An Explainable Federated Learning Approach for Heart Disease Classification

Tanjila Alam Sathi

Department of Computer Science and Engineering
Islamic University of Technology (IUT)
March, 2025.

An Explainable Federated Learning Approach for Heart Disease Classification

by

Tanjila Alam Sathi

Supervisor
Dr. Md. Azam Hossain
Associate Professor
Department of Computer Science and Engineering
Islamic University of Technology

MASTER OF SCIENCE
IN
COMPUTER SCIENCE AND ENGINEERING

Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Board Bazar, Gazipur-1704, Bangladesh.
March, 2025.

CERTIFICATE OF APPROVAL

The thesis titled,"**An Explainable Federated Learning Approach for Heart Disease Classification**" submitted by Tanjila Alam Sathi, St. No. 201041021of Academic Year 2020-21 has been found as satisfactory and accepted as partial fulfillment of the requirement for the degree Master of Science in Computer Science and Engineering on March 11, 2025.

Board of Examiners:

Dr. Md. Azam Hossain — Chairman
Associate Professor, — (Supervisor)
Department of Computer Science and Engineering,
Islamic University of Technology (IUT), Gazipur.

Prof. Dr. Md. Hasanul Kabir — Member
Professor and Head, — (Ex-Officio)
Department of Computer Science and Engineering,
Islamic University of Technology (IUT), Gazipur.

Prof. Dr. Abu Raihan Mostofa Kamal — Member
Professor,
Department of Computer Science and Engineering,
Islamic University of Technology (IUT), Gazipur.

Prof. Dr. A.B.M. Aowlad Hossain — Member
Professor, — (External)
Department of Electronics and Communication Engineering,
Khulna University of Engineering and Technology (KUET), Khulna, Bangladesh.

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Tanjila Alam Sathi** under the supervision of **Dr. Md. Azam Hossain**, Associate Professor, Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Dr. Md. Azam Hossain
Associate Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: March 11, 2025.

Tanjila Alam Sathi
Student No.: 201041021
Date: March 11, 2025.

Dedicated to Allah (SWT), who has bestowed upon me with this knowledge, verily I belong to HIM, and, verily to HIM I will return

Table of Contents

Page

Acknowledgement **x**

Abstract **xi**

1 Introduction **1**

1.1 Motivation and Scope 1
1.2 Problem Statement 3
1.3 Research Challenges 3
1.4 Research Contributions 4
1.5 Organization 4

2 Background Study and Related Works **5**

2.1 Cardiovascular Disease 5
2.1.1 Arrhythmia 5
2.1.2 Types of Arrhythmia 6
2.1.3 Ischemia 6
2.1.4 Types of Ischemia 7
2.2 Electrocardiogram (ECG) Signal 8
2.3 Federated Learning 10
2.4 Related Works 11
2.4.1 ML and DL based Models 11
2.4.2 Federated Learning based Models 13

3 Proposed Methodology **14**

3.1 Overview 14
3.2 Materials and Methods 16
3.3 Dataset Acquisition 16
3.4 ECG Feature Extraction 17
3.4.1 ECG Fiducial Features 18
3.5 Data Augmentation 20

3.6 Federated Learning System 20
3.6.1 Flower for Federated Learning System 20
3.6.2 Federated Averaging Algorithm(FedAvg) 21
3.7 Proposed Federated System for Client side and Server side 22
3.7.1 Client-Side Pseudocode 22
3.7.2 Server-Side Pseudocode 23
3.8 Using Deep Learning Models in Client Side for Heart Disease Classification 25
3.8.1 ANN Model for Federated Learning 25
3.8.2 ANN Model Architecture 25
3.8.3 LSTM Model for Federated Learning 26
3.8.4 LSTM Model Architecture 27
3.9 Explainability with XAI Technique 29
3.9.1 Shapley Additive Explanations (SHAP) 30
3.9.2 Local Interpretable Model-Agnostic Explanations (LIME) . . 31
3.10 Evaluation Metrics 32
3.10.1 Accuracy 32
3.10.2 Precision 32
3.10.3 Recall 33
3.10.4 F1-Score 33
3.10.5 Loss 33
3.10.6 Misclassification Rate 34
3.10.7 False Positive Rate(FPR) 34
3.10.8 False Negative Rate(FNR) 34
3.10.9 AUC-ROC Score 34

4 Results and Discussions **36**
4.1 Experimental Setup 36
4.1.1 Environment 36
4.1.2 Dataset Split 36
4.1.3 Batch Size 37
4.1.4 Epoch and Learning Rate 37
4.1.5 Optimizer and Loss Function 37
4.1.6 Data Validation 37
4.2 Performance Evaluation and Comparison of FL-ANN and FL-LSTM Models 38
4.3 Impact of Number of Clients on the Models 42
4.4 Impact of Communication Rounds on the Models 43

4.5 AUC-ROC curve of the models 44
4.6 Model Explanation using SHAP 46
4.6.1 FL-ANN Model Explanation using SHAP 46
4.6.2 FL-LSTM Model Explanation using SHAP 47
4.7 Model Explanation using LIME 48
4.7.1 FL-ANN Model Explanation using LIME 48
4.7.2 FL-LSTM Model Explanation using LIME 49
4.7.3 Interpretation of ECG Fiducial Features for Arrhythmia and Ischemia 49
4.8 Comparison with State-of-the-Art Methods 52
4.9 Discussion 54

5 Conclusion **55**
5.1 Summary 55
5.2 Future Works 56

References **56**

List of Publications **64**

List of Figures

Page

2.1 (a) Cardiac Arrhythmia (b) Types of Arrhythmia. (Courtesy of [1]) . . 6
2.2 Cardiac Ischemia with narrowed coronary arteries. (Courtesy of [2]) . 8
2.3 The placement of ECG electrodes on the chest,arms,and legs.(Courtesy of [3]) 9
2.4 A standard 12-lead ECG signal.(Courtesy of [4]) 9
2.5 ECG wave features.(Courtesy of [5]) 10

3.1 A comprehensive system architecture for arrhythmia,ischemia, and healthy heart conditions recognition, employing ECG data alongside interpretable AI techniques. (a) Data Acquisition: ECG data collected from arrhythmia, ischemia, and healthy groups. (b) Feature Node: Extraction of fiducial features from ECG. (c) Model Framework: Utilization of Federated learning with deep learning for prediction of arrhythmia, ischemia, and healthy heart conditions. (d) Personalized Diagnostics: Explanation tailored for individual CVD cases [6]. 15
3.2 Federated Learning Architecture.(Courtesy of [7]) 21
3.3 ANN Model Architecture 27
3.4 LSTM Model Architecture 29

4.1 Performance comparison and analysis of FL-ANN and FL-LSTM Models in 16 rounds based on evaluation metrics. 42
4.2 FL-ANN clients vs accuracy 43
4.3 FL-LSTM clients vs accuracy 44
4.4 FL-ANN rounds vs accuracy 44
4.5 FL-LSTM rounds vs accuracy 45
4.6 Receiver Operating Characteristic (ROC) curves with AUC scores of three classes for (a) FL-ANN Model (b) FL-LSTM model 46

4.7 Interpreting ECG feature contributions in arrhythmia, ischemia, and healthy heart conditions using SHAP plots, ranked by importance. (a) Waterfall SHAP Plots for Individual Feature Importance of FL-ANN Model; (b) Bar SHAP Plots for Feature Importance of FL-ANN Model 47
4.8 Interpreting ECG feature contributions in arrhythmia, ischemia, and healthy heart conditions using SHAP plots, ranked by importance. (a) Waterfall SHAP Plots for Individual Feature Importance of FL-LSTM Models; (b) Bar SHAP Plots for Feature Importance of FL-LSTM Models 47
4.9 Individualized Interpretation of ECG-based heart disease prediction using the LIME Approach with FL-ANN model: (a) The cells highlighted in blue indicate the features that had the greatest influence on predicting arrhythmia risk. (b) The cells highlighted in orange indicate the features that had the greatest influence on predicting healthy individuals. (c) The cells highlighted in green indicate the features that had the greatest influence on predicting ischemia risk. 50
4.10 Individualized Interpretation of ECG-based heart disease prediction using the LIME Approach with FL-LSTM model: (a) The cells highlighted in blue represent the features that had the most significant impact on predicting arrhythmia risk. (b) The cells highlighted in orange represent the features that had the most significant impact on predicting healthy individuals. (c) The cells highlighted in green represent the features that had the most significant impact on predicting ischemia risk. 51

List of Tables

Page

3.1 Dataset Description 17

4.1 FL-ANN model 39
4.2 FL-LSTM model 40
4.3 Comparison with state-of-the-art methods for heart disease classification 53

Acknowledgment

First, I sincerely thank Almighty Allah (SWT) for giving me this knowledge and enabling me to complete this thesis for the Master of Science in Engineering degree on time, and for blessing me with good health and energy to carry out this research successfully.

I express my sincere gratitude, appreciation, and gratitude to Dr. Md. Azam Hossain, Associate Professor in the Department of Computer Science and Engineering, Islamic University of Technology, for his assistance, close supervision, academic guidance, instructions, encouragement, and valuable criticism throughout the research project.

I would never have achieved this milestone without the unwavering support of Jinat Alam and Rafsan Jany, for whom I am incredibly grateful. I express my deep admiration and sincere gratitude to the entire faculty and staff of the Islamic University of Technology's Department of Computer Science and Engineering for their enthusiastic collaboration and genuine assistance in conducting the project.

Finally, I would like to express all my appreciation to my beloved parents and my spouse for their unwavering support and inspiration.

Abstract

Cardiovascular disease (CVD) is a life-threatening medical condition that affects the heart and blood vessels, leading to substantial financial and social burdens. Electrocardiograms (ECG) are extensively used to detect and treat heart diseases as these are rapid, secure, non-invasive, and cost-effective. The ECG signal has several segments, including the P-wave, QRS complex, QT interval, ST segment, and T-wave, together with the associated onset, offset, and peak points referred to as fiducial points. Abnormalities in these components can be indicative of various heart diseases. Medical professionals face challenges when using the conventional approach to ECG recording, including comprehending the intricate nature of the ECG to interpret, dealing with signal interference, and handling the frequent coexistence of other health conditions alongside heart disease. However, accurate identification of fiducial points associated with the ECG wave is crucial for proper assessment of heart disease. To automate the diagnosis of heart diseases by analyzing ECG signals, various machine learning and deep learning classification models are used where data must be centralized. The majority of contemporary research encounters challenges due to data privacy, limited dataset, and lack of collaboration and proper explanation of the result. To address these challenges, this thesis presents an explainable federated learning model for the classification of heart disease using ECG fiducial features. This approach is the first of its kind to classify three distinct heart conditions: arrhythmia, ischemia and healthy states while ensuring both data privacy, as well as the interpretability of the outcome. We introduced two models for classifying three distinct heart conditions: Federated Learning with Artificial Neural Networks (FL-ANN) and Federated learning with long-short-term memory (FL-LSTM), which achieved impressive accuracies of 87% and 90%, respectively. In addition, explainable artificial intelligence(XAI) models can improve trust by providing visual interpretations of their results and decisions. The proposed explainable models highlighted the fiducial features of the ECG, including P-wave height (P-H), R-wave height (R-H), the QRS complex, heart rate (HR), QT interval, and corrected QT interval (QTc), which provide invaluable insights into the heart's electrical activity, serve as vital markers for cardiac disorders, and significantly improve diagnostic precision. Our explainable federated learning models facilitate data collaboration while safeguarding the privacy of sensitive information. By ensuring the explainability of the outcome, these models are designed to support healthcare professionals, helping them make better and more accurate diagnostic decisions.

Chapter 1

Introduction

1.1 Motivation and Scope

Cardiovascular disease (CVD), which primarily affects the heart or blood vessels, is the leading cause of mortality worldwide. CVDs are expected to become more prevalent, resulting in an escalation of mortality rates. The projected number of deaths associated with CVD is expected to increase from around 18.9 million in 2020 to more than 22.2 million in 2030 and 32.3 million in 2050 [8]. Among cardiovascular diseases, ischemia and arrhythmia are two of the most prevalent, contributing to the highest mortality rate [9]. Ischemia is a condition of the coronary arteries where a portion of the myocardium does not get adequate blood flow, leading to sudden cardiac death. Cardiac arrhythmia, which becomes more common in elderly people, is characterized by an irregular heart rate (HR) resulting from irregular electrical activity in the atria. An early and accurate prognosis of cardiac diseases can prevent fatalities and improve the quality of life.

