approaches in smart healthcare: concepts, taxonomies, challenges and open issues, Cluster Comput 26 (2023). doi:0.1007/s10586-022-03658-4.

[13] Z. Mufeng, W. Yining, L. Tao, Federated learning for arrhythmia detection of non-iid ecg, in: 2020 IEEE 6th International Conference on Computer and Communications (ICCC), 2020, pp. 1176–1180. doi:10.1109/ICCC51575.2020.9344971.

[14] R. R. Hossain, C. S. Abanti, K. Marufa, R. A. M. Golam, Privacy-preserving knee osteoarthritis classification: A federated learning approach with gradcam visualization, in: 2023 26th International Conference on Computer and Information Technology (ICCIT), 2023, pp. 1–6. doi:10.1109/ICCIT60459.2023.10441001.

[15] B. Lin, J. Wang, Y. Dou, Y. Zhang, W. Yue, G. Yu, J. Yin, Fedcce: A class-level contribution explainable federated learning based on comparable prototypes collaboration for multi-site medical image classification, in: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2023, pp. 2085–2090. doi:10.1109/BIBM58861.2023.10385560.

[16] C. R. S., Favaro, L. C., Federated learning applied to arrhythmia detection on electrocardiograms, in: 2023 IEEE Virtual Conference on Communications (VCC), 2023, pp. 305–310. doi:10.1109/VCC60689.2023.10474963.

[17] D. M. J. Gutierrez, H. M. Hassan, L. Landi, A. Vitaletti, I. Chatzigiannakis, Application of federated learning techniques for arrhythmia classification using 12-lead ecg signals, 2024. arXiv:2208.10993.

[18] R. N. Asif, A. Ditta, H. Alquhayz, S. Abbas, M. A. Khan, T. M. Ghazal, S.-W. Lee, Detecting electrocardiogram arrhythmia empowered with weighted federated learning, IEEE Access 12 (2024) 1909–1926. doi:10.1109/ACCESS.2023.3347610.

[19] M. Bochicchio, S. N. Zeleke, Privacy-preserving federated learning for in-home monitoring of elderly using wearable biometric sensors. in:francesca fracasso, francesca gasparini and frida milella, (eds), in: Proceedings of the 4th Italian Workshop on Artificial Intelligence for an Ageing Society, AIxAS 2023, Nov 9,2023, volume 3623, 2023.

[20] Z. Ying, G. Zhang, Z. Pan, C. Chu, X. Liu, Fedecg: A federated semi-supervised learning framework for electrocardiogram abnormalities prediction, Journal of King Saud University - Computer and Information Sciences 35 (2023). doi:10.1016/j.jksuci.2023.101568.

[21] A. Raza, K. P. Tran, L. Koehl, S. Li, Designing ECG monitoring healthcare system with federated transfer learning and explainable AI, Knowl. Based Syst. 236 (2022) 107763. doi:10.1016/J.KNOSYS.2021.107763.

[22] G. AL, A. LA, G. L, H. JM, I. PC, M. RG, M. JE, M. GB, P. CK, S. HE., Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals., Circulation. 101(23) (2000).

[23] M. R. Islam, M. Qaraqe, K. Qaraqe, E. Serpedin, Cat-net: Convolution, attention, and transformer based network for single-lead ecg arrhythmia classification, biomedical signal processing and control 93 (2024). doi:10.1016/j.bspc.2024.106211.

[24] Y. Lin, I. Koprinska, M. Rana, Temporal convolutional attention neural networks for time series forecasting, in: International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021, IEEE, 2021, pp. 1–8. URL: https://doi.org/10.1109/IJCNN52387.2021.9534351. doi:10.1109/IJCNN52387.2021.9534351.

[25] Y. Zhao, J. Ren, B. Zhang, J. Wu, Y. Lyu, An explainable attention-based TCN heartbeats classification model for arrhythmia detection, Biomed. Signal Process. Control. 80 (2023) 104337. doi:10.1016/J.BSPC.2022.104337.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008.

[27] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, N. D. Lane, Flower: A friendly federated learning research framework, CoRR abs/2007.14390 (2020). arXiv:2007.14390.

[28] W. Chorney, H. Wang, Towards federated transfer learning in electrocardiogram signal analysis, Comput. Biol. Medicine 170 (2024). doi:10.1016/J.COMPBIOMED.2024.107984.

[29] P. Thölke, Y. Mantilla-Ramos, H. Abdelhedi, C. Maschke, A. Dehgan, Y. Harel, A. Kemtur, L. M. Berrada, M. Sahraoui, T. Young, A. B. Pépin, C. E. Khantour, M. Landry, A. Pascarella, V. Hadid, E. Combrisson, J. O'Byrne, K. Jerbi, Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data, NeuroImage 277 (2023). doi:10.1016/J.NEUROIMAGE.2023.120253.

[30] S. K. Berkaya, A. K. Uysal, E. S. Gunal, S. Ergin, S. Günal, M. B. Gülmezoglu, A survey on ECG analysis, Biomedical Signal Process Control. 43 (2018) 216–235. doi:10.1016/J.BSPC.2018.03.003.

[31] L. Atul, ECG made easy, Jaypee Brothers Medical Publishers, 2019.

[32] R. N. Asif, A. Ditta, H. Alquhayz, S. Abbas, M. A. Khan, T. M. Ghazal, S. Lee, Detecting electrocardiogram arrhythmia empowered with weighted federated learning, IEEE Access 12 (2024) 1909–1926. doi:10.1109/ACCESS.2023.3347610.