Application of Federated Learning Techniques for Arrhythmia Classification Using 12-Lead ECG Signals

Daniel Mauricio Jimenez Gutierrez, Hafiz Muhammed Hassan, Lorella Landi, Andrea Vitaletti, and Ioannis Chatzigiannakis
Sapienza University, Rome, Italy
jimenezgutierrez@diag.uniroma1.it, vitaletti@diag.uniroma1.it, ichatz@diag.uniroma1.it

Abstract. Artificial Intelligence-based (AI) analysis of large, curated medical datasets is promising for providing early detection, faster diagnosis, and more effective treatment using low-power Electrocardiography (ECG) monitoring devices information. However, accessing sensitive medical data from diverse sources is highly restricted since improper use, unsafe storage, or data leakage could violate a person’s privacy. This work uses a Federated Learning (FL) privacy-preserving methodology to train AI models over heterogeneous sets of high-definition ECG from 12-lead sensor arrays collected from six heterogeneous sources. We evaluated the capacity of the resulting models to achieve equivalent performance compared to state-of-the-art models trained in a Centralized Learning (CL) fashion. Moreover, we assessed the performance of our solution over Independent and Identical distributed (IID) and Non-IID federated data. Our methodology involves machine learning techniques based on Deep Neural Networks and Long-Short-Term Memory models. It has a robust data preprocessing pipeline with feature engineering, selection, and data balancing techniques. Our AI models demonstrated comparable performance to models trained using CL, IID, and Non-IID approaches. They showcased advantages in reduced complexity and faster training time, making them well-suited for cloud-edge architectures.

Keywords: Federated Machine Learning · Deep Learning · Low-power devices · Arrhythmia and Multilead ECG classification · 12-channel ECG · Medical Data Privacy.

1 Introduction

Significant advances in Artificial Intelligence (AI) based medical data analysis, i.e., big data, machine learning (ML), deep learning (DL), etc., drive innovation in MedTech [21, 43]. Numerous studies demonstrate how scalable data digitization, combined with data technologies, can significantly contribute to timeliness and resource effectiveness of prevention, diagnosis, and treatment [7]. Modern AI
models feature millions of parameters that need to be learned from sufficiently large, curated datasets to achieve clinical-grade accuracy while being safe, fair, equitable, and generalizing well to unseen data [61]. Recent publications provide evidence on how large volumes of data managed by independent and possibly heterogeneous low-power devices, often on a cross-border basis, are shared to drive AI systems that can outperform specialists or, at the very least, help them reduce their effort significantly [36]. Despite its advantages, the practical application of this technological evolution in healthcare environments faces certain difficulties [48]: the rate at which AI technologies are being developed is out of step with the rate at which they are being adopted in practice. The integration of AI in healthcare operations and services creates a solid and urgent demand to integrate medical data obtained from various edge devices [24]. However, since health data is susceptible and its use is strictly restricted, data like this is challenging to obtain [60]. Indeed, this is exacerbated in the case of personal information, where improper use, unsafe storage, data leakage, or abuse could violate a person’s privacy. Even if data anonymization could circumvent these limits, it is now widely accepted that deleting metadata such as a patient’s name or date of birth is frequently insufficient to maintain privacy [52].

Recently, Federated learning (FL) has been introduced as a decentralized training paradigm of an AI model that enables collaborative learning without exchanging the data itself [64]. FL utilizes distributed consensus algorithms that allow organizations to gain insights collaboratively so that patient data are not transferred outside the organization’s firewalls. Consider, for example, multiple hospitals (equipped with low-powered medical devices) collaboratively learning from a shared medical predictive model while keeping data sources in their original location in each hospital, decoupling the ability to do ML from the need to store the data in a single place. Some recent studies examine the performance levels achieved by AI models trained using FL techniques on isolated single-organizational data. They compare them with models trained on centrally hosted data, indicating comparable results [58]. Therefore FL offers a promising path with a potentially significant impact on large-scale precision medicine, resulting in models that make impartial judgments, best reflect an individual’s physiology, and are sensitive to uncommon illnesses while respecting governance and privacy issues.

In this work, we apply the FL paradigm to train AI models that can assist doctors in closely and efficiently monitoring arrhythmias, a common Cardiovascular disease (CVDs) recognized as one of the leading causes of death globally. Arrhythmia is a disorder of the heart rate (pulse) or heart rhythm, being too slow, fast or irregular. Today, the most effective tool to identify variations in heart rhythm along with other patterns of the heart’s electrical impulses is based on low-power Electrocardiography (ECG) monitoring devices [44]. Such devices use multiple sensors, and the ECG leads to capture the heart’s electrical activity from different angles and positions. Automatic analysis of the ECG signal has

---

1 https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
become one of the essential tools in diagnosing heart disease. It has been used as the basis for the design of algorithms for many years [55].

We introduce a multi-step privacy-preserving methodology for training AI models for arrhythmia detection and classification using 12-lead ECG recordings. The process can be applied to various ML techniques, ranging from deep neural networks to recurrent neural networks with long short-term memory. Moreover, the methodology is generic enough to be applied to several classification problems. A detailed feature extraction stage is included that examines the morphology of all the channels of the ECG signals both in the time domain and the frequency domain and the cardiac rhythm of consecutive heartbeats, resulting in about 650 features. The methodology introduces a novel feature selection stage where a subset of the features is selected so that the overall training time of the model and performance achieved is optimized.

We apply the FL paradigm over the PhysioNet in Cardiology Challenge 2020 dataset [49], which includes high-definition 12-lead ECG recordings obtained from 43,059 patients recruited by six disparate geographical centers. In contrast to other publicly available datasets of ECG recordings, the one considered here (a) provides the most extensive set of cardiac abnormalities, thus contributes significantly to training AI models that are capable of recognizing a broad range of cardiac arrhythmias; (b) the recordings provided by the six centers are heterogeneous as they include a different number of patients and with different distribution of arrhythmias. We evaluate the capacity of the resulting AI models to achieve equivalent performance when compared to state-of-the-art models trained (a) when the same data is collected in a central place – this allows us to study the performance following the currently dominant, centralized training approach; (b) when the data is available at multiple homogeneous devices, that it is Independent and Identically Distributed among the devices (IID) – this allows us to study the performance in a simplified federated environment; (c) when the data is available at multiple heterogeneous devices, that it is non-independent and identically distributed among the devices (Non-IID) – this is the most realistic scenario since the distribution of the different types of arrhythmias recorded varies between hospitals.

2 Previous related work

2.1 Arrhythmia detection

During the past years, the development of AI models has been based on free access ECG datasets, such as PhysioNet [40]. Probably the most diffused dataset for ECG signal analysis and arrhythmia detection among all those freely available is the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) Arrhythmia database [41]. The latter is because, first, the dataset includes recordings of all five arrhythmia classes suggested by the AAMI standards [38]. Second, as two cardiologists annotate all the recordings included, it makes it suitable for supervised and unsupervised learning techniques. On the other hand, the limitations of the MIT-BIH dataset in terms of (a) signal quality, since only
two ECG leads were used during the recordings, and (b) dataset size, since the different types of arrhythmias identified on the 48 patients, inevitably restrict the generality of the trained AI-models.

In recent years, machine-learning methods have rapidly detected cardiac abnormalities in 12-lead ECGs [6, 2]. Newly emerging deep-learning models have further achieved comparable performance to clinical cardiologists on many ECG analysis tasks [22] and arrhythmia/disease detection [10]. However, as high-quality, real-world ECG data is challenging to acquire, most deep-learning models are designed to detect only a tiny fraction of cardiac arrhythmias, owing to the limitations of the datasets. In addition, due to the database’s structure and annotations, many of those [56, 33] focus only on heartbeat classification. Most of the ones [50, 53] that use the MIT-BIH database for rhythmic arrhythmia classification, on the other side, only deal with a small number of arrhythmias, so their good performance results are not comparable with those of this paper due to the limited range of arrhythmias detected and the differences in database dimensions.