The electrocardiogram (ECG) is an imaging representation of the electrical signals produced by the heart. It has become an established diagnostic method for evaluating the prognosis for various cardiac conditions. ECG serves as a quick, safe, non-invasive, low-cost, and pain-free test for diagnosing ischemia, arrhythmia, healthy individual, and other CVD conditions [10]. Five distinct waves, namely P, Q, R, S, and T, comprise the ECG waveform, making up a heartbeat. The ECG signal is composed of multiple segments, including the P-wave, QRS complex, QT interval, ST segment, and T-wave, as well as the accompanying onset, offset, and peak points known as fiducial points [11]. ECG or vital sign capabilities are now accessible in the majority of fitness trackers and wellness gadgets, thanks to recent advancements in the Internet of Things (IoT), wearable technology, digital twins, and Healthcare 4.0 in medicine [12]. As a result, real-time monitoring of heart activity provides a cheap and effective way of predicting high-risk patients with cardiovascular disease [13]. The World Health

Foundation's World Heart Vision 2030 emphasized that everyone, regardless of nation, geography, origin, race, gender, age, education, or income, has the right to cardiovascular health and well-being [8]. This includes health promotion, preventive care, and the therapy of cardiovascular disease, including rehabilitation. Furthermore, the government and healthcare institutions are looking for a novel and efficient approach to heart disease prognosis and treatment [14].

Traditionally, healthcare practitioners often use conventional 12-lead ECG equipment with several electrodes to capture ECGs at medical facilities and hospitals. These ECG procedures require skilled medical personnel and clinical environments to accurately assess the heart's electrical activity in multiple spatial positions. Long clinical preparation and expert skill demand may delay the prognosis of acute heart disease [12]. Furthermore, numerous tests are necessary to accurately predict cardiac disease. Due to a lack of competence among healthcare professionals, incorrect predictions may occur, making early detection difficult. This procedure becomes substantially more challenging when the ECG data contains noise [15]. As a result, there is a considerable emphasis on developing machine learning (ML) models for automatically detecting cardiac conditions. Advancements in machine learning (ML) and deep learning (DL) primarily drive the effective integration of artificial intelligence (AI) into healthcare and hospitals. AI has completely revolutionized our approach to real-world operations traditionally performed by humans. These models have many applications, including disease diagnosis, medication discovery, tailored treatment planning, activity identification, predictive analytics, and other areas [16].

However, data must be present in a centralized location for the existing ML or DL models. Medical facilities are now forced to rely on their own data sources, which may be skewed by factors like clinical specialty, instrument usage, or patient demographics in particular. If they don't, they'll have to combine data from other organizations to get all the information they require, which means handling regulatory concerns. So, the bulk of current research on heart disease classification faces obstacles due to lack of collaboration, data privacy issues and limited dataset. Therefore, while respecting patients' data privacy, federated learning combined with deep learning can be utilized to automatically classify heart diseases. Federated learning(FL) is a decentralized method that enables several organizations to work together to build a model without exchanging data. By facilitating cooperative model training among various healthcare facilities without exchanging patients' data, FL preserves patient privacy and complies with laws such as HIPAA and GDPR. Through federated learning, AI algorithms may learn from a large variety of data that are spread across multiple locations.

Nevertheless, these models are complex and very challenging to grasp since they

have a tendency to obscure their underlying logic and internal decision-making mechanisms. For this reason, these models are often referred to as black boxes because of their lack of interpretability. As a consequence, users find it more difficult to accept the results owing to lack of transparency and interpretability, which limits their implementation in sensitive areas such as healthcare, finance, and law. Users in these fields need explanations to understand and interpret the results [10].

Recently, significant progress has been made in the field of Explainable Artificial Intelligence (XAI), which has evolved into a technique for improving the interpretability and reliability of the models. Researchers often use feature importance strategies to explain the relationship between features and the outcome and these techniques provide information on the contribution of each feature to the output [17]. There are two approaches for determining the importance of features: global and local explanation methods. The global explanation emphasizes feature-level importance scores, which quantify the contribution of a particular input feature to the output of a model and a local explanation describes how particular features influence a particular observation [10]. So based on the above motivation, an explainable federated learning model is developed in this work for heart disease classification using ECG features.

1.2 Problem Statement

Based on the above discussion, this work aims to develop an explainable federated learning-based heart disease classification model to efficiently classify heart diseases while maintaining privacy and to provide the explainability of the result produced. The specific objectives of this research are:

i. To develop a federated learning model for heart disease classification.

ii. To apply effective XAI methods to understand and explain the results.

iii. To analyze the comparative performance against different existing standard methods.

1.3 Research Challenges

- The first research challenge is to build a reliable heart disease dataset that includes data for three distinct types of heart disease. This involves addressing the scarcity of publicly available datasets and overcoming issues such as class imbalance, missing data, and irrelevant data, all of which can pose significant challenges for the model.

- To develop a federated learning (FL) model for heart disease classification while ensuring privacy and enabling data collaboration, the challenge lies in satisfying a cross-domain scenario.
- To offer clinical interpretation of the FL models' outcome is also a challenging task.

1.4 Research Contributions

The key contributions of this work can be described in three folds:

i. To address the research challenge, we combined three publicly available and widely used datasets: MIT-BIH Arrhythmia [18], European ST-T [19], and Fantasia [20], to create a larger and more reliable dataset for training our model.

ii. To address the privacy concerns and enable data collaboration, a federated learning model was developed to automatically classify three different heart conditions, namely arrhythmia, ischemia, and healthy.

iii. To explain the importance of ECG features and provide visual interpretation of the results, SHAP [21] and LIME [22] interpretable frameworks were employed to offer clinical interpretation of the models' outcome.
To our knowledge, this study represents the first attempt to classify three distinct heart conditions using an explainable federated learning model.

1.5 Organization

The rest of the dissertation is organized as follows. Chapter 2 discusses the background study and related works of heart disease classification. It also identifies the problems persistent in the existing literature. Chapter 3 presents the proposed federated learning (FL) model architecture for heart disease classification. Chapter 4 presents the results of FL models and XAI system to provide the explanation of the output. Chapter 5 highlights the contributions and provides direction for future research scopes.

Chapter 2

Background Study and Related Works

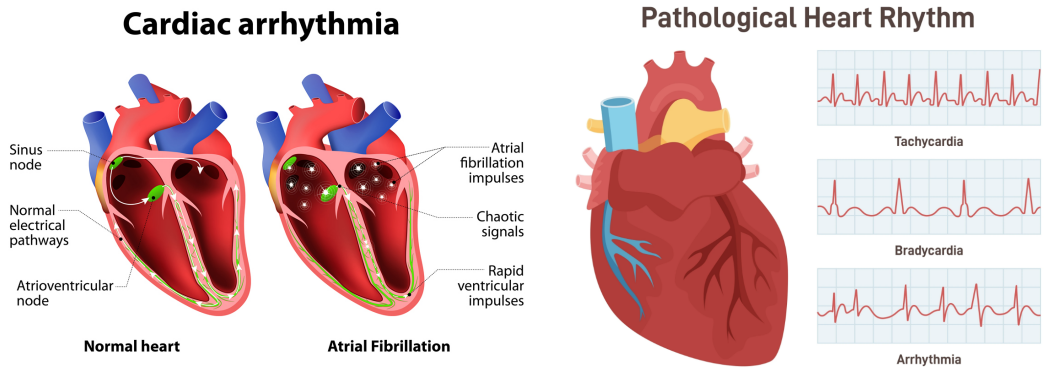
In this chapter, we have provided an overview of the background and relevant works related to heart disease classification. We discussed about various types of cardiovascular diseases, delved into the significance of electrocardiograms(ECG) in diagnosing these conditions, and the application of federated learning systems.

2.1 Cardiovascular Disease

Cardiovascular disease, a leading cause of morbidity and mortality worldwide, encompasses a group of conditions that affect the heart and blood vessels. These disorders can impact one or more areas of the cardiovascular system, including the heart and various blood vessels. Four out of every five deaths occur due to heart failure, according to the World Health Organization (WHO) [8]. An early and accurate prognosis of cardiac diseases can prevent fatalities. Over the years, researchers have developed a variety of techniques aimed at accelerating the early diagnosis of cardiovascular diseases such as coronary artery disease, arrhythmia, ischemia, heart failure or stroke.

2.1.1 Arrhythmia

An irregular heartbeat, medically referred to as arrhythmia or atrial fibrillation, is a condition characterized by an abnormal heart rhythm. The heart may beat too rapidly, too slowly, or in an irregular pattern. It is normal for heart rate to rise during physical exertion and to decrease during periods of rest or sleep. However, a frequent irregular heartbeat could indicate that body isn't receiving enough blood from the heart [23]. A person may experience lightheadedness, fainting, or other symptoms. In Figure 2.1(a), a normal heart and a heart with arrhythmia is shown where it is seen that arrhythmia occurs due to rapid ventricular impulses and irregular or chaotic signals of the heart [1].



(a) **(b)**

Figure 2.1: (a) Cardiac Arrhythmia (b) Types of Arrhythmia. (Courtesy of [1])

2.1.2 Types of Arrhythmia

There are many types of arrhythmias as shown in Figure 2.1(b). It depends on which part of the heart is affected and whether these cause a slow, fast, or irregular heart rate. Arrhythmias may happen in the atria (upper chambers of the heart) or the ventricles (lower chambers of the heart) [1].

- Bradycardia: Bradycardia occurs when the resting heart rate is less than 60 beats per minute [1]. Some people may naturally have slower heart rates, particularly those who are youthful or physically fit [1].
- Tachycardia: When the resting heart rate exceeds 100 beats per minute, it is referred to as tachycardia and also some may have an irregular heartbeat [1].
- A premature or extra heartbeat happens when the signal to beat comes too early which causes a pause, and when heart settles back into its normal rhythm, the beat gets stronger [1]. It can be felt as if it missed a beat. This is a common type of arrhythmia that can lead to other types of arrhythmias [1].

If left untreated, arrhythmias can negatively affect the heart, brain, or other vital organs. They may lead to serious complications such as heart failure, cardiac arrest, or a potentially fatal stroke. Cardiac arrest occurs when the heart suddenly and unexpectedly ceases to beat, and without immediate medical intervention, it can result in death within minutes.

2.1.3 Ischemia

Ischemia refers to a condition in which there is insufficient blood flow to a specific area of the body due to the obstruction of blood vessels. This condition arises when an

organ, such as the heart, does not receive an adequate supply of blood and oxygen. As illustrated in Figure 2.2, the narrowing of the coronary arteries which supply blood to the heart muscle is known as ischemic heart disease, also referred to as coronary heart disease (CHD) or coronary artery disease. This occurs when reduced blood flow limits the oxygen supply to the heart muscle. The reduced blood flow is usually the result of a partial or complete blockage of the heart's arteries (coronary arteries) [2]. A heart attack, also known as a myocardial infarction (MI), occurs when the heart muscle's whole blood supply is cut off, causing the heart muscle cells to die [2].

2.1.4 Types of Ischemia

There are many types of ischemia depending on which area of the body has less blood flow than the regular flow [2]. As, blood carries oxygen to the cells and tissues, so when there is less oxygen, it causes ischemia.

- Myocardial Ischemia: When heart muscle does not receive enough blood, it is known as myocardial ischemia. This indicates that heart muscle is not receiving enough blood to perform its necessary functions. The accumulation of fat and cholesterol, or plaque, that prevents enough blood from passing through coronary arteries is frequently the cause. Treatments for myocardial ischemia include medications and surgery [2].

- Ischemic stroke: It occurs when there is a blockage that prevents blood flow to all parts of the brain. Without blood supply,the affected parts stop working and begin to die. An ischemic stroke can result in death or irreversible brain damage if blood flow doesn't return promptly enough. This is a medical emergency that requires prompt attention [2].

- Transient Ischemic Attacks(TIAs): These are just as dangerous as a stroke, despite being mistakenly referred to as "mini-strokes." A stroke is frequently imminent when a transient ischemic attack (TIA) occurs. A TIA is a serious medical condition that should not be ignored. More significantly, it's an opportunity to receive care that may help stop another stroke [2].