To encourage more multidisciplinary research, PhysioNet in Cardiology Challenge 2020 (Challenge 2020) [49] provided high-quality 12-lead ECG data obtained from multiple centers with a large set of cardiac abnormalities. The aim of Challenge 2020 was to identify clinical diagnoses from 12-lead ECG recordings, providing an opportunity to employ various advanced methods to address clinically important questions that are either unsolved or not well-solved [4].

Forty-one teams got selected during this challenge, and others were discarded due to failing to achieve high scores or appear at the conference. One can observe that most teams use common techniques by looking at the 41 teams’ papers published in PhysioNet [49]. Among these techniques, signal processing, DNNs, convolutional neural networks (CNNs), and end-to-end and multi-binary classifications are used by all of the top 10 teams. In addition, there are several important points 1) deep-learning methods were more popular than traditional methods in Challenge 2020; 2) all the teams that employed deep-learning methods used CNNs; and 3) none of the top 10 teams used hand-labeled features (except demographic features); they all adopted end-to-end models instead.

One can also notice that the three highest-ranking teams used the model ensemble [46, 67, 69], but only 14 out of 41 teams employed this strategy. It is also important to note that model ensemble only helps if used for a single model rather than structurally different models. Most of the team also used only age and sex as features rather other using demographic features or 12-lead ECG-based features. The training data in Challenge 2020 suffer from heavy class imbalance, so most teams used threshold optimization [14, 67, 17] and weighted loss [39, 12] to handle the imbalance class issue. In addition, over-sampling [70], down-sampling [27], and other methods have been employed in Challenge 2020.

2.2 Federated learning

FL applied to health has gained interest in the last few years. Specifically, employing FL to classify ECG arrhythmias shows relevant advances and results
[59]. For example, [51] implemented explainable artificial intelligence (XAI) and convolutional neural networks (CNN) for ECG-based healthcare. They employed the baseline Massachusetts Institute of Technology - Boston’s Beth Israel Hospital (MIT-BIH) Arrhythmia database. The trained classifier obtained an accuracy of up to 94.5% and 98.9%.

Utilizing the same MIT-BIH, [29] proposed an algorithm to identify the best number of epochs per training round in FL. They showed that using its algorithm’s suggested number of epochs decreases the training time and resource consumption while keeping the accuracy level. In detail, their approach obtained an accuracy close to 97.5%.

Using ECG collected from different medical institutions, the authors of [66] obtained a dataset where the arrhythmia’s distribution is diverse from one hospital to another. The latter may lead to the non-convergence of the FL-based algorithm. To address that challenge, they optimized the FL-based algorithm using a sharing strategy for partial ECG data of each medical institution combined with the elastic weight consolidation (EWC) algorithm. As a result, their model’s performance measured through the F1-Score got 0.7.

FL can be classified into different types depending on how the data is divided among the devices (data partitioning) [64]. One of those classifications is the so-called Horizontal data partitioning, in which the data of various local nodes share the same properties (i.e., features). However, there is a limited sample space intersection (i.e., the patients differ from device to device). Moreover, the most common FL architecture is horizontal partitioning [42].

FL has relevant advantages over centralized (traditional) systems [11]. The most remarkable is that the data remains locally (in each hospital) and does not require transmitting personal information along the network. It only sends the weights of the models. A second advantage is that models continuously enhance by utilizing local nodes’ input, eliminating the requirement to aggregate data for continual learning. In third place, FL generates technological efficiency. This technique employs less complex hardware since FL models do not require a single central server to train data.

FL, on the other hand, faces several difficulties. Firstly, models could be poisoned by providing model updates resulting from mislabeled data, even though only models, not raw data, are sent to the central server. FL is vulnerable to adversarial assaults in the form of backdoors during training, as seen in [62]. A backdoor aims to tamper with the trained model’s performance on specified sub-tasks. Secondly, device-specific characteristics may prevent models from generalizing from some local nodes, decreasing the global model’s performance [11]. The mentioned difficulties and how to tackle them are out of the scope of this work.

3 Methodology

In this study, we assume a group of healthcare organizations (equipped with low-powered medical devices) that wish to collectively train an arrhythmia classifica-
tion AI-based module without sharing their medical records. For each organization, we envision a single, local module that coordinates all the activities related to collecting, storing, and analyzing medical records. Moreover, we assume that each organization can access high-definition electrocardiogram monitoring devices to record the patient’s heart activity. Each monitoring session produces a short 12-channel ECG recording, i.e., about 16 seconds, that is transferred from the device to the local, private database of the organization. One or more healthcare experts examine the ECG recordings and provide a diagnosis stored in the local database.

We assume that the healthcare organizations have agreed upon a single, global, trusted cloud server. The role of the global server is to coordinate all the activities of the local modules. The global server does not maintain any database with medical records. The only information stored is related to the standard AI model and the system’s operating parameters. We also assume that all organizations have agreed on a standard length for the ECG recordings, e.g., 16 seconds, and a common sampling frequency, e.g., 257KHz.

Fig. 1: High-level overview of the proposed federated learning methodology and software and hardware components.

Remark that several different interconnection architectures are used in ECG technologies available in the market and studied in the relevant literature, such as wireless technologies like WI-FI or BLE, or wire technologies such as USB, or non-volatile memory formats. Such interconnections aspects are beyond the scope of this paper.

Note that in case the recording has a different sampling frequency, various algorithms exist in the relevant literature to change the sampling frequency to a lower or a higher one without affecting the accuracy. [68]

Note that ECG recordings that are longer than the agreed length can be split into multiple ones without loss of generality.
Periodically, the global server starts a **global training session** by notifying all the local servers (a.k.a, clients or local nodes) of the organizations that participate in the federation. Upon receiving this notification, each organization’s device independently goes through a **local training session**. During a local training session, all the records available in the local, private database are analyzed using local computing resources based on a **processing pipeline** made up of four steps. First, each recording passes through a **Feature Extraction** phase, where they are analyzed independently, and critical information connected to the heartbeats that will help the training of the AI model is extracted. Second, a **feature normalization** step follows, where statistics are used to scale the features to improve the robustness of the data. In coordination with the global server, the local servers compute the necessary robust measures over the federated dataset without revealing sensitive information. Third, all the original records, along with the normalized features, are examined. To this end, a **feature selection** is carried out to remove redundant features that may hinder the performance of the trained model and reduce the computational cost since they use a smaller number of features to train the local model. The fourth and final **data balancing** step examines the diagnosis attached to all the records in the local database to identify and remove any imbalances found between the representation of arrhythmia classes, thus increasing the generalization power of the model.

When all the local servers have completed the processing pipelines, under the coordination of the global server, they start training their local models. When the training is concluded, the weights of the resulting model are transmitted to the global server. The global server examines all the individual weights using a **weight aggregation method** and sends the resulting model to all organizations. This process is repeated until either the **distributed optimization converges** or a certain number of steps is reached.

A high-level overview of the above methodology and the interconnection of the software and hardware elements that make up the federated architecture are depicted in Fig. 1.

Regarding the data partitioning, the FL type shown in Fig. 1 is a **Horizontal FL** since each local node collects the same features. Additionally, regarding the ML models, the FL approach exposed is homogeneous because each hospital’s device applies the same model with different data. Finally, regarding scale, the FL defined is known as cross-device because the number of local nodes is high, but they do not have enough computing power [65].