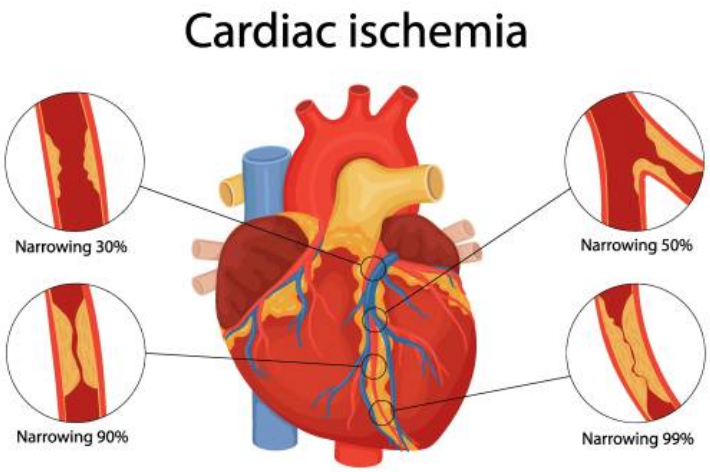


Figure 2.2: Cardiac Ischemia with narrowed coronary arteries. (Courtesy of [2])

2.2 Electrocardiogram (ECG) Signal

Over the years, researchers have developed a wide range of methods in an effort to accelerate the automated diagnosis of neurological and cardiovascular disorders [17]. The electrical activity of the heart is measured by an ECG device through sensors and electrodes that are placed on the patient's arms, legs, and chest, as shown in Figure 2.3. The electrical signals captured by these electrodes are associated with a 12-lead ECG machine that records the combined electrical activity of the heart from various angles over some time [24]. The three bipolar leads, out of the twelve leads, evaluate the potential differences between one arm and the leg and between both arms. The remaining nine unipolar electrodes are made up of six limb leads (I, II, III, aVR, aVL, and aVF) that aid in viewing the heart in the vertical plane and six chest leads (V1 to V6) that observe the heart in the horizontal plane [4], as shown in Figure 2.3. A standard ECG record is shown in Figure 2.4. A single cycle of an ECG contains a pattern of waves, as shown in Figure 2.5. When the sinoatrial (SA) node triggers an impulse, the atrial fibers depolarize to produce a potential difference called a P wave, leading to atrial contraction [25]. In a normal ECG, as shown in Figure 2.5, a P wave has a duration of about 0.08 s [25]. A P wave is observable in leads II and V1. Additionally, it appears inverted in lead aVR and upright in leads I and II, as illustrated in Figure 2.4. Following the depolarization of the atrial fibers, the electrical impulse is transmitted to the ventricular fibers, resulting in their rapid depolarization. The depolarization causes additional electrical changes because the ventricular walls are thick; this is known as the QRS complex, and it is made up of Q, R, and S waves. Additionally, the QRS complex lasts for around 0.08 seconds [25]. A T wave is then created as the ventricles

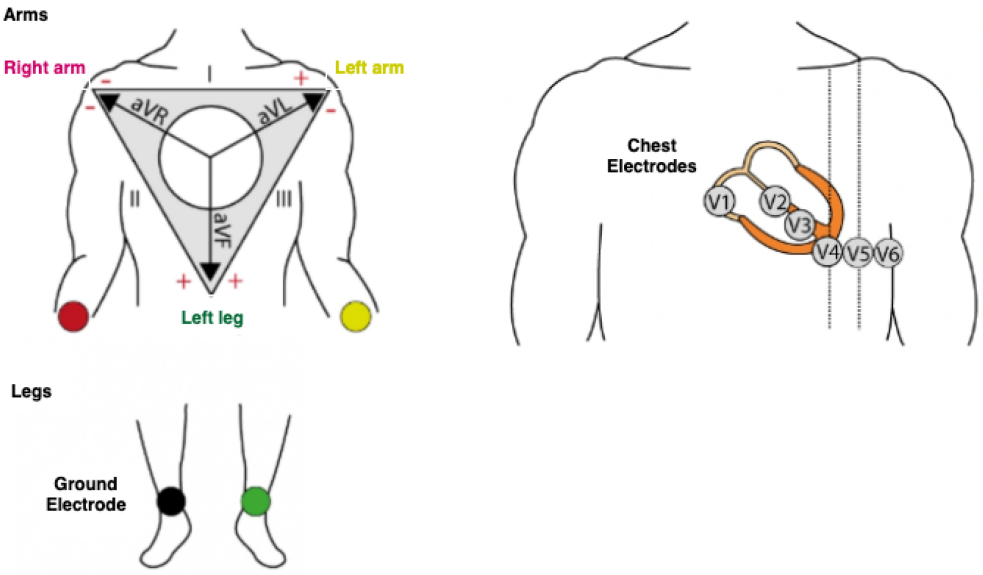


Figure 2.3: The placement of ECG electrodes on the chest,arms,and legs.(Courtesy of [3])

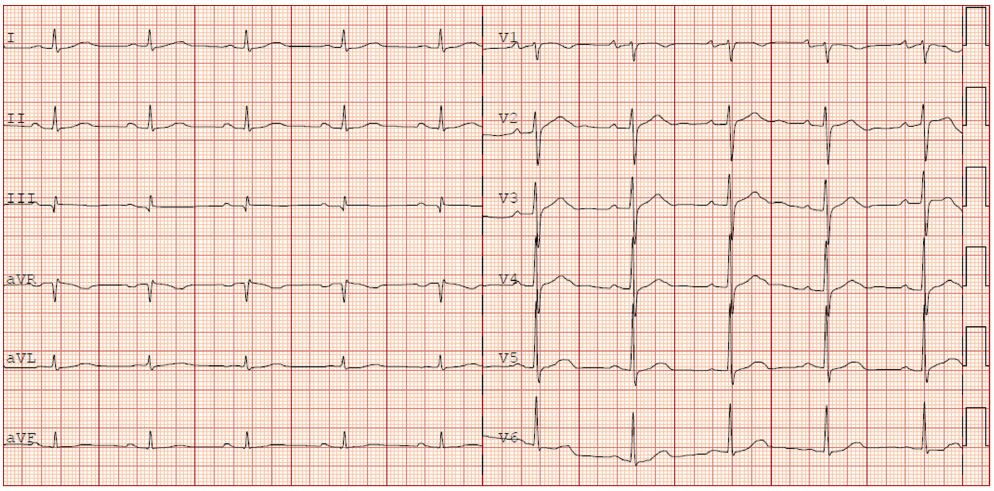


Figure 2.4: A standard 12-lead ECG signal.(Courtesy of [4])

repolarize. In a typical ECG, the T wave lasts roughly 0.16 seconds. Due to atrial fiber repolarization occurring concurrently with ventricular fiber depolarization, the atrial repolarization is absent from the pattern [25].

As depicted in Figure 2.5, the PR interval is the period between the P wave and the QRS complex. Impulse transmission periods between the atrioventricular (AV) and SA nodes are shown by the PR interval. It includes contraction, atrial depolarization, and depolarization waves through the conduction system. In contrast, the ST segment, which lasts roughly 0.22 seconds, occurs when the ventricular myocardium depolarizes. The QT interval, which lasts approximately 0.38 seconds, is the time between ventricular depolarization and repolarization [25]. The absence of a significant amount of potential difference in the ventricular myocardial cells is indicated by the

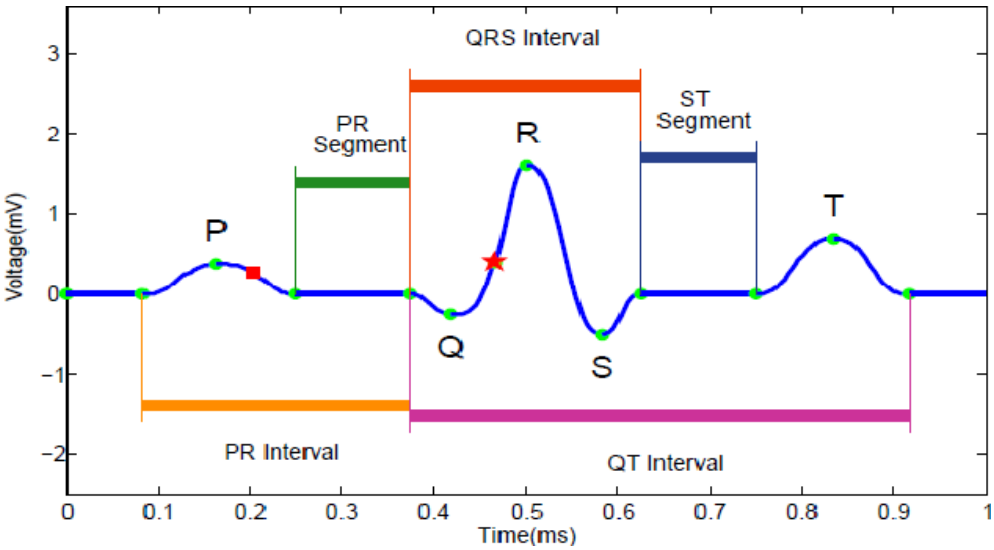


Figure 2.5: ECG wave features.(Courtesy of [5])

TP segment, an isoelectric area. This is the ventricular myocardial cell's resting state, which spans the interval between the conclusion of repolarization and the start of the subsequent depolarization [26]. Any variation from this typical cardiac cycle can be a sign of conduction system issues and heart diseases.

2.3 Federated Learning

Federated learning model allow many organizations to work together on model building without having to exchange private clinical data directly. Over a number of training cycles, the shared models are exposed to a far greater variety of data than any one business could possibly have on hand. This method increases the predicted accuracy for disease outbreaks, diagnostics, and individualized treatment plans by utilizing a variety of datasets from several sources to create strong, generalizable models. By keeping data localized, FL reduces the risk of data breaches and lowers data transfer costs, making it a resource-efficient solution. During successive training iterations, the shared models encounter a much more extensive variety of data compared to what is available within a single organization. This approach allows for the creation of robust, generalizable models by leveraging diverse datasets from various sources, which enhances predictive accuracy for disease outbreaks, diagnostics, and personalized treatment plans.

2.4 Related Works

In this section, we presented a comprehensive overview of the background and relevant studies on heart disease classification.

2.4.1 ML and DL based Models

Various machine learning (ML) and deep learning (DL) techniques use ECG features to diagnose various cardiovascular diseases [27]. Murthy et al. [28] investigated the European ST-T database [19] and compared the Artificial Neural Networks (ANN), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) classifiers to predict cardiac ischemia by extracting morphological features from ECG. Using the same databese in [29], SVM and Kernel Density Estimation (KDE) techniques were applied to identify ischemia in ECGs.Several cardiovascular conditions are linked to cardiac arrhythmia, which is evidenced by irregular heartbeats and a collection of abnormal heart rhythms. An assessment for irregularities in the R-R interval or the absence of P waves is regarded as a biomarker associated with arrhythmia [30]. The availability of the open-source ECG MIT-BIH arrhythmia database [18] has enabled scientists to classify cardiac arrhythmia from various statistical and morphological features extracted from ECG data using various machine learning (ML) models [31–34]. Kuila et al. [35] used a Linear Discriminate (LD) classifier to categorize different heartbeats, based on fiducial parameters such as RR interval, heartbeat interval, and segmentation ECG morphological characteristics. Sraitih et al. [36] leveraged an ensemble technique of three supervised machine learning models—Random Forest (RF), KNN, and SVM—to automatically identify arrhythmia in ECG data. Khan et al. [37] presented considerable insight into existing methodologies and limitations of arrhythmia classification techniques and applied various deep learning architectures to classify arrhythmia. The study [38] surveyed and explored the current trend of deep learning approaches for classifying arrhythmia using ECG data. The majority of the existing approaches rely on traditional machine learning or deep learning models, which are constrained by the availability of limited dataset and the model's opaqueness. Tougui et al. [39] investigated six data mining and machine learning techniques for cardiovascular disease detection and demonstrated that the artificial neural network (ANN) outperforms Logistic Regression, Support Vector Machine, K Nearest Neighbors, and Random Forest. Ghosh et al. [15] merged five datasets to create a large dataset and employed two feature selection methods Relief and LASSO, revealing that the Random Forest Bagging Method (RFBM) showed superior performance. Deep residual convolutional neural networks (CNN) were utilized to classify heartbeats using PhysionNet's MIT-BIH and

PTB Diagnostics datasets to improve arrhythmia predictions [40]. In [41], a framework composed of two deep-learning models, CNN and LSTM, is used to predict five micro-classes of arrhythmias.

Two machine learning models Support Vector Machine (SVM) and Multi-layer perceptron classifier(MLP) were used to classify various cardiovascular diseases from extracted features [42]. Atrial Fibrillation (AF) is one of the dreadful arrhythmia. To identify Atrial Fibrillation from the heart rate segment that is obtained from the ECG signal, Yao et. al [43] utilized the multi-scale convolutional neural networks(MCNN). Liaqat et al. [44] implemented machine learning and deep learning algorithms to identify Atrial Fibrillation. Acharya et al. [45] proposed a convolutional neural network (CNN) technique to identify distinct ECG components for the automated identification of arrhythmia [45]. To identify 3 different types of arrhythmia, discrete wavelet transform and thresholding methods were utilized with the help of dynamic features of ECG signal in [46]. Oh et al. [47] proposed an automated system to classify various cardiovascular diseases by using a combination of deep learning LSTM and CNN without hand-crafted features. Sattar et al. [48] proposed a CNN model that utilized the digitized form of ECG signals to classify ECG signals for the classification of cardiac arrhythmia. Though Deep learning networks generate exceptional results, the drawbacks lie in the data imbalance and security issue of patients' health data.

The current state of machine learning and deep learning models lacks the capability to provide adequate explanations, thereby impeding their extensive adoption and implementation in critical domains like healthcare. Recent research has focused extensively on the application of XAI technologies. These interpretation approaches facilitate human experts to rely on the outputs of models, debug and troubleshoot models, and elude various model biases [49]. Gunning et al. [50] reported that users can comprehend, adequately trust, and administer the next generation of artificially intelligent by utilizing a variety of XAI tools.

Dave et al. [51] supported the prospects of XAI in the healthcare domain and presented examples based on a heart disease dataset to clarify the preferred XAI techniques to create trustworthiness in healthcare. Abdullah et al. [52] presented a comprehensive survey on the uses of explainable machine learning techniques on ECG signal-based heart disease classification. The work [53, 54] involves the use of important global features and the identification of features with significant clinical relevance based on their value in the output classification of the machine learning model.The research conducted by [55] examines the use of ensemble classifiers within the explainable artificial intelligence (XAI) framework to predict heart diseases. The article [56] proposed employing time series derivatives in XAI approaches to measure the local

feature importance in the temporal domain on the MIT-BIH arrythmia dataset. Matteo et al. [57] employed Random Forest (RF) to classify myocardial infarction and used the LIME technique to interpret the key input segments for detection. Moreover, existing conventional methods are often compared with "black box" since even the model designers are unable to explain why a particular decision or prediction is made by these models. Explainable artificial intelligence (XAI) is necessary because the results generated from black-box models have led to a lack of accountability, transparency, and trust.

2.4.2 Federated Learning based Models

Data must be present in a centralized location for the existing machine learning or deep learning models. So the bulk of current research on heart disease classification faces obstacles due to limited dataset, lack of collaboration, and data privacy issues. In addition, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) requires worldwide rules to protect patient-sensitive health information from unauthorized disclosure. To address these challenges, Federated Learning (FL) is emerging as a popular privacy-preserving collaborative learning approach. Unlike traditional machine learning methods, FL trains an algorithm across multiple independent sessions, each utilizing its own distinct dataset. It can be used to process electronic health records (EHR) to build a shared model while keeping the data distributed across many geographic locations [58]. Various FL algorithms like Federated Averaging(FedAvg), Federated Stochastic Variance Reduced Gradient are evaluated on the MNIST dataset and FedAvg achieved superior performance [59]. In a federated setting [7], a CNN-based classifier is developed to categorize an input ECG signal for various arrhythmia detection and compared the result using noisy and clean data from MIT-BIH Arrhythmia dataset. A framework is used based on federated matched averaging (FedMA) with modified Artificial Bee Colony (MABC) optimization algorithm to enhance the diagnostic procedure for the prediction of heart disease [60]. A hybrid technique comprised of a modified artificial bee colony and support vector machine (MABC-SVM) with FedMA is developed in [61] for heart disease prediction where the MABC-SVM technique is used by the local model and FedMA is utilized by the global model. From the above analysis, it is observed that currently there are very few federated learning research studies for heart disease classification and the vast majority of these studies lack an explanation of the results despite the fact that it is very important in healthcare.

Chapter 3

Proposed Methodology

The amazing ability of deep learning models to automatically learn features has resulted in their wide usage in heart disease classification tasks. However, most of these models lack privacy and provide no explanation of the outcome. With the goal of tackling these limitations, we have proposed an explainable federated learning approach for heart disease classification that ensures state-of-the-art-level performance along with applicability in local devices. In this chapter, we discussed the pipeline of the proposed explainable federated learning model for heart disease classification.

3.1 Overview

The proposed architecture aims to develop an explainable federated learning model for heart disease classification that accurately classifies subjects into arrhythmia, ischemia, and healthy groups and provides clinical interpretation with the help of explainable artificial intelligence. As illustrated in Figure 3.1(a), this study merged and integrated three widely used ECG databases: the MIT-BIH Arrhythmia, the European ST-T(EDB), and FANTASIA. Fiducial features of ECG recordings were extracted to identify heart disease. A number of reference features had been obtained from each electrocardiogram (ECG) record, as illustrated in Figure 3.1(b). Subsequently, the data SMOTE technique [62] was applied to balance the total dataset and then distributed among clients. Figure 3.1(c) shows the models of this work that were utilized on the server side and the client side. It receives ECG feature data as input and output the class labels. This work employed federated learning, a collaborative approach where multiple decentralized devices work together to train a shared model. The central server initialized and distributed a global model to the decentralized clients. Two deep learning algorithms were deployed, that is, ANN and LSTM, for the client models, which were trained locally using their own data. After local training, the clients' model updates are sent (such as gradients or parameter changes) back to the central

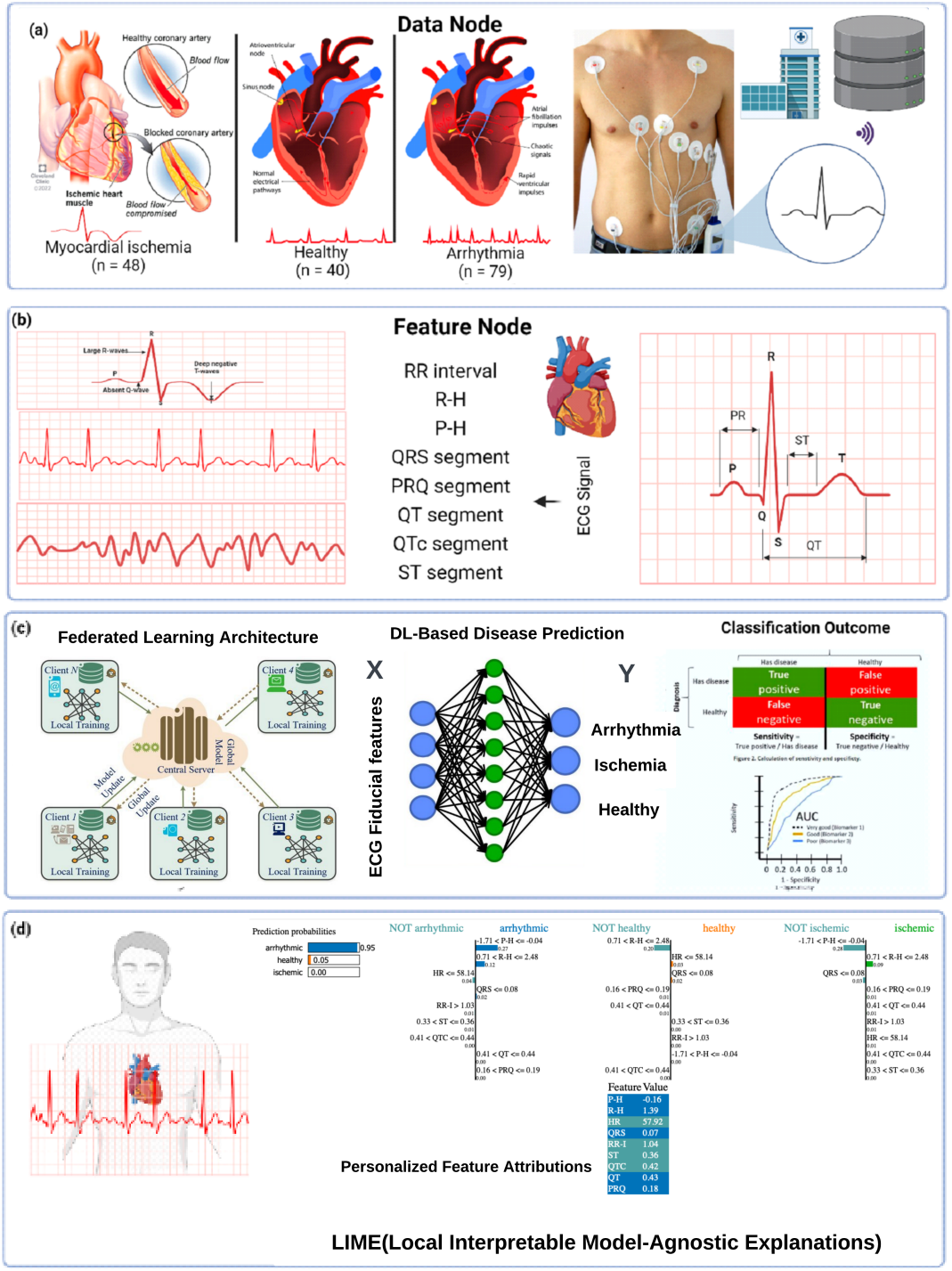


Figure 3.1: A comprehensive system architecture for arrhythmia,ischemia, and healthy heart conditions recognition, employing ECG data alongside interpretable AI techniques. (a) Data Acquisition: ECG data collected from arrhythmia, ischemia, and healthy groups. (b) Feature Node: Extraction of fiducial features from ECG. (c) Model Framework: Utilization of Federated learning with deep learning for prediction of arrhythmia, ischemia, and healthy heart conditions. (d) Personalized Diagnostics: Explanation tailored for individual CVD cases [6].

server. These updates were then aggregated by the server to form an improved global model with the Federated Averaging (FedAvg) algorithm [59], typically by averaging the received updates. This updated global model was sent back to the clients, and the process was repeated. Explainable artificial intelligence (XAI) models were applied to offer a clinical interpretation of the outcome, as shown in Figure 3.1(d). The LIME and SHAP explanation models were applied to the testing data set to identify the contribution of each feature to the model outcome. The application of explainable artificial intelligence (XAI) methods, such as SHAP and LIME, enhances the clinical applicability of these models by offering transparency in decision making. These XAI methods elucidate the complex relationships between the models and key ECG fiducial characteristics, providing a clear understanding of how each model uses specific parameters to predict ischemia, arrhythmia, or general heart conditions. The proposed explainable models highlighted the fiducial features of the ECG, including P-wave height (P-H), R-wave height (R-H), the QRS complex, heart rate (HR), QT interval, and corrected QT interval (QTc), which provide invaluable insights into the heart's electrical activity and serve as vital markers for cardiac disorders. This transparency is essential for physicians and researchers, allowing them to understand and trust the models' predictions regarding heart disease diagnosis and treatment.

3.2 Materials and Methods

This section provides a comprehensive description of the proposed approach, encompassing the acquisition of ECG datasets, data prepossessing and the integration of datasets and the extraction of relevant features. Furthermore, the FL based classification model and the explainable artificial intelligence (XAI) methods employed for model explanation are also described. The general pipeline of the proposed approach is depicted in Figure 3.1.