### 3.1 Problem Formulation

We assume that an ECG recording $s_i$ has $c \in [1, 12]$ channels with fixed samplings rate, producing $d$ samples per channel, $s_i = \{x_{i1}^1, x_{i2}^1, \ldots, x_{id}^1, x_{i1}^2, x_{i2}^2, \ldots, x_{id}^2, \ldots, x_{i1}^{12}, x_{i2}^{12}, \ldots, x_{id}^{12}\}$ where $x_{cj}^j$ is the $j^{th}$ sample of channel $c$, where $j \in [1, d]$. The ECG multiclass arrhythmia classification task considered here takes as input a local data set of $n$ 12-channel ECG recordings ($\mathcal{S} = \{s_1, \ldots, s_n\}$). It outputs a sequence of labels, one for each signal $\mathcal{L} = \{l_1, \ldots, l_n\}$, where each $l_i \in \mathcal{A}$, and $\mathcal{A}$ represents the different rhythm classes considered. We focus on $|\mathcal{A}| = 27$ diagnoses of arrhythmia...
classes as defined by the AAMI standard [38] that are of clinical interest and more likely to be recognizable from ECG recordings [3]. Table 1 summarizes the complete list of diagnoses considered.

| Diagnosis                        | Abbreviation |
|----------------------------------|--------------|
| 1st degree AV block              | IAVB         |
| Abnormal QRS                     | abQRS        |
| Atrial fibrillation              | AF           |
| Left atrial enlargement          | LAE          |
| Left axis deviation              | LAD          |
| Left bundle branch block         | LBBB         |
| Left ventricular hypertrophy     | LVH          |
| Low QRS voltages                 | LQRSV        |
| Myocardial infarction            | MI           |
| Myocardial ischemia              | MIs          |
| Nonspecific ST abnormality       | NSSSTTA      |
| Old myocardial infarction        | OldMI        |
| Pacing rhythm                    | PR           |
| Premature atrial contraction     | PAC          |
| Prolonged QT interval            | LQT          |
| Q wave abnormal                  | QAb          |
| Right bundle branch block        | RBBB         |
| Sinus arrhythmia                 | SA           |
| Sinus bradycardia                | SB           |
| Sinus rhythm                     | NSR          |
| Sinus tachycardia                | STach        |
| ST depression                    | STD          |
| ST elevation                     | STE          |
| ST interval abnormal             | STIAb        |
| T wave abnormal                  | TAb          |
| T wave inversion                 | Tinv         |
| Ventricular ectopics             | VEB          |

Table 1: Diagnoses and abbreviations for the selected diagnoses

3.2 Feature Extraction

First, each 12-channel ECG recording is analyzed by examining the morphology of all the heart cycles included. The QRS complex represents a heart cycle, illustrated in Fig. 2, which corresponds to the depolarization of the right and left ventricles of the human heart. Detecting the QRS complex is essential for time-domain signal analysis, that is, the heart rate variability. For each found QRS complex, a series of representative morphological features\(^5\) is extracted in terms of slope, amplitude, and width, such as the R peak, the highest amplitude

\(^5\) Code for morphological features: https://github.com/physionetchallenges/python-classifier-2020/blob/master/get_12ECG_features.py
of the complex: \textit{PR interval}, the period between the P peak and the R peak; \textit{QRS interval}, the period between the Q and the S segments; \textit{ST interval}, the period between the S and T segments; \textit{QT interval}, the period between the Q and T segments; the \textit{RR interval}, the period between two consecutive R peaks.

![Typical ECG waveform](image)

Fig. 2: Typical ECG waveform from the Lead II position for one cardiac cycle.

An alternative type of feature based on \textit{Spectral analysis} was employed in this work. The latter is a frequency-based signal analysis that uses a Discrete Wavelet Transform (DWT) followed by spectral analysis\(^6\). For each of the coefficients of the DWT, as well as the original complete signal, the percentiles (5, 25, 50, 75, 95), mean, standard deviation, variance, skewness, and kurtosis, are calculated along with Shannon’s Entropy (information of the distribution). The DWT coefficients were obtained using the \texttt{waveform} function of the \texttt{pywt} Python package. The age and sex of patients are also included as features, resulting in a total of 650 features (14 morphological and demographical and 636 spectral). Note that this step is done only once for each record, and the extracted features are stored on the local database for future training sessions.

### 3.3 Feature Normalization

Towards enhancing the performance of the AI models, the values of each feature are centered and scaled independently according to the median value and the interquartile range (IQR) between the first and third quartiles, i.e., the 25th and 75th percentiles, using the \texttt{RobustScaler} Python function. The specific formula is as follows:

\[
\text{norm} (f_{x_i}^e) = \frac{f_{x_i}^e - \bar{f}_{x_i}^e}{\bar{f}_{q_{75}} - \bar{f}_{q_{25}}}
\]

\(^6\) Code for spectral features: \url{https://github.com/onlyzdd/ecg-diagnosis/blob/dfa9033d5ae7be135a563f567e666db2b86d76d/expert_features.py}
Where $f$ is one of the 636 spectral features, $\bar{x}^c$ is the median, $f_{q_{25}}$ and $f_{q_{75}}$ are the 25th and 75th percentiles computed over all the values in the entire federated data set for the given feature $f$ in channel $c$.

### 3.4 Feature Selection

Since the extraction of the majority of the features is conducted over each of the 12 channels and due to the nature of multi-channel ECG recording, in many cases, there is redundancy between pairs of features. Moreover, environmental noise from nearby appliances or noise originating from muscular activity frequently produces abnormalities in a recording identified and removed by the ECG monitor at the hardware level [57]. The latter leads to temporal loss of some values of one or more of the 12 channels [35]. Since redundant or missing features can potentially lower the performance of the classification model, it is crucial to identify and remove them. Besides improving our model’s predictive performance, it will also reduce the overall computational requirements of AI model training [16].

Several methods have been proposed for feature selection in the relevant bibliography [20]. Here we offer a simple and automated way to rank features based on their importance. We calculate the embedded feature importance scores based on tree ensemble models. The central concept here is that given the feature sub-sampling and bootstrapping in an ensemble of trees, the importance scores of two or more redundant features are expected to be spread evenly among the trees. Then the resulting scores of the candidate features are re-examined following a sequential boosting process based on the XG-Boost model [15]. The boosting process allows re-examining some of the features that received a low rank due to specific limitations of the tree-based feature importance calculation in large datasets, datasets with missing values, and datasets with both numerical and categorical features [5].

### 3.5 Dataset Balancing

Various nationwide register-based cohort studies conducted over a very long period, such as in Denmark [63] and Canada [54], indicate an inequality of arrhythmia incidence rate between highest and lowest income groups and the existence of significant geographical variation even in countries with free access to healthcare and even when accounting for socioeconomic differences at an individual level. Therefore when the organizations do not share their medical records, the observed distribution of the local diagnoses of the arrhythmia classes will differ across each healthcare organization. As each organization is training its local model based only on the data locally available, it is expected that the resulting model will have certain biases towards the most populated classes, learning too much about their characteristics and failing to recognize the differentiating factors of the others, resulting in poor performance for the minority classes [23].

Various techniques have been proposed to overcome this problem in the relevant literature, such as the Random Over Sampling (ROS) [34], Random
Under Sampling (RUS) [23], and Synthetic Minority Oversampling Technique (SMOTE-SMT) [13]. The approaches implemented were ROS + RUS and SMT + RUS. The latter aimed to overcome the well-known problem of ROS resulting from the generation of exact copies of the minority class samples.

In more detail, for a local dataset made up of \( n \) recordings and diagnoses represented as \( D = \{ (s_1, l_1), \ldots , (s_n, l_n) \} \). Then \( b_u \) is the number of signals belonging to the under-balanced classes, and \( b_o \) is the number of over-balanced classes. In this case, \( b_u \leq b_o \). Clearly \( b_u + b_o = n \). Our balancing method will make a total of \( \tau = (m_l - m_s) \cdot \beta \) random duplications (or interpolations) and eliminations, where \( \beta \in [0,1] \), and \( m_l \) and \( m_s \) represent the numbers of recordings in the largest class and the smallest class, respectively. Here \( \beta \) is user-defined. For example, if \( \beta = 1 \), the resulting dataset is completely balanced.