3.3 Dataset Acquisition

The study aims to develop federated learning based classification model that accurately recognize subjects into arrhythmia, ischemia, and healthy groups. To achieve this, three widely used ECG databases was integrated: the MIT-BIH Arrhythmia Database (MIT-BIH) [18], the European ST-T Database (EDB) [19], and the FANTASIA [20]. The following Table 3.1 provides detailed description of the three datasets. The MIT-BIH dataset was the first dataset used for the assessment of arrhythmia. It consists of 48 recordings obtained from a sample of 47 patients, including 25 males and 22 females. The age range for the male patients is 32 to 89 years, while the age range for the female

patients is 27 to 89 years. The European ST-T Database (EDB) comprises ECG records specifically focused on myocardial ischemia. The EDB database consists of a total of 90 annotated segments of ambulatory ECG recordings obtained from a cohort of 79 individuals which consists of 70 male participants ranging in age from 30 to 84, as well as 8 female participants ranging in age from 55 to 71. The Fantasia Database comprises electrocardiogram (ECG) and respiratory recordings, together with beat annotations, obtained from a group of 20 young and 20 elderly participants who were assessed as being in a healthy condition during a period of rest. Every subgroup of individuals consists of an equal number of male and female participants.

Table 3.1: Dataset Description

Database	Data Label	Male or Young	Age Range	Female or Elderly	Age Range
MIT-BIH Arrhythmia Database [18]	Arrhythmia	25 male	32 to 89 years	22 female	32 to 89 years
European ST-T Database (EDB) [19]	Ischemia	70 male	30 to 84 years	8 female	55 to 71 years
FANTASIA Database [20]	Healthy	20 young	21 to 34 years	20 elderly	68 to 85 years

3.4 ECG Feature Extraction

The non-invasive electrocardiogram (ECG) is a reliable tool for the early diagnosis of cardiac conditions. The analysis of ECG begins with preprocessing, which removes various noises. After preprocessing, the features are extracted in either the temporal or frequency domains. AcqKnowledge version 5.0 (Biopac Systems Inc., Goleta, CA, USA) [63] was used to extract fiducial features from ECG data which is an interactive, user-friendly tool that allows users to see, measure, convert, playback, and analyze data in real time. Using a variant of the Tompkins QRS detector[1], this program also includes added rules that enable it to automatically extract the fiducial features of an ECG signal on a cycle-by-cycle basis: onset of P, P peak, end of P, onset of QRS, peak of QRS, end of QRS, onset of T, peak of T, and end of T. In this work, fiducial features from ECG recordings are extracted to identify heart diseases. In order to extract features, only one channel was selected from the recordings. All electrocardiogram (ECG) streams were sampled at 200 Hz to match the optimized sampling rate of QRS

detection algorithms [64] and all premature, missing, or ectopic beats are filtered out. By comparing its results with manual classification, the system is adjusted to human ECGs.

3.4.1 ECG Fiducial Features

ECG studies have reported quantitative ECG readings in clinical applications to investigate the relationship between cardiac, neurological, and functional outcomes of ischemic stroke and researchers employed ECG fiducial characteristics to assess different cardiac abnormalities [65]. The ECG signal is composed of multiple segments, including the P-wave, QRS complex, and T-wave, as well as the accompanying onset, offset, and peak points known as fiducial points. A number of reference features have been obtained from each electrocardiogram (ECG) record as illustrated in Figure 2.5. The features include the measurement of time and voltage for P-wave, Q-wave, and S-wave events as well as the extraction of QRS events from various points and intervals within the waveform of the ECG signal cycle. The various ECG reference features with equations are described below:

- Heart rate is measured using the RR interval, which is expressed in heart rate per minute (BPM). It can be calculated as:

$$\text{HR}_i = \frac{60}{RR_i} \tag{3.1}$$

- R height (R-H): The amplitude of the R wave in the electrocardiogram (ECG) cycle is referred to as the R-height (R-H) and is measured in millivolts (mV).

- PH height (P-H): The amplitude of a single-cycle P-wave peak measured in mV.

- PR interval: The interval of time between the onset of the P wave and the start of the QRS complex, measured in seconds. Let P_i represents the sample index of the i-th P wave and QRS_{start_i} represents the sample index wave where the i-th QRS complex starts. The PR interval between P_i and QRS_{start_i} can be calculated as:

$$\text{PR}_i = QRS_{start_i} - P_i \tag{3.2}$$

- RR interval: The time between consecutive R peaks of the ECG waveform, measured in seconds. Let R_i represents the sample index of the i-th R wave. The RR

interval between R_i and R_{i+1} can be calculated as:

$$\text{RR}_i = R_{i+1} - R_i \tag{3.3}$$

- QRS complex: The period from the start of the Q wave to the end of the S wave. This includes the R wave. Let QRS_{start_i} represents the sample index where the i-th QRS complex starts and QRS_{end_i} represents the sample index where the i-th QRS complex ends. The QRS duration for the i-th complex can be calculated as:

$$\text{QRS}_{duration_i} = QRS_{end_i} - QRS_{start_i} \tag{3.4}$$

- ST segment: The time between the end of the QRS complex and the start of the T wave, measured in seconds. Let T_{start_i} represents the sample index where the i-th T wave starts and QRS_{end_i} represents the sample index where the i-th QRS complex ends. The ST segment duration for the i-th complex can be calculated as:

$$\text{ST}_i = T_{start_i} - QRS_{end_i} \tag{3.5}$$

- QT interval: The time from the start of the QRS complex to the end of the T wave, measured in seconds. Let QRS_{start_i} represents the sample index where the i-th QRS complex starts and T_{end_i} represents the sample index where the i-th T wave ends. The QT interval for the i-th complex can be calculated as:

$$\text{QT}_i = T_{end_i} - QRS_{start_i} \tag{3.6}$$

- Corrected QT interval (QTc): Modified QT interval period adjusted for the RR interval.

Stroke patients are often diagnosed with cardiac abnormalities such as myocardial ischemia. Fiducial abnormalities on the ECG are often identified as ST-segment depression, an extended QT interval, flat or inverted T-waves, and U-waves [12]. The intensity and duration of myocardial ischemia episodes can be obtained from ECG signal. These episodes are represented by variations in the ST segment, which can be either elevations or depressions. Acute myocardial ischemia can be identified by examining the ST segment of each cardiac cycle on an ECG [66]. Pueyo et al. [67] suggested that the QRS slope can be an additional source of information for the diagnosis and characterization of ischemia along with the ST segment.

3.5 Data Augmentation

The total number of fiducial features obtained for each category is as follows: 581,219 for ischemic cases, 267,360 for healthy cases, and 40,238 for arrhythmic cases. So, data augmentation methods were employed to expand this imbalanced dataset and enhance the model's robustness to different types of noise. To make the model resistant to post estimation inaccuracies, SMOTE technique [62] is used. Through the generation of synthetic examples for the minority class, this data augmentation technique helps to lessen the effects of class imbalance and improves models' dependability and performance. These augmentations were applied exclusively to the training set and executed dynamically during runtime.

3.6 Federated Learning System

Federated learning is a collaborative method in which several decentralized devices, or nodes, collectively train a shared model. In this setup, a central server initializes a global model and sends it to multiple client devices. Each client subsequently conducts local training with its own data, ensuring data privacy by retaining the raw data on the device. Once local training is complete, the clients transmit only the model updates such as gradients or parameter adjustments back to the central server. The server subsequently combines these updates, usually by averaging them, to produce an enhanced version of the global model. This updated global model is sent back to the clients, and the process is repeated iteratively as depicted in Figure 3.2. This method enables the global model to learn from distributed data while preserving data privacy and security, as the raw data stays on the individual devices.

3.6.1 Flower for Federated Learning System

Federated Learning (FL) is made easier by the Flower framework [68], which allows several clients (servers or devices) to operate together to train a model without exchanging data. The flower framework is designed to be flexible and extensible, enabling it to handle a large number of clients, making it scalable for real-world applications. It allows for custom implementations of key components such as client selection and model aggregation, enabling users to tailor the FL process to their specific needs. By keeping data on local devices and reducing the need for centralized data storage, the Flower framework provides an efficient and collaborative solution for decentralized model training, ensuring data privacy and operational efficiency [68].

Figure 3.2: Federated Learning Architecture.(Courtesy of [7])

3.6.2 Federated Averaging Algorithm(FedAvg)

According to [59], FedAvg is an algorithm used in federated learning to compute a weighted average of model updates from clients on a central server which enables local nodes to carry out multiple mini-batch updates on their own data before sharing the updated model parameters. Algorithm 1 orchestrates the training process via a central server that maintains the global model w_t, where t indicates the communication round.

Algorithm 1 Federated Averaging Algorithm

Initialize X_0
for each cycle $t = 0, 1, \ldots$ **do**
  $m \leftarrow \max(\lfloor C \cdot K \rfloor, 1)$
  $W_t =$ set of m clients
  for each client $k \in W_t$ **in parallel do**
    $X_{k,t+1} = \text{UpdateClient}(k, X_t)$
  $X_{t+1} = \sum_{k \in W_t} \frac{n_k}{n_\sigma} X_{k,t+1}, \quad n_\sigma = \sum_{k \in W_t} n_k$

3.7 Proposed Federated System for Client side and Server side

Federated learning model is used to take a weighted average of the updated parameters from the clients on the server. On the client side, it enables local nodes to perform multiple batch updates on their local data and exchange the updated weights. Then it coordinates training by using a central server to host the shared global model.

3.7.1 Client-Side Pseudocode

The client-side pseudocode describes the local training process on each client. Federated learning allows clients to train a model locally using their private datasets without sharing the raw data with a central server, thus preserving privacy. As shown in algorithm 2, each client follows these steps:

- **Input Parameters:** Each client i begins with its own dataset D^i, a predefined batch size B, local epochs E, learning rate η, and model parameters θ^i, which are randomly initialized at the start.
- **Epoch Loop:** The training process in each client continues for E local epochs. In each epoch, the client divides its dataset D^i into mini-batches of size B, which are used in the training process.
- **Mini-Batch Training:** For each mini-batch b, the client computes a forward pass, denoted as $h^i = f(x^i, \theta^i)$, where x^i represents the input data and f is the model function parameterized by θ^i. Then, the loss $\mathcal{L}(y^i, \hat{y}^i)$ is computed between the predicted values $\hat{y}^i$ and the actual ground truth y^i, from which the gradients of the loss with respect to the model parameters are calculated as $\Delta\theta^i = \frac{\partial \mathcal{L}(y^i, \hat{y}^i)}{\partial \theta^i}$.
- **Model Update:** After computing the gradients, the client updates its local model parameters using stochastic gradient descent (SGD). The updated model parameters are obtained using the rule $\theta^i \leftarrow \theta^i - \eta \cdot \Delta\theta^i$, where η is the learning rate that controls how much the model parameters should change with each gradient step.
- **Sending Updated Parameters:** Once all local epochs are completed, the client sends the updated model parameters θ^i to the central server for aggregation.

Algorithm 2 Client Pseudocode

Require: Client dataset D^i, Batch size B, Local epochs E, Learning rate η, Model parameters θ^i
Ensure: Updated local model θ^i
Initialize: Random weights $\theta^i = \text{rand}()$
for epoch $e = 1$ to E **do**
    Divide D^i into mini-batches of size B
    for each mini-batch b **do**
        Compute forward pass $h^i = f(x^i, \theta^i)$
        Compute gradients $\Delta\theta^i = \frac{\partial \mathcal{L}(y^i, \hat{y}^i)}{\partial \theta^i}$
        Update local model $\theta^i \leftarrow \theta^i - \eta \cdot \Delta\theta^i$
    end for
end for
Send updated model parameters θ^i to the server

3.7.2 Server-Side Pseudocode

The server plays a crucial role in aggregating the model updates from the clients and updating the global model, which is subsequently shared with the clients in the next training round. The server-side algorithm 3 operates as follows:

- **Input Parameters:** The server receives local model parameters $\theta^1, \theta^2, \ldots, \theta^N$ from N clients and uses an aggregation function, such as Federated Averaging (FedAvg), to combine the updates.
- **Global Model Initialization:** At the start, the server initializes a global model θ^g with random weights.
- **Aggregation Loop:** For each round $r = 1$ to R, the server collects the local model parameters θ^i from each client. It then aggregates these parameters to update the global model using the FedAvg algorithm. The global model parameters θ^g are updated as the average of the client models:

$$\theta^g = \frac{1}{N} \sum_{i=1}^{N} \theta^i \tag{3.7}$$

  where N is the number of participating clients.
- **Global Model Deployment:** After each aggregation, the server deploys the updated global model to the clients for the subsequent round of local training.