The proposed algorithm uses an appropriate mix of random duplication (or interpolation) of the minority, under-balanced class samples, and random elimination of the majority, over-sampled classes. Thus, after identifying the changes that need to be made, the algorithm randomly chooses a sample \( x_i \). If it belongs to an under-balanced class, it will duplicate it, while if it belongs to an over-balanced class, it will eliminate it. The resulting approach is simple to implement and has shallow requirements in terms of computational resources.

### 3.6 Multiclass Arrhythmia Classification

**Model 1 - (DNN) Deep Neural Networks** In this context, deep neural networks (DNN) refer to organic or artificial systems of neurons. DNN can adapt to changing input and produce the best possible result without modifying the output criteria because they can adjust to variable inputs [18]. DNN, an artificial intelligence-based concept, has gained popularity for developing health classification systems.

![DNN architecture](image)

Fig. 3: DNN architecture
As depicted in Fig. 3, the DNN defined, inspired by [45], has one input layer, three hidden layers, and one output layer. For the input layer, 120 units were considered since they represent the number of features used in the training set. The three hidden layers contain 500 hidden units each. Furthermore, the last layer was formed by 27 neurons, related to the number of diagnoses to predict. Moreover, the hidden layers employed the ReLu activation and the SoftMax for the output layer. The hidden layers, neurons, and activation functions arose after the hyperparameter fine-tuning method, using the possibilities in Table 2. In detail, one model was trained for every combination of parameters. Then, Accuracy, Precision, Recall, and F1-Score were calculated over the test set and the fitted model. Finally, the best approach ended up as the model obtaining the highest F1-Score among all the combinations.

| Hyperparameter                          | Possible values | Best approach |
|-----------------------------------------|-----------------|---------------|
| Number of hidden layers                 | 2, 3, 5, 10     | 3             |
| Number of hidden neurons                | 50, 100, 500, 1000 | 500          |
| Activation function (Hidden layers)     | ReLu, Tanh, SeLu | ReLu          |
| Activation function (Output layer)      | Sigmoid, SoftMax | SoftMax       |

Table 2: Fine-tuning grid for DNN

**Model 2 - (LSTM) Long-Short Term Memory** Long-short-term memory (LSTM) networks are a type of DNN. It is a Recurrent Neural Networks (RNNs) class that can learn long-term dependencies and functionals to solve sequence prediction issues. Besides, for single data points like pictures, LSTM has feedback connections, which means it can process a complete data sequence [25].

Fig. 4 depicts the architecture used for the LSTM methodology, motivated by [45]. Then, 120 neurons were employed in the input layer because those variables were used to predict. Moreover, both the LSTM cell and the output layer were composed of 27 neurons, as many as the number of classes predicted. Regarding the hyperparameters mentioned, their values were selected after the fine-tuning process, providing the best performance. Table 3 depicts the possibilities tested during the hyperparameter fine-tuning procedure. Notice that the number of LSTM cell neurons has an upper limit of 27, corresponding to the number of classes (arrhythmias) considered in this work.

### 3.7 Distributed Weight Aggregation

When the weights from each local model arrive at the global node, they are aggregated. Federated Average Aggregation (FedAvg) is typically used for that task [32]. The latter enables local nodes to perform multiple batch updates on local data, exchanging averaged updated weights rather than raw data [37].
In an FL environment exists a concept so-called \textbf{communication round} (comm round) \cite{8}. The latter begins when a model is trained inside each local node, and later the models’ weights are passed to the global node to be aggregated there. Finally, the communication round finishes when each local model gets updated with the new weights. Then, as a result, in each comm round, the model’s performance increases. Moreover, it should get stable after some trials \cite{9}.

The training process in the proposed architecture works as follows. First, each local node trains the model using the correspondent local data. In the second place, every client sends the weights of the obtained model to the global server. The weights are aggregated and sent back to each client in the latter. Finally, each local node updates its parameters. The previous process is repeated until the performance of the global model converges.

\section{Experimental settings and fine-tuning}

\subsection{Data}

We utilized the 2020 PhysioNet Challenge dataset \cite{3} to classify cardiac arrhythmias from ECG records, consisting of 12-lead ECG recordings from 43,059 patients. Furthermore, it adds data from different hospitals, making it suitable for an FL approach to address privacy concerns. It offers heterogeneous data
and a wide range of 82 cardiac arrhythmias, representing a realistic scenario. The data are from five sources summarized in Table 4.

| Dataset                        | Number of Recordings | Mean Duration (seconds) | Mean Age (years) | Sex (male / female) | Sample Frequency (Hz) |
|--------------------------------|----------------------|-------------------------|------------------|---------------------|-----------------------|
| CPSC (all data)                | 13256                | 16.2                    | 61.1             | 53%/47%             | 500                   |
| CPSC Training                  | 6877                 | 15.9                    | 60.2             | 54%/46%             | 500                   |
| CPSC-Extra Train               | 3453                 | 15.9                    | 63.7             | 53%/46%             | 500                   |
| Hidden CPSC                    | 2926                 | 17.4                    | 60.4             | 52%/48%             | 500                   |
| INCART                         | 72                   | 1800.0                  | 56.0             | 54%/46%             | 257                   |
| PTB                            | 516                  | 110.8                   | 56.3             | 73%/27%             | 1000                  |
| PTB-XL                         | 21837                | 10.0                    | 59.8             | 52%/48%             | 500                   |
| G12EC (all data)               | 20678                | 10.0                    | 60.5             | 54%/46%             | 500                   |
| G12EC Training                 | 10344                | 10.0                    | 60.5             | 54%/46%             | 500                   |
| Hidden G12EC                   | 10344                | 10.0                    | 60.5             | 54%/46%             | 500                   |
| Undisclosed                     | 10000                | 10.0                    | 63.0             | 53%/47%             | 300                   |

Table 4: Relevant information for each data set

Fig. 5: Feature importance from XG-Boost algorithm (only the 50 best)

**Processing Pipeline fine-tuning** The bar plot in Fig. 5 shows the most important features for predicting classes using the XG-Boost method. After obtaining importance scores for all 650 features, trial-and-error was conducted to remove unimportant features, ending with 120 features keeping the Accuracy and F1-Score. The best features included entropy for leads 9, 11, and 10, the lead six’s 5th percentile, and the median for leads 1 and 2.

\[
\text{feature}_s = l(lead_i), c_\text{coefficient}_j, \text{operation}_k
\]  

(2)

Each 636 spectral feature is based on lead denoted as \( l \), their coefficient as \( c \), and operation applied on it. Also, lead, represents ECG lead number from 00-11,
coefficient$_j$ represents a coefficient number from DWT (there are five coefficients in total 1-5), and operation$_k$ represents the operation name like average, standard deviation as std, variance as var, percentiles represents as n5(percentile 5), n25(percentile 25), n50(percentile 50), n75(percentile 75), n95(percentile 95) and entropy applied on coefficient got represented as c1 − c5. In Fig. 5, names of features are represented by the 2 mnemotechnic structure.

Fig. 6: (Left) Selected diagnoses distribution. (Right) Diagnoses distribution for ROS and SMOTE datasets.

The ROS and SMT approaches were applied to the dataset with a down-sampling to have as many recordings as the filtered dataset, with 43,200 recordings for each method. Thus, as depicted in Fig. 6, the distribution of the labels is much more similar among the arrhythmia categories.

4.2 Scenarios

We present an FL approach for classifying 12-channel ECG arrhythmias. Two FL scenarios, IID and Non-IID, were developed and tested (see Fig. 7). The IID combined six databases into one dataset, resulting in 41,894 (a.k.a. filtered dataset) recordings after selecting the most representative diagnoses. The dataset was randomly split into train (90%), validation (5%), and test (5%) sets. Stratified random splitting was used for the label distributions to be the same for each partition. Different partition sizes, such as 2, 4, 6, 8, and 10, are possible depending on the 12-leads ECG data, which will be explored in section 5.4.

The Non-IID approach has the same two steps of the IID (data appending and splitting of Fig. 7). The main difference is the generation of the local nodes. Thus, to get those clients, an Unstratified random sampling (with replacement) method [28] was employed. As a result, each local node has a varied distribution
Fig. 7: IDD and Non-IID architecture

of diagnoses. Furthermore, some recordings may be missing entirely from the synthesized dataset.

Fig. 8: Label distribution for the filtered dataset by each local node

As depicted in Fig. 8, the distribution among the four local nodes seems IID. The latter means that the diagnoses along the devices are equally distributed, containing 9,426 recordings per local node. The same occurs for the ROS and SMOTE datasets, as is shown in Fig. 9.

For the Non-IID approach, the method drew four random samples with repetition from the 41,894 recordings (local nodes) dataset. The given data in each local node (or client) is Non-IID approaches due to the unstratified random sampling, as shown in Fig. 10. This scenario is far more plausible since most 12-channel ECGs shared across several devices are Non-IID.
4.3 Performance metrics

For measuring the goodness-of-fit for each method proposed, five metrics were employed. Those metrics are related to the concepts extracted from a confusion matrix [26]. The following are explanations about each term, using as an example a fictitious diagnosis A:

- **True negatives (TN):** The model predicted a patient wouldn’t have diagnosis A, and he did not have it.
- **True positives (TP):** These were examples when the model predicted yes (the recording has the diagnosis A), and the patient does have it.
- **False positives (FP):** The model predicted that a patient would have the diagnosis A, but he does not. (also referred to as a "Type I error.")
- **False negatives (FN):** The model predicted that a patient would not have diagnosed A, yet he does. (This is often referred to as a "Type II error.")

Then, the metrics to measure the classification power of the algorithms proposed are:

**Accuracy:** \( \text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \): A metric measuring correct predictions out of all observations. Useful for balanced data where false positives and false negatives are similar.

**Precision:** \( \text{Precision} = \frac{TP}{TP + FP} \): The ratio of accurately predicted positives to total expected positives. It reflects the accuracy of identifying patients with a specific diagnosis and relates to a low false-positive rate.
Recall: \( \text{Recall} = \frac{TP}{TP + FN} \): The ratio of successfully predicted positives to all actual class observations. It measures the identification of patients with arrhythmia A.

F1-Score: \( \text{F1-Score} = \frac{2 \times (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}} \): A weighted average of Precision and Recall, considering false positives and false negatives. It is more valuable than accuracy, particularly when dealing with unequal class distribution or significant differences in the cost of false positives and negatives.

It is relevant to highlight that the current work deals with a multi-class problem (more than two classes as labels). Due to it, for Precision, Recall, and F1-Score, the final metrics values were calculated as the weighted average of the metric measured in each pairwise class [19].

In FL is relevant to measure the **Execution time** that a model takes to be trained [29]. The latter helps compare the FL architecture to the centralized one, expecting the FL approach to be faster in each local node. This time accounts for the time elapsed since the model started to train until it finished and returned the results. During the experimentation of this work, the execution times were measured in **minutes**. The hardware specification used to train and evaluate the models is shown in Table 5.

| Component          | Specification                                      |
|--------------------|----------------------------------------------------|
| Disk size          | 108 GB                                             |
| Processors’ model  | Intel(R) Xeon(R) CPU @ 2.20GHz                     |
| Number of processors | 2                                                    |
| Memory             | 13.2 GB                                            |
| Operating System   | (Linux) Ubuntu 18.04.5 LTS                         |
| GPU                | GeForce RTX 3070 8GB                               |
| Python version     | 3.7.13                                             |

Table 5: Hardware specification used to train models

5 Results

5.1 Performance of Centralized

A Centralized (traditional) Learning approach was employed to evaluate the dataset’s performance without dividing it among hospitals and devices. Feature engineering, data balancing, and classification methods (DNN, LSTM) were applied to the complete dataset. The model of the second-best team (TEAM2) [68] in the Physionet competition was replicated for fair comparisons\(^7\). Imbalanced datasets for each arrhythmia type were addressed using ROS and SMT techniques. ROS was also applied to the TEAM2 strategy but did not yield better results.

\(^7\) TEAM2’s code: [https://github.com/ZhaoZhibin/Physionet2020model](https://github.com/ZhaoZhibin/Physionet2020model)
results. The SMT strategy was not tested within the TEAM2 approach. Table 6 depicts the tested and best approaches we got regarding the classification metrics.

| Characteristic          | Scenarios                                                                 | Best approach               |
|-------------------------|---------------------------------------------------------------------------|-----------------------------|
| Data Split              | % Train-% Validation-% Test:                                               | Option 4: 90%-5%-5%          |
|                         | 60%-20%-20% Option 2: 70%-10%-10% Option 3: 80%-10%-10% Option 4: 90%-5%-5% |                             |
| Features normalization  | Option 1: MinMaxScaler Option 2: StandardScaler Option 3: RobustScaler     |                             |
| Sampling rate           | Option 1: 257Hz Option 2: 500Hz                                           | Option 1: 257Hz             |
| Features employed       | Option 1: Baseline features (morphological and demographic) Option 2: Baseline features (morphological and demographic) and Spectral features | Option 2: Baseline features and Spectral features |

Table 6: Scenarios tested for centralized learning

Fig. 11: (Left) F1-Score for methods employed in CL on the test set. (Right) Execution times for the methods used in CL.

Within the results in Fig. 11, the best model was TEAM2’s approach, which got an F1-Score close to 0.63. Nevertheless, the DNN over the ROS data had similar behavior, with the mentioned metric at 0.61. Finally, LSTM does not perform well compared to the other models, with metrics between 0.39 and 0.47 for all the scenarios.

Finally, as shown in Fig. 11, TEAM2’s approach took almost 122 minutes to run. On the other hand, DNN and LSTM took lower execution times (close to 89 minutes on average). Thus, TEAM2 is the slowest method, although it generates the best results. On the contrary, DNN is a fast method, and the performance is NOT quite different from TEAM2.
5.2 Performance of IID Federated

Then, the modeling part was executed in the IID FL paradigm using four clients. Therefore, DNN and LSTM methodologies classified the 12-leads ECG’s arrhythmias. Hence, the aggregation technique employed to get the weights of the global model was FedAvg [31].

As depicted in Fig. 12, the best results arise using the TEAM2 method with an F1-Score of 0.58. On the other hand, DNN got an F1-Score of 0.54, placing it as the second-best option. In addition, the best performance for LSTM was obtained with the original data, although it is worst than the TEAM2 and DNN ROS models. Then, applying oversampling for LSTM does not improve the result of the models. However, using ROS, the performance of the FL model increases.

Considering Fig. 12 results, the TEAM2 generates the slowest procedure with a time of 78 minutes. In addition, the second less time-consuming approach was the DNN ROS with 32 minutes, where using oversampling techniques makes the execution time increase [30]. Compared to the CL approach, the TEAM2 FL method has a faster execution with similar performance. Similarly, DNN ROS and SMOTE ran fast, performing comparably to their CL versions.