- **Model Interpretability:** Once the final global model is obtained, model interpretability techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are applied. SHAP values are computed to quantify the contribution of each feature x_j to the model's prediction, represented as:

$$\phi_j(x) = \frac{\partial f(x)}{\partial x_j} \tag{3.8}$$

These values help explain the importance of individual features in the model's decision-making process. LIME is also used for instance-wise explanations by approximating the global model $f(x)$ locally for specific data points, providing interpretable insights into how the model behaves on specific instances. This allows us to analyze and interpret model predictions with a focus on transparency and explainability.

Algorithm 3 Server Pseudocode

Require: Model parameters from clients $\theta^1, \theta^2, \ldots, \theta^N$, Aggregation function FedAvg
Ensure: Global model parameters θ^g
Initialize: Global model $\theta^g = \text{rand}()$
for each round $r = 1$ to R **do**
    for each client i **do**
        Receive local model θ^i
    end for
    Aggregate local models $\theta^g = \frac{1}{N} \sum_{i=1}^{N} \theta^i$
end for
Deploy the global model θ^g
Apply SHAP and LIME on the final global model θ^g
Compute SHAP values $\phi_j(x) = \frac{\partial f(x)}{\partial x_j}$ for feature attribution
Use LIME for instance-wise explanations $E(x) \approx f(x)$
Analyze and interpret model predictions using SHAP and LIME explanations

3.8 Using Deep Learning Models in Client Side for Heart Disease Classification

Deep learning, a subfield of machine learning, has gained significant attention in recent years, largely due to the success of deep neural networks. We selected two models ANN and LSTM, where LSTM is one of the most high-achieving models of the PhysioNet 2020 competition [69]. Although transformers and convolutional neural networks (CNNs) are frequently utilized in the ECG-based identification of cardiac diseases, it is important to emphasize that they are not taken into account in this work. This is because their primary purpose is to handle signals and image-like datasets while our work is shown using tabular data. In order to examine the limited impact of hyperparameter adjustment, we chose to examine two comparatively straightforward models to build our explainable federated learning model.

3.8.1 ANN Model for Federated Learning

One popular method for detecting ECG anomalies is the Artificial Neural Network (ANN), which is chosen because of its ability to develop hierarchical representations of data through layers of interconnected neurons. Artificial Neural Networks (ANNs) are computer models inspired by the structure and function of the human brain. They are made up of layers of interconnected nodes, referred to as neurons. Typically, these layers include an input layer, one or more hidden layers, and an output layer. ANNs are trained using a process called backpropagation, where the network minimizes the difference between expected and actual outputs by learning from the data and adjusting the weights of the connections between neurons. The input data is passed forward through the network, the output is computed, compared to the actual output, and the error is then propagated backward through the network to update the weights.

3.8.2 ANN Model Architecture

The following section analyzes the structure and functionality of an Artificial Neural Network (ANN) model used in the context of federated learning. The model consists of a series of layers, including input, dense, dropout, and output layers as shown in Figure 3.3. Each layer plays a significant role in determining how the model learns from data across different clients in the federated learning framework. The ANN model can be described as a basic feedforward network composed of three main layers :

- **Input Layer:** The input layer processes the data from the client's local dataset. In this case, it flattens the input into a 1-dimensional vector of size 9 (represented

as `Flatten [9]`). This step is necessary for preparing the input data for the dense layers that follow. The flattened vector could represent feature extraction from the client's local data (e.g., sensor readings, time-series data, or other structured data).

- **Dense Layer 1:** The first hidden layer is a fully connected (dense) layer consisting of 128 units with a ReLU activation function. ReLU adds non-linearity to the model, allowing it to capture and learn more complex patterns. The dense layer applies weights to the flattened input and computes the outputs, which are propagated through the network. The presence of 128 units provides significant capacity for the model to capture intricate patterns in the data, making it suitable for complex tasks in federated learning.

- **Dropout Layer:** A dropout layer with a dropout rate of 0.2 is applied after the first dense layer to reduce the risk of overfitting. Dropout works by randomly setting a fraction (20% in this case) of the neurons to zero during each forward pass of training. This regularization technique is especially important in federated learning, where each client's local data might be limited or exhibit high variability. By using dropout, the model generalizes better across different clients and reduces overfitting to individual client datasets.

- **Output Layer:** The final layer is a dense output layer with 3 units and a softmax activation function. This indicates that the model performs classification into three categories (e.g., `arrhythmic`, `healthy`, `ischemic` in a medical context). The softmax function outputs a probability distribution over the three classes, ensuring that the predicted class corresponds to the highest probability.

3.8.3 LSTM Model for Federated Learning

Secondly, the Long Short-Term Memory (LSTM) methodology [70], is chosen because this model can effectively model the dynamic nature of heart rhythm variations over time, allowing accurate detection and classification of arrhythmias [71]. While the LSTM cell and the output layer are made up of the same number of neurons as the number of classes to predict, the input layer employs as many neurons as the number of variables to predict. After the fine-tuning procedure, the values of the activation function of the output layer emerged; the ReLu and SoftMax functions performed the best. Unlike conventional feedforward neural networks, LSTMs can benefit from temporal dependency across data sequences since they have feedback connections. Gradients may vanish or explode when training traditional RNNs on data sequences.

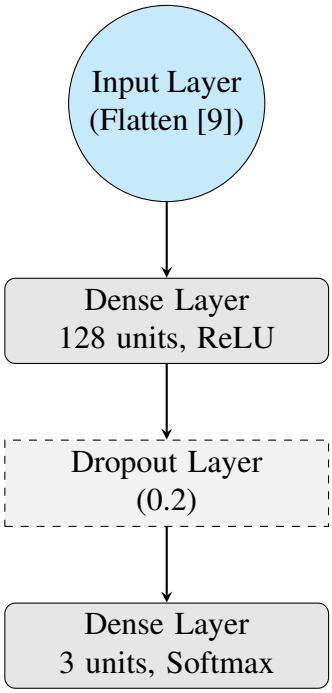


Figure 3.3: ANN Model Architecture

LSTM is designed to tackle just this kind of issue. LSTM networks add memory cells, which have the capacity to store information across lengthy sequences. Every memory cell consists of three main components: an input gate, an output gate, and a forget gate. These gates make it easier to regulate the internal flow of information.

3.8.4 LSTM Model Architecture

This section provides an analysis of a Long Short-Term Memory (LSTM) model designed for federated learning. The model consists of several LSTM layers, dropout layers for regularization, and a dense output layer for classification as shown in Figure 3.4. This architecture is suitable for time-series data or sequential data processing tasks, which are common in federated learning environments.The model is structured as follows:

- **Input Layer:** The input layer takes data in the shape of `(9, 1)`, representing 9 time steps with 1 feature per time step. This input format is typical for sequential data such as sensor readings or biomedical time-series data.
- **LSTM Layer 1:** The first LSTM layer has 64 units with a ReLU activation function and is configured to return sequences. Returning sequences is important because it allows the output of this LSTM layer to be passed to the next LSTM layer while preserving the sequence structure. This layer is crucial for learning

temporal dependencies in the data, as LSTM units are specifically designed to capture long-term dependencies in time-series data.

- **Dropout Layer 1:** After the first LSTM layer, a dropout layer with a rate of 0.2 is applied to prevent overfitting. In federated learning, each client's local data might be limited or highly variable, so dropout ensures the model generalizes well across different clients.

- **LSTM Layer 2:** The second LSTM layer has 32 units, also using the ReLU activation function and returning sequences. Reducing the number of units in this layer helps in gradually compressing the learned representations while still maintaining the temporal relationships.

- **Dropout Layer 2:** A second dropout layer with the same rate (0.2) follows the second LSTM layer, further ensuring that the model does not overfit, especially on smaller client datasets.

- **LSTM Layer 3:** The third LSTM layer has 16 units with a ReLU activation function, but it does not return sequences. Instead, it outputs a single vector that summarizes the learned features from the previous LSTM layers. This layer acts as a final compression of the temporal features before they are passed to the output layer.

- **Dropout Layer 3:** Another dropout layer with a rate of 0.2 is applied after the third LSTM layer to maintain regularization.

- **Output Layer:** The final layer consists of a dense layer with 3 units and a softmax activation function, signifying that the model is designed for multi-class classification. The output provides a probability distribution across 3 classes (e.g., `arrhythmic`, `healthy`, `ischemic` in a medical context), with the class having the highest probability being chosen as the predicted output.

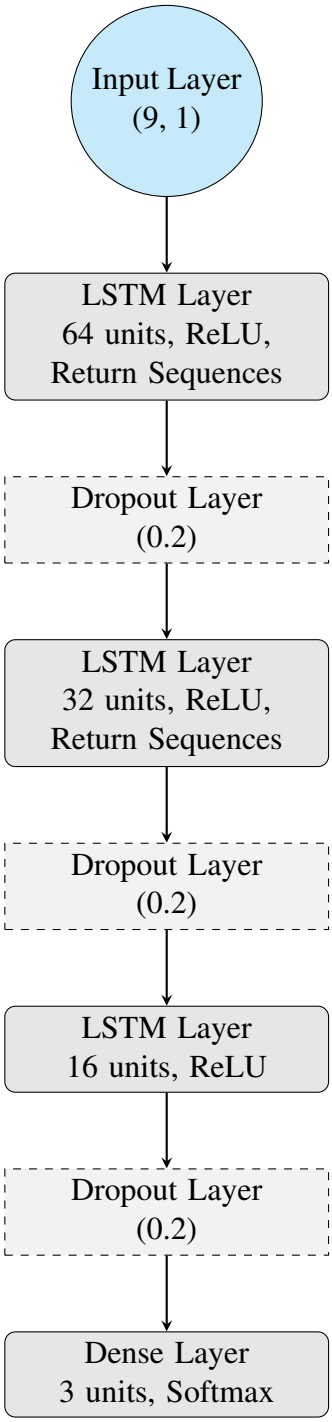


Figure 3.4: LSTM Model Architecture

3.9 Explainability with XAI Technique

In recent years, significant progress has been made in the field of explainable artificial intelligence (XAI), with a focus on understanding the reasoning behind the decisions made by machine learning models. Various XAI techniques have been used to explain the output and behavior of machine learning models [55]. They may be classified as

globally or locally interpretable based on whether they explain the whole model or individual input samples. In this study, Shapley Additive Explanation(SHAP) method was employed to interpret the entire model after the ML model's training. Furthermore, Local interpretable model-agnostic explanations (LIME) were utilized to provide an explanation for a particular input sample. Following subsections briefly discusses SHAP and LIME techniques.

3.9.1 Shapley Additive Explanations (SHAP)

Lundberg and Lee introduced the SHapley Additive exPlanations (SHAP) method in 2016 to explain the significance of individual features on the classification or prediction outcomes of a black box machine learning model. SHAP is derived from the Shapley values, which are the optimum values in game theory. There are other techniques available for calculating Shapley values, such as linear SHAP, kernel SHAP, and Deep SHAP [21]. The linear SHAP method provides an explanation for the feature significance in linear machine learning models. Depending on the complexity of the ML model, Shapley values can be produced to determine the feature importance in the model output prediction or classification. There are various methods for deriving Shapley values, including linear SHAP, kernel SHAP, and Deep SHAP [21].

The feature importance in linear ML models is explained by the linear SHAP. Given that S is a subset of all features and that S_F is defined as $\{X_1, X_2, \ldots, X_k, \ldots, X_M\}$ where X_k denotes the characteristics of the dataset at the k-th column in a dataset of size $N \times M$. There are two methods used to incorporate feature X_i into a model f's output. The model is first trained while feature X_i is present; the resultant model is represented as $f_{S\cup\{i\}}$. Next, the model is retrained without feature X_i; the resulting model is then represented as f_S. Then, the Shapley value, i, for the feature X_i is determined using Equation (1) [21]:

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}\left(f_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}}) - f_S(\mathbf{x}_S)\right) \tag{3.9}$$

where $f_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}})$ signifies the marginal value of f for the feature values present in S plus feature X_i, and $f_S(\mathbf{x}_S)$ denotes the marginal value of f for the feature values present in S.

Equation (1) calculates the disparity over all feasible subsets $S \subseteq F \setminus \{i\}$, where S is the number of features from the total number of features, F, and $|F|$ is the number of features in S.