Fig. 13: Metrics along communication rounds for TEAM2 (the best model) on the test set
Fig. 13 establishes the behavior of each metric for each communication round. As depicted, the performance stabilizes after the eighth comm round. Notice that all the metrics start low in the first comm round, but the measurements get steady after some updates. DNN and LSTM methods performed similarly, but for the sake of the extension, it is not included in this work.

| Method                  | Accuracy | Precision | Recall | F1-Score | Time | Accuracy | Precision | Recall | F1-Score | Time |
|-------------------------|----------|-----------|--------|----------|------|----------|-----------|--------|----------|------|
| Competence Team #2 [68] | 0.64     | 0.64      | 0.64   | 0.63     | 122  | 0.63     | 0.64      | 0.63   | 0.58     | 78   |
| Inspirational DNN [45]  | 0.50     | 0.46      | 0.50   | 0.47     | 88   | -        | -         | -      | -        | -    |
| Inspirational LSTM [45] | 0.50     | 0.45      | 0.50   | 0.46     | 89   | -        | -         | -      | -        | -    |
| DNN                     | 0.50     | 0.47      | 0.50   | 0.47     | 89   | 0.46     | 0.53      | 0.45   | 0.49     | 22   |
| DNN ROS                 | 0.61     | 0.66      | 0.61   | 0.61     | 90   | 0.55     | 0.59      | 0.54   | 0.54     | 32   |
| DNN SMOTE               | 0.60     | 0.64      | 0.60   | 0.60     | 91   | 0.52     | 0.58      | 0.52   | 0.51     | 31   |
| LSTM                    | 0.51     | 0.44      | 0.51   | 0.39     | 91   | 0.48     | 0.58      | 0.48   | 0.52     | 22   |
| LSTM ROS                | 0.38     | 0.51      | 0.38   | 0.39     | 91   | 0.39     | 0.48      | 0.38   | 0.38     | 25   |
| LSTM SMOTE              | 0.39     | 0.49      | 0.39   | 0.39     | 89   | 0.39     | 0.47      | 0.39   | 0.38     | 25   |

Table 7: Metrics for CL and IID Federated Learning (FL -IID) in the test data set. Execution time for preprocessing and training in minutes.

Table 7 depicts the metrics obtained over the test dataset and the execution time of the training phase and all the state-of-the-art competitors. The CL TEAM2 model outperforms all the proposals by at least two percentage points, but it is the slowest method. Concerning the FL architecture, the TEAM2 FL is close enough to the centralized TEAM2 model, showing that the FL applied over an IID set behaves well, although the execution time is better for DNN ROS.

5.3 Performance of Non-IID Federated

The modeling phase was done again with the four Non-IID datasets extracted. The 12-channel ECG arrhythmias were then classified using the TEAM2, DNN, and LSTM techniques following an FL paradigm.

We employed a CL strategy to compare the Non-IID FL and CL implementations. In terms of classification performance, the best results were for TEAM2 with the original datasets, as shown in Fig. 14, while the DNN ROS and DNN LSTM techniques also performed well, with an F1-Score of around 0.58 on the test set. On the other hand, the highest performance for LSTM was with the original data, yet it was worse than the DNN ROS model.

As shown in Fig. 14, the TEAM2 method was the slowest, clocking in at 74 minutes. Compared to the CL method, the TEAM2 methodology with FL is faster and produces equivalent results. The same behavior applies to DNN ROS and SMOTE; however, while LSTM performs worse than TEAM2 and DNN, it is significantly quicker in producing the results.
5.4 Performance changing local nodes

We implemented a model’s performance evaluation by changing the number of local nodes diverse to 4 (i.e., 2, 4, 6, 8, 10). Per each client, we trained a CL to determine how well the FL training fits regarding that CL for IID and Non-IID scenarios. Still, for the sake of the extension of this document, only the Non-IID result is reported because the performances of both methods were quite similar.

Fig. 15 represents the F1-Score for all the methods while changing the number of local nodes. TEAM2 CL, DNN CL, and LSTM CL are reference models trained with all the data appended to a single dataset, where the best solution when considering two local nodes is DNN ROS. On the other hand, with four or more clients best solution is TEAM2, keeping steady behavior when increasing the number of clients. Nevertheless, when considering six or more nodes, LSTM on the original data is a good option, similar to the TEAM2 metric.

Typically, the higher the local nodes’ quantity, the faster the algorithms run, as depicted in Fig. 15. Moreover, all the techniques drastically decreased the running time when increasing the number of clients, but the performance also decreased.
6 Discussion

Analyzing the results obtained from the experiments in the CL setting, the highest classification performance was reached by TEAM2, which combined Convolutional Recurrent Neural Networks (CRNN) and cross-validation. Nevertheless, our proposed DNN on the augmented ROS dataset had a competitive classification performance. In addition, the execution time for DNN was more than twice faster than TEAM2. Why can a DNN model perform similarly to a CRNN while reducing the execution time? This is possible since we employed a solid pre-processing pipeline that extracted and selected powerful spectral and morphological features to classify arrhythmia accurately. Moreover, the architecture of DNN was considerably less complex (fewer hidden layers, fewer hidden neurons, not using cross-validation) than TEAM2’s approach.

Concerning the FL IID experiments results, TEAM2 and DNN using the ROS balanced data have similar F1-Score to their CL versions while having a decrease of around 36% and 64% in the execution time, respectively. Why does employing FL cause such execution time reduction? That is an expected behavior supporting scientific evidence [1] since running models in an FL fashion can be done parallelly in each client, removing the burden of transmitting large amounts of data, and causing the execution time to drop substantially.

Regarding the FL non-IID experiment results, TEAM2 and DNN using ROS again reached a classification score similar to their CL approaches, causing an increase of 85% and 65% in the execution time, respectively. How come that FL IID causes a decrease in the execution time, but FL non-IID increases it? The answer to this question involves the aggregation method employed (FedAvg) and the diversity of the data among the clients. Given that the data is non-IID by construction, it causes the time taken for the models to converge to be more significant since, in each communication round, the weights to aggregate are much more different from one another, which is something that does not occur in the IID (and not quite real) setting.

In the context of FL Non-IID, using less than four clients, the most accurate models were DNN and TEAM2. Even so, when considering six or more clients, LSTM exhibited a comparable performance while drastically decreasing the execution time. Why does the number of clients involved in the FL experiment cause LSTM to outperform DNN and become more similar to TEAM2’s approach? The FL approach under DNN may overfit when training with fewer data in each client since the model becomes too specialized for a smaller dataset. At the same time, the LSTM algorithm has been studied as a more robust solution to overfitting (mitigating the vanishing gradient problem) [47]. In this case, the model is not as complex as the DNN, having a better performance while increasing the number of clients (decreasing the number of samples per client). CRNN may also overfit the given data, but that effect is mitigated by using cross-validation in TEAM2’s solution, causing a comparable performance between TEAM2 and LSTM.
7 Conclusions and future work

Our team implemented a 12-lead arrhythmia classification leveraged on an FL paradigm over multiple ML and AI algorithms to evaluate their performance compared to a centralized (traditional) approach. Although the solution from TEAM2 provided good results, our proposal, using DNN over augmented datasets (ROS), produced a comparable performance with the benefit of a drastic reduction in the execution time. In addition, considering a Non-IID scenario over a few clients, TEAM2 and DNN ROS attained equivalent behaviors to their respective CL and FL approaches. At the same time, our DNN ROS solution acutely reduced the execution time. Alternatively, using our proposed LSTM for larger clients demonstrated steady and analogous behavior to the obtained in a CL with a drastic reduction of the execution time compared to the state-of-the-art.

In future work, the FL architecture could use a different data partition. For example, each original database can be used as a client with a Non-IID distribution of arrhythmias. Moreover, alternative aggregation methods may deal with the Non-IID property, such as Fedprox, SCAFFOLD, or FedNova [31]. Another venue for future research is using Catboost and XG-Boost in an FL architecture to check their performance for the 12-leads ECG arrhythmia classification.