3.9.2 Local Interpretable Model-Agnostic Explanations (LIME)

LIME was first introduced by Ribeiro et al. [22]. To identify which features contribute most significantly to the output of a black-box machine learning model, LIME uses a locally interpretable surrogate model to approximate the complex, non-linear model. This approach is based on the assumption that complex models behave in a linear manner on a local scale. Therefore, it becomes feasible to approximate the behavior of the complex model in the vicinity of specific instances that require explanation.

The penalty function $x(z)$, which calculates the proximity of perturbed examples $z \in \mathbb{R}$ around an instance feature vector, x, assesses the neighborhood significance and as a result, g, a surrogate model, best approximates f among a class of potential interpretable models G, i.e., $g \in G$, given f, a black box ML model to be explained [22].The explanation $\xi(x)$ for an instance feature vector x produced by LIME is obtained by minimizing the objective function $L(f, g, x) + \Omega(g)$, as given in Equation 3.10:

$$\xi(x) = \arg\max(L(f, g, x) + \Omega(g)) \quad [g \in G] \tag{3.10}$$

where L is a locality-aware loss function for measuring how g is unfaithful in closely resembling f in the locality defined by x, and $\Omega(g)$ is a measure of g's complexity. In LIME, the original feature space of the dataset, $x \in X$, is sampled to create a set of d interpretable representation features, $x' \in \{0, 1\}^{b'}$, respectively [22]. A label for the explanation model, $f(z)$, is produced using a binary vector representation of perturbed instances z' around non-zero parts of x'. The mapping of the binary vector representation of features to the original real-valued representation is performed via a mapping function h_x, such that $h_x : z' \rightarrow z$, i.e., $z = h_x(z')$. Thus, using this dataset, Z, of perturbed samples with their labels, i.e., $\{(z', f(z))\}$, the locality-aware loss function is defined as Equation 3.11:

$$L(f, g, x) = \sum_{z,z' \in Z} x(z)(f(z) - g(z'))^2 \tag{3.11}$$

A small amount of literature has attempted to demonstrate the use of LIME in comparing ECG signal-based ML model outputs for heart disease categorization [56, 57]. Although it relies on the complexity of the local surrogate models, LIME offers a straightforward explanation. The local surrogate models interpret the data using features sampled from the source dataset. This procedure increases the value of LIME methods, particularly when sophisticated features are used to train the black box ML model. The LIME's feature importance ratings do not, however, add up to produce the ambiguous prediction probability. Furthermore, they do not offer a comprehensive

justification of the sophisticated ML model that has been learned across the complete range of feature values.

Additionally, the LIME methodologies were hampered by the instabilities that make it difficult to replicate the explanations due to the random perturbations of feature instances. Additionally, biases can be concealed by manipulating LIME [72]. To counteract this instability and the ensuing LIME unfaithfulness, various strategies have been put forth in the literature [73–76].

3.10 Evaluation Metrics

3.10.1 Accuracy

The ratio of the overall number of accurate predictions to the total number of predictions is known as accuracy. The accuracy is calculated using the test dataset, which remains inaccessible to the model throughout the training process, in order to provide a more objective and reliable estimate of the model's generalization capability.

$$\text{Accuracy} = \frac{A}{B} \times 100\% \tag{3.12}$$

Here, B is the number of samples in the test set and A is the number of samples for which the class labels were correctly predicted by the model.

3.10.2 Precision

Precision is formally defined as the ratio of the total number of true positive predictions to the sum of true positive and false positive predictions across all classes. In the context of a multiclass classification problem, precision evaluates the proportion of correctly identified samples among all those predicted to belong to a particular class. This metric is also referred to as the Positive Predictive Value (PPV).

Precision for each class can be calculated considering the one-vs-all strategy.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.13}$$

Here, TP is the number of correctly classified samples, and FP is the number of wrongly classified samples.

3.10.3 Recall

Recall is defined as the ratio of the total number of true positive predictions to the sum of true positive and false negative predictions across all classes. In multiclass classification tasks, recall quantifies the proportion of actual class instances that were correctly identified by the model. This metric is also commonly referred to as sensitivity.

Recall for each class can be calculated considering the one-vs-all strategy.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.14}$$

Here, TP is the number of correctly classified samples, and FN is the number of wrongly classified samples.

3.10.4 F1-Score

The F1-Score is the harmonic mean of precision and recall, incorporating both false positive and false negative predictions into its calculation. In scenarios involving imbalanced datasets, a high F1-Score is particularly important, as it reflects the model's effectiveness in minimizing both false positives and false negatives.

F1-Score for each class can be calculated using the following formula:

$$\text{F1-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.15}$$

3.10.5 Loss

Loss measures the error of a model's prediction compared to the actual value. It varies depending on the task and chosen loss function.

$$\text{Loss} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{3.16}$$

Here, y_i is the true label (usually 0 or 1 for binary classification), $\hat{y}_i$ is the predicted probability for class 1 and N is the number of samples.

3.10.6 Misclassification Rate

The misclassification rate is the proportion of incorrect predictions.

$$\text{Misclassification Rate} = \frac{\text{Number of Incorrect Predictions}}{\text{Total Number of Predictions}} = \frac{FP + FN}{TP + TN + FP + FN} \tag{3.17}$$

3.10.7 False Positive Rate(FPR)

FPR is the proportion of actual negatives that were incorrectly predicted as positive.

$$\text{FPR} = \frac{\text{False Positives}}{\text{True Negatives} + \text{False Positives}} = \frac{FP}{TN + FP} \tag{3.18}$$

3.10.8 False Negative Rate(FNR)

FNR is the proportion of actual positives that were incorrectly predicted as negative.

$$\text{FNR} = \frac{\text{False Negatives}}{\text{True Positives} + \text{False Negatives}} = \frac{FN}{TP + FN} \tag{3.19}$$

3.10.9 AUC-ROC Score

Each of the widely used metrics, including precision and recall, has its own set of limitations. Precision is a metric that assesses the accuracy of a classification task and is solely dependent on true positive and false positive predictions; one true positive prediction can yield a precision score of 1.0. Conversely, recall relies exclusively on genuine positive and false negative responses and is all about completeness. Therefore, if all the samples are predicted to be positive, the recall will be 1.0, but the precision will be quite poor. In this context, the True Positive Rate (TPR) and False Positive Rate (FPR) are utilized to construct the Receiver Operating Characteristic (ROC) curve, with the area under the ROC curve (AUC-ROC) serving as an evaluation metric. These techniques enable models to be assessed based on how effectively they distinguish between classes. To evaluate the model's performance and its ability to differentiate between classes, the False Positive Rate (FPR) and True Positive Rate (TPR) are calculated for a series of predictions made by the model at varying thresholds. In the ROC curve, each probability threshold is represented as a point, and these points are connected to form the curve. A diagonal line connecting FPR 0 and TPR

0 (coordinates: 0,0) to FPR 1 and TPR 1 represents a model without discriminating power. (coordinate: 1,1).

Chapter 4

Results and Discussions

Federated learning-based architectures have achieved state-of-the-art performance in different heart disease classification tasks. However, this high performance poses the requirement of a huge amount of training data, which is a very difficult constraint to satisfy. With the goal of alleviating this limitation, we have proposed an explainable federated learning model by merging three heart disease datasets that can produce high performance. In this chapter, the results and outcome of the proposed models are discussed.

4.1 Experimental Setup

In this section, we discuss the experimental environment and different hyperparameters related to the training and testing process of our heart disease classification pipeline.

4.1.1 Environment

The proposed system was implemented in Google Colab using the PyTorch framework. The environment provides a NVIDIA Tesla T4 GPU with 15 GB GPU RAM and CUDA version 12.2. An Intel Xeon CPU with 2 cores, each having a base clock speed of 2.30 GHz. The total usable memory of the system was 13 GB.

4.1.2 Dataset Split

To determine the efficacy of our models, we need a unbiased measurement approach. To do this, we divided the data into two categories: training data and test data. For our merged dataset, we allocated 67% of the data for training and 33% of the data for testing. We trained at least 10 clients in each round. Per client got maximum 10% of training data or 6.7% of the whole dataset.

4.1.3 Batch Size

A batch size of 32 was used, adhering to the mini-batch gradient descent approach. Although this batch size is relatively small compared to the total number of frames in the training set, the training samples were efficiently accommodated in memory without any issues.

4.1.4 Epoch and Learning Rate

The system was trained in 15 cycles, each consisting of 30 clients. The learning rate (LR) was set at 0.001 at the beginning to ensure divergence from the local minima.

4.1.5 Optimizer and Loss Function

The Adam optimizer is employed to mitigate the impact of noise resulting from the estimation of heart diseases [77]. Additionally, it is ideal for our scenario, as it can manage large datasets. In this case, a sparse categorical cross-entropy loss function is used, which is an extension of the categorical cross-entropy loss and is applicable when there are two or more label classes.

4.1.6 Data Validation

Data validation is a method used to assess a model's performance while ensuring it generalizes effectively across different data subsets. In this study, 10-fold cross-validation is employed, where the dataset is divided into 10 equal subsets, or folds. The model is trained 10 times, each time using 9 folds for training and the remaining fold for testing. The performance metrics (e.g., accuracy, precision, recall, etc.) are then averaged across the 10 folds to provide a final evaluation score. This technique helps in assessing how well the model performs on different parts of the data, reducing the risk of overfitting.

In a federated setting, applying 10-fold cross-validation requires partitioning the data not only across clients but also across the different folds. Each participating client (device) has its local dataset, and the global dataset is divided into 10 subsets (folds). These subsets should ideally be distributed across clients, ensuring that each fold has data from different clients. For each fold, the model is trained on the data from 9 folds (across different clients) and tested on the data from the 1 remaining fold. Since federated learning operates by aggregating local model updates rather than sharing raw data, each client's model will train on its local data, and model updates will be sent to a central server for aggregation. After each round of training, the central server

aggregates the updates and computes a global model. The global model is evaluated based on the fold that is being used for testing in that iteration.

4.2 Performance Evaluation and Comparison of FL-ANN and FL-LSTM Models

Table 4.1 and Table 4.2 provide a detailed analysis of the performance of two federated learning models: FL-ANN and FL-LSTM models repectively. The comparison is made based on several evaluation metrics, including accuracy, AUC, F1 Score, False Negative Rate (FNR), False Positive Rate (FPR), Loss, Misclassification Rate (MC Rate), Precision and Recall, over multiple rounds of training. The results indicate that both models show improvements in performance as the number of training rounds increases. However, the FL-LSTM model consistently outperforms the FL-ANN model on most metrics, especially as training progresses.

- **Accuracy**: The FL-LSTM model exhibits higher accuracy compared to the FL-ANN model on most rounds. For example, in the final round (round 15), the FL-LSTM achieves an accuracy of 90.91%, while the FL-ANN reaches 87.98%. This indicates that the LSTM model is better at learning the underlying patterns in the data.

- **F1 Score**: The F1 score, which balances precision and recall, is consistently higher for the FL-LSTM model throughout training rounds. By Round 15, the FL-LSTM achieves an F1 score of 0.9088, compared to the FL-ANN's 0.8787. The ability of the LSTM model to capture temporal dependencies may contribute to its higher F1 score, especially in tasks where sequence information is critical.

- **False Negative Rate (FNR)**: The FL-LSTM model demonstrates a lower FNR in all rounds, indicating that it makes fewer errors in classifying actual positives. In Round 15, the FNR for the FL-LSTM is 0.0347, whereas the FL-ANN has an FNR of 0.0272. This shows that both models perform well in minimizing false negatives, but the LSTM model has a slightly better balance in precision and recall, leading to a lower FNR in earlier rounds.

- **False Positive Rate (FPR)**: The FL-LSTM model maintains a lower FPR in most rounds. In Round 15, the FPR for the FL-LSTM is 0.0757, while for the FL-ANN it is 0.0834. This difference indicates that the LSTM model is better at avoiding false positive classifications.

- **Loss**: The loss metric steadily decreases for both models as training progresses, with the FL-LSTM showing a faster reduction in loss compared to the FL-ANN. By Round 15, the FL-LSTM has a lower loss (0.2446) compared to the FL-ANN (0.3233). A lower loss indicates that the FL-LSTM model better fits the training data, resulting in improved generalization.