Acknowledgements

This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU and PNRR351 TECHNOPOLE - NEXT GEN EU Roma Technopole - Digital Transition, FP2 - Energy transition and digital transition in urban regeneration and construction.
Bibliography

[1] AbdulRahman, S., Tout, H., Ould-Slimane, H., Mourad, A., Talhi, C., Guizani, M.: A survey on federated learning: The journey from centralized to distributed on-site learning and beyond. IEEE Internet of Things Journal 8(7), 5476–5497 (2020)

[2] Al-Zaiti S, Besomi L, B.Z.: Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. National Library of Medicine (7 2020). https://doi.org/10.1038/s41467-020-17804-2

[3] Alday, E.A.P., Gu, A., Shah, A.J., Robichaux, C., Wong, A.K.I., Liu, C., Liu, F., Rad, A.B., Elola, A., Seyedi, S.: Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. Physiological measurement 41(12), 124003 (2020)

[4] Alday, E.A.P., Gu, A., Shah, A.J., Robichaux, C., Wong, A.K.I., Liu, C., Liu, F., Rad, A.B., Elola, A., Seyedi, S., Li, Q., Sharma, A., Clifford, G.D., Reyna, M.A.: Classification of 12-lead ECGs: the PhysioNet/computing in cardiology challenge 2020. Physiological Measurement 41(12), 124003 (dec 2020). https://doi.org/10.1088/1361-6579/abc960, https://doi.org/10.1088/1361-6579/abc960

[5] Alsahaf, A., Petkov, N., Shenoy, V., Azzopardi, G.: A framework for feature selection through boosting. Expert Systems with Applications 187, 115895 (2022). https://doi.org/https://doi.org/10.1016/j.eswa.2021.115895, https://www.sciencedirect.com/science/article/pii/S0957417421012513

[6] Ana Minchole, Julia Camps, A.L.: Machine learning in the electrocardiogram. National Library of Medicine p. S61–S64 (2019). https://doi.org/10.1016/j.jelectrocard.2019.08.008

[7] Arnold, D., Wilson, T.: What doctor? why ai and robotics will define new health. In: PwC (2017)

[8] Asad, M., Moustafa, A., Ito, T.: Fedopt: Towards communication efficiency and privacy preservation in federated learning. Applied Sciences 10, 1–17 (04 2020). https://doi.org/10.3390/app10082864

[9] Asad, M., Moustafa, A., Ito, T., Aslam, M.: Evaluating the communication efficiency in federated learning algorithms (2020). https://doi.org/10.48550/ARXIV.2004.02738, https://arxiv.org/abs/2004.02738

[10] Attia, Z.I., Noseworthy, P.A., Lopez-Jimenez, F., Asirvatham, S.J., Deshmukh, A.J., Gersh, B.J., Carter, R.E., Yao, X., Rabinstein, A.A., Erickson, B.J., Kapa, S., Friedman, P.A.: An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet 394(10201), 861–867 (2019). https://doi.org/https://doi.org/10.1016/S0140-6736(19)31721-0, https://www.sciencedirect.com/science/article/pii/S0140673619317210
[11] Bogdanova, A., Attoh-Okine, N., Sakurai, T.: Risk and advantages of federated learning for health care data collaboration. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering 6, 04020031 (09 2020). https://doi.org/10.1061/AJRUA6.0001078

[12] Bos, M.N., Van De Leur, R.R., Vranken, J.F., Gupta, D.K., Van Der Harst, P., Doevendans, P.A., Van Es, R.: Automated comprehensive interpretation of 12-lead electrocardiograms using pre-trained exponentially dilated causal convolutional neural networks. In: 2020 Computing in Cardiology. pp. 1–4 (2020). https://doi.org/10.22489/CinC.2020.253

[13] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357 (2002)

[14] Chen, J., Chen, T., Xiao, B., Bi, X., Wang, Y., Duan, H., Li, W., Zhang, J., Ma, X.: Se-ecgnet: Multi-scale se-net for multi-lead ecg data. In: 2020 Computing in Cardiology. pp. 1–4 (2020). https://doi.org/10.22489/CinC.2020.085

[15] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 785–794. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785, https://doi.org/10.1145/2939672.2939785

[16] Dhal, P., Azad, C.: A comprehensive survey on feature selection in the various fields of machine learning. Applied Intelligence pp. 1–39 (2022)

[17] Fayyazifar, N., Ahderom, S., Suter, D., Maiorana, A., Dwivedi, G.: Impact of neural architecture design on cardiac abnormality classification using 12-lead ecg signals. In: 2020 Computing in Cardiology. pp. 1–4 (2020). https://doi.org/10.22489/CinC.2020.161

[18] Gallo, C.: Artificial Neural Networks: tutorial, chap. Artificial Neural Networks: tutorial, p. 10 (01 2015)

[19] Grandini, M., Bagli, E., Visani, G.: Metrics for multi-class classification: an overview (2020). https://doi.org/10.48550/ARXIV.2008.05756, https://arxiv.org/abs/2008.05756

[20] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research 3(Mar), 1157–1182 (2003)

[21] Hamet, P., Tremblay, J.: Artificial intelligence in medicine. Metabolism 69, S36–S40 (2017)

[22] Hannun AY, Rajpurkar P, H.M.: Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine p. 65–69 (7 2019). https://doi.org/10.1038/s41591-018-0268-3

[23] He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on knowledge and data engineering 21(9), 1263–1284 (2009)

[24] He, J., Baxter, S.L., Xu, J., Xu, J., Zhou, X., Zhang, K.: The practical implementation of artificial intelligence technologies in medicine. Nature medicine 25(1), 30–36 (2019)
Application of FL techniques for 12-lead ECG arrhythmia classification

[25] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9, 1735–80 (12 1997). https://doi.org/10.1162/neco.1997.9.8.1735

[26] Hossin, M., M.N, S.: A review on evaluation metrics for data classification evaluations. International Journal of Data Mining and Knowledge Management Process 5, 01–11 (03 2015). https://doi.org/10.5121/ijdkp.2015.5201

[27] Hsu, P.Y., Hsu, P.H., Lee, T.H., Liu, H.L.: Multi-label arrhythmia classification from 12-lead electrocardiograms. In: 2020 Computing in Cardiology. pp. 1–4 (2020). https://doi.org/10.22489/CinC.2020.134

[28] Hsu, T.M.H., Qi, H., Brown, M.: Measuring the effects of non-identical data distribution for federated visual classification (2019). https://doi.org/10.48550/ARXIV.1909.06335, https://arxiv.org/abs/1909.06335

[29] Ibrahim, L., Selimi, M., Freitag, F.: Bepoch: Improving federated learning performance in resource-constrained computing devices. In: 2021 IEEE Global Communications Conference (GLOBECOM). pp. 1–6 (2021). https://doi.org/10.1109/GLOBECOM46510.2021.9685095

[30] Jamali-Rad, H., Abdizadeh, M., Singh, A.: Federated learning with taskonomy for non-iid data. IEEE Transactions on Neural Networks and Learning Systems (2022)

[31] Li, Q., Diao, Y., Chen, Q., He, B.: Federated learning on non-iid data silos: An experimental study. arXiv preprint arXiv:2102.02079 (2021)

[32] Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z.: On the convergence of fedavg on non-iid data (2019). https://doi.org/10.48550/ARXIV.1907.02189, https://arxiv.org/abs/1907.02189

[33] Lin, C.C., Yang, C.M.: Heartbeat classification using normalized rr intervals and wavelet features. In: 2014 international symposium on computer, consumer and control. pp. 650–653. IEEE (2014)

[34] Ling, C.X., Li, C.: Data mining for direct marketing: Problems and solutions. In: Kdd. vol. 98, pp. 73–79 (1998)

[35] Mariappan, P.M., Raghavan, D.R., Aleem, S.H.A., Zobaa, A.F.: Effects of electromagnetic interference on the functional usage of medical equipment by 2g/3g/4g cellular phones: A review. Journal of Advanced Research 7(5), 727–738 (2016)

[36] McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al.: International evaluation of an ai system for breast cancer screening. Nature 577(7788), 89–94 (2020)

[37] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)

[38] for the Advancement of Medical Instrumentation, A.: Testing and reporting performance results of cardiac rhythm and st segment measurement algorithms (2013), american National Standard 2013, ANSI/AAMI EC57:2012

[39] Min, S., Choi, H.S., Han, H., Seo, M., Kim, J.K., Park, J., Jung, S., Oh, I.Y., Lee, B., Yoon, S.: Bag of tricks for electrocardiogram classification with deep neural networks. In: 2020 Computing in Cardiology. pp. 1–4 (2020). https://doi.org/10.22489/CinC.2020.328
[40] Moody, G., Mark, R., Goldberger, A.: Physionet: a web-based resource for the study of physiologic signals. IEEE Engineering in Medicine and Biology Magazine 20(3), 70–75 (2001). https://doi.org/10.1109/51.932728

[41] Moody, G.B., Mark, R.G.: The impact of the mit-bih arrhythmia database. IEEE Engineering in Medicine and Biology Magazine 20(3), 45–50 (2001)

[42] Mori, J., Teranishi, I., Furukawa, R.: Continual horizontal federated learning for heterogeneous data (2022). https://doi.org/10.48550/ARXIV.2203.02108, https://arxiv.org/abs/2203.02108

[43] Müller, H., Holzinger, A., Plass, M., Brcic, L., Stumptner, C., Zatloukal, K.: Explainability and causability for artificial intelligence-supported medical image analysis in the context of the european in vitro diagnostic regulation. New Biotechnology 70, 67–72 (2022)

[44] Murat, F., Yildirim, O., Talo, M., Baloglu, U.B., Demir, Y., Acharya, U.R.: Application of deep learning techniques for heartbeats detection using ecg signals-analysis and review. Computers in biology and medicine 120, 103726 (2020)

[45] Murat, F., Yildirim, O., Talo, M., Baloglu, U.B., Demir, Y., Acharya, U.R.: Application of deep learning techniques for heartbeats detection using ecg signals-analysis and review. Computers in Biology and Medicine 120, 103726 (2020). https://doi.org/https://doi.org/10.1016/j.compbiomed.2020.103726, https://www.sciencedirect.com/science/article/pii/S0010482520301104

[46] Natarajan, A., Chang, Y., Mariani, S., Rahman, A., Boverman, G., Vij, S., Rubin, J.: A wide and deep transformer neural network for 12-lead ecg classification. In: 2020 Computing in Cardiology. pp. 1–4 (2020). https://doi.org/10.22489/CinC.2020.107

[47] Ookura, S., Mori, H.: An efficient method for wind power generation forecasting by lstm in consideration of overfitting prevention. IFAC-PapersOnLine 53(2), 12169–12174 (2020)

[48] Panch, T., Mattie, H., Celi, L.A.: The “inconvenient truth” about ai in healthcare. NPJ digital medicine 2(1), 1–3 (2019)

[49] Perez Alday EA, Gu A, S.A.: Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. Physiol Meas (2020 Nov 11). https://doi.org/10.1088/1361-6579/abc960, https://moody-challenge.physionet.org/2020/papers/

[50] Rajkumar, A., Ganesan, M., Lavanya, R.: Arrhythmia classification on ecg using deep learning. In: 2019 5th international conference on advanced computing & communication systems (ICACCS). pp. 365–369. IEEE (2019)

[51] Raza, A., Tran, K.P., Koehl, L., Li, S.: Designing ecg monitoring healthcare system with federated transfer learning and explainable ai. Knowledge-Based Systems 236, 107763 (2022)

[52] Rocher, L., Hendrickx, J.M., De Montjoye, Y.A.: Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications 10(1), 1–9 (2019)
[53] Rohmantri, R., Surantha, N.: Arrhythmia classification using 2d convolutional neural network. International Journal of Advanced Computer Science and Applications 11(4) (2020)

[54] Rosychuk, R.J., Mariathas, H.H., Graham, M.M., Holroyd, B.R., Rowe, B.H.: Geographic clustering of emergency department presentations for atrial fibrillation and flutter in alberta, canada. Academic Emergency Medicine 22(8), 965–975 (2015)

[55] da S. Luz, E.J., Schwartz, W.R., Cámara-Chávez, G., Menotti, D.: Ecg-based heartbeat classification for arrhythmia detection: A survey. Computer Methods and Programs in Biomedicine 127, 144 – 164 (2016). https://doi.org/https://doi.org/10.1016/j.cmpb.2015.12.008, http://www.sciencedirect.com/science/article/pii/S0169260715003314

[56] Sahoo, S., Kanungo, B., Behera, S., Sabut, S.: Multiresolution wavelet transform based feature extraction and ecg classification to detect cardiac abnormalities. Measurement 108, 55–66 (2017)

[57] Serhani, M.A., T. El Kassabi, H., Ismail, H., Nujum Navaz, A.: Ecg monitoring systems: Review, architecture, processes, and key challenges. Sensors 20(6), 1796 (2020)

[58] Sheller, M.J., Reina, G.A., Edwards, B., Martin, J., Bakas, S.: Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In: International MICCAI Brainlesion Workshop. pp. 92–104. Springer (2018)

[59] Tang, R., Luo, J., Qian, J., Jin, J.: Personalized federated learning for ecg classification based on feature alignment. Security and Communication Networks 2021 (2021)

[60] Van Panhuis, W.G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A.J., Heymann, D., Burke, D.S.: A systematic review of barriers to data sharing in public health. BMC public health 14(1), 1–9 (2014)

[61] Wang, F., Casalino, L.P., Khullar, D.: Deep learning in medicine—promise, progress, and challenges. JAMA internal medicine 179(3), 293–294 (2019)

[62] Wang, H., Sreenivasan, K., Rajput, S., Vishwakarma, H., Agarwal, S., Sohn, J.y., Lee, K., Papailiopoulos, D.: Attack of the tails: Yes, you really can backdoor federated learning (2020). https://doi.org/10.48550/ARXIV.2007.05084, https://arxiv.org/abs/2007.05084

[63] Wodschow, K., Bihrmann, K., Larsen, M.L., Gislason, G., Erbsøll, A.K.: Geographical variation and clustering are found in atrial fibrillation beyond socioeconomic differences: a danish cohort study, 1987–2015. International journal of health geographics 20(1), 1–10 (2021)

[64] Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST) 10(2), 1–19 (2019)

[65] Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST) Volume 10 Issue 2, Article No. 12 (2019).
[66] Zhang, M., Wang, Y., Luo, T.: Federated learning for arrhythmia detection of non-iid ecg. In: 2020 IEEE 6th International Conference on Computer and Communications (ICCC). pp. 1176–1180. IEEE (2020)

[67] Zhao, Z., Fang, H., Relton, S.D., Yan, R., Liu, Y., Li, Z., Qin, J., Wong, D.C.: Adaptive lead weighted resnet trained with different duration signals for classifying 12-lead ecgs. In: 2020 Computing in Cardiology. pp. 1–4 (2020). https://doi.org/10.22489/CinC.2020.112

[68] Zhao, Z., Fang, H., Relton, S.D., Yan, R., Liu, Y., Li, Z., Qin, J., Wong, D.C.: Adaptive lead weighted resnet trained with different duration signals for classifying 12-lead ecgs. In: 2020 Computing in Cardiology. pp. 1–4. IEEE (2020)

[69] Zhu, Z., Wang, H., Zhao, T., Guo, Y., Xu, Z., Liu, Z., Liu, S., Lan, X., Sun, X., Feng, M.: Classification of cardiac abnormalities from ecg signals using se-resnet. In: 2020 Computing in Cardiology. pp. 1–4 (2020). https://doi.org/10.22489/CinC.2020.281

[70] Zisou, C., Sochopoulos, A., Kitsios, K.: Convolutional recurrent neural network and lightgbm ensemble model for 12-lead ecg classification. In: 2020 Computing in Cardiology. pp. 1–4 (2020). https://doi.org/10.22489/CinC.2020.417