- **Misclassification Rate (MC Rate)**: As expected, the Misclassification Rate is the inverse of accuracy, and thus the FL-LSTM exhibits a lower MC Rate throughout the rounds. By Round 15, the FL-LSTM's MC Rate is 0.0909, whereas the FL-ANN's MC Rate is 0.1202. The lower misclassification rate of the FL-LSTM model highlights its superior performance in correctly predicting the class labels.

- **Precision and Recall**: Both Precision and Recall follow similar trends to the other metrics, with the FL-LSTM model showing better results overall. In Round 15, the FL-LSTM achieves a precision of 0.9101 and a recall of 0.9091, while the FL-ANN achieves a precision of 0.8880 and recall of 0.8798. This further demonstrates that the FL-LSTM model is better at correctly identifying positive instances and maintaining a higher balance between precision and recall.

To sum up, the table carefully assesses the effectiveness of two Heart Disease Classification models: FL-ANN and FL-LSTM. FL-LSTM is the best-performing model which has the greatest scores in accuracy (0.92), recall (0.92), F1 Score (0.92), precision (0.92), and loss (0.2068). This model is the best option for classifying heart diseases because of its remarkable ability to separate positive cases from negatives and to capture all relevant cases while maintaining a high accuracy in recognizing positive instances.

Table 4.1: FL-ANN model

Round	Accuracy	F1 Score	FNR	FPR	Loss	MC Rate	Precision	Recall
0	0.4986	0.4071	0.5057	0.00001	1.5655	0.5014	0.4406	0.4986
1	0.6981	0.6955	0.4691	0.0063	0.6438	0.3019	0.7486	0.6981
2	0.8242	0.8240	0.1281	0.0753	0.4762	0.1758	0.8355	0.8242
3	0.8514	0.8505	0.0612	0.0925	0.4176	0.1486	0.8599	0.8514
4	0.8575	0.8564	0.0497	0.0912	0.3880	0.1425	0.8657	0.8575
5	0.8629	0.8617	0.0467	0.0845	0.3668	0.1371	0.8713	0.8629
6	0.8679	0.8669	0.0493	0.0764	0.3527	0.1321	0.8755	0.8679

7	0.8712	0.8702	0.0627	0.0526	0.3472	0.1288	0.8818	0.8712
8	0.8710	0.8699	0.0480	0.0715	0.3368	0.1290	0.8799	0.8710
9	0.8702	0.8691	0.0338	0.0945	0.3355	0.1298	0.8789	0.8702
10	0.8735	0.8724	0.0326	0.0896	0.3295	0.1265	0.8817	0.8735
11	0.8736	0.8726	0.0274	0.0985	0.3271	0.1264	0.8817	0.8736
12	0.8771	0.8761	0.0377	0.0775	0.3240	0.1229	0.8853	0.8771
13	0.8793	0.8781	0.0343	0.0730	0.3215	0.1207	0.8877	0.8793
14	0.8773	0.8759	0.0347	0.0731	0.3268	0.1227	0.8868	0.8773
15	0.8798	0.8787	0.0272	0.0834	0.3233	0.1202	0.8880	0.8798

Table 4.2: FL-LSTM model

Round	**Accuracy**	**F1 Score**	**FNR**	**FPR**	**Loss**	**MC Rate**	**Precision**	**Recall**
0	0.3336	0.1668	0.0000	1.0000	1.1331	0.6664	0.1113	0.3336
1	0.7133	0.7014	0.1891	0.1100	0.6420	0.2867	0.7792	0.7133
2	0.8444	0.8436	0.0596	0.1157	0.3861	0.1556	0.8566	0.8444
3	0.8650	0.8644	0.0342	0.1299	0.3443	0.1350	0.8700	0.8650
4	0.8738	0.8736	0.0506	0.1024	0.3135	0.1262	0.8784	0.8738
5	0.8939	0.8937	0.0372	0.0932	0.2877	0.1061	0.8959	0.8939
6	0.8990	0.8988	0.0318	0.0939	0.2760	0.1010	0.9007	0.8990
7	0.8887	0.8889	0.0874	0.0580	0.3035	0.1113	0.8910	0.8887
8	0.9007	0.9006	0.0443	0.0823	0.2731	0.0993	0.9024	0.9007
9	0.8974	0.8976	0.0649	0.0763	0.2790	0.1026	0.8991	0.8974
10	0.9015	0.9015	0.0531	0.0707	0.2665	0.0985	0.9034	0.9015
11	0.9033	0.9032	0.0511	0.0675	0.2625	0.0967	0.9036	0.9033
12	0.9039	0.9039	0.0436	0.0769	0.2852	0.0961	0.9052	0.9039
13	0.9011	0.9010	0.0391	0.0878	0.2674	0.0989	0.9027	0.9011
14	0.9023	0.9023	0.0559	0.0615	0.2627	0.0977	0.9040	0.9023
15	0.9091	0.9088	0.0347	0.0757	0.2446	0.0909	0.9101	0.9091

Figure 4.1 provides a visual representation of the performance of the FL-ANN and FL-LSTM models across various metrics, including accuracy, loss, AUC, F1 score,

precision, recall, false negative rate (FNR), false positive rate (FPR), and misclassification rate. The comparison analysis is based on 16 rounds of the federated learning model.

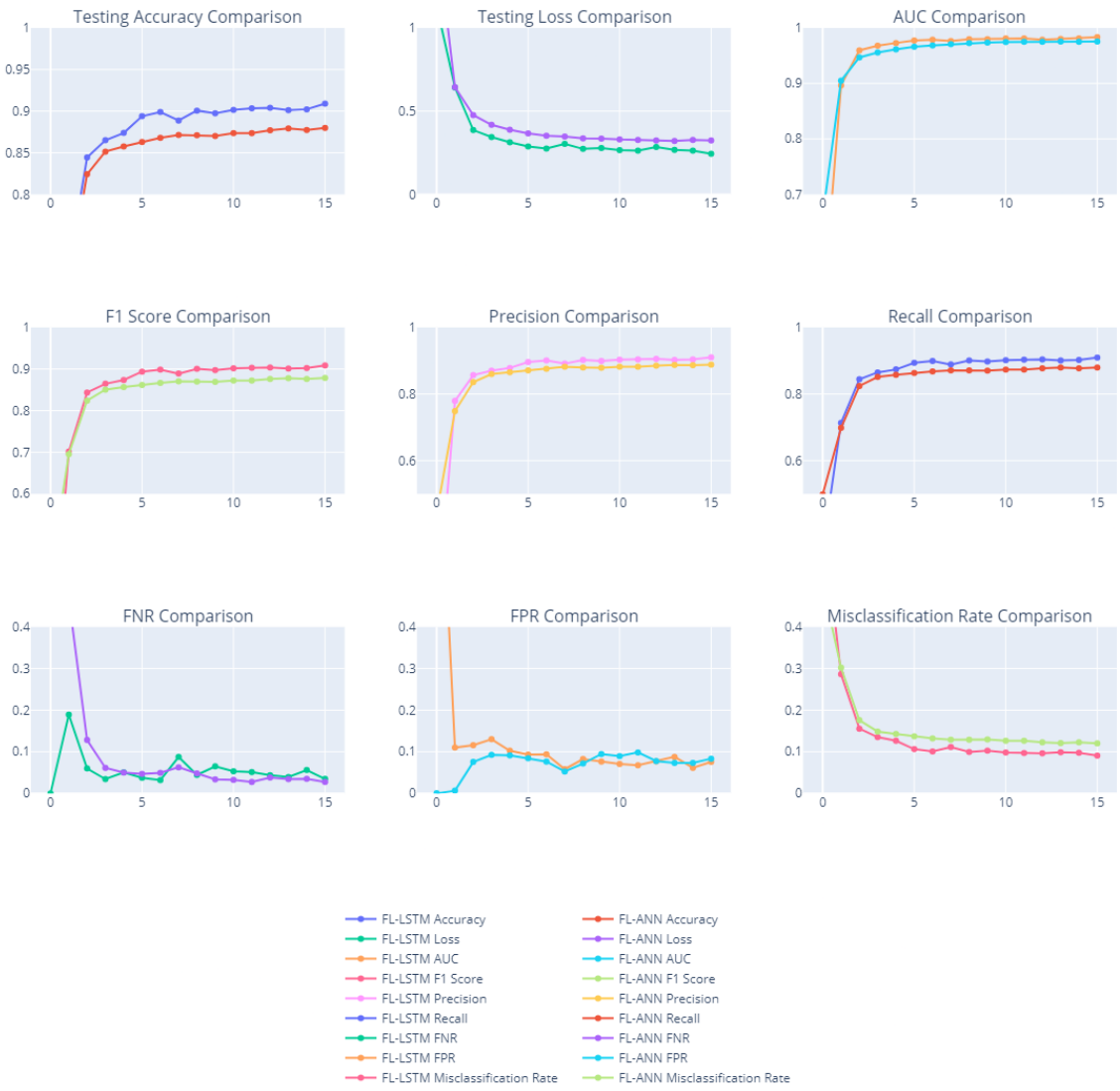


Figure 4.1: Performance comparison and analysis of FL-ANN and FL-LSTM Models in 16 rounds based on evaluation metrics.

4.3 Impact of Number of Clients on the Models

We investigate how the effectiveness of our two models FL-ANN and FL-LSTM are affected by varying number of clients. As shown in Figure 4.2 and Figure 4.3, we take into consideration N = 10, 20, 30, 40, and 50 clients for the simulation. As the number of clients increases, it can be observed in Figure 4.2 and Figure 4.3 that there is a tendency for the framework's performance to decline. This happens as a result

of more diverse data and less successful local models for aggregated clients resulting from larger client counts. Although having more clients may cause the framework to perform worse, this is not our main concern. To simulate a realistic scenario, this work still continues to be effective for a larger number of clients.

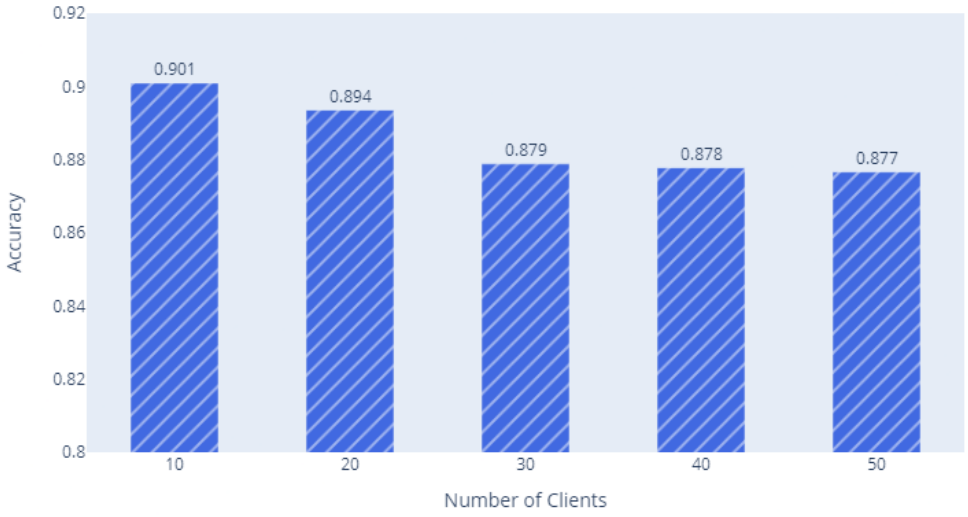


Figure 4.2: FL-ANN clients vs accuracy

4.4 Impact of Communication Rounds on the Models

We assess how the performance of two models, that is, FL-ANN and FL-LSTM are affected by varying the number of communication rounds. We quantify the precision of the framework taking into account T = 5, 15, 25, 35, 45 for the FL-ANN model and T = 10, 20, 30, 40, 50 for FL-LSTM model. As the number of communication rounds rises, overall the framework performs better, as seen in Figure 4.4 and Figure 4.5. The improved aggregation of local models and more precise global models result from the increased communication between the server and the clients. It is seen that the enhancement in performance becomes less significant after about 35 or 40 rounds, suggesting that the framework converges rapidly and more communication provides diminishing rewards. As a result, in practical settings, keeping high model accuracy while lowering communication overhead can be achieved by roughly adjusting the number of communication rounds to 40.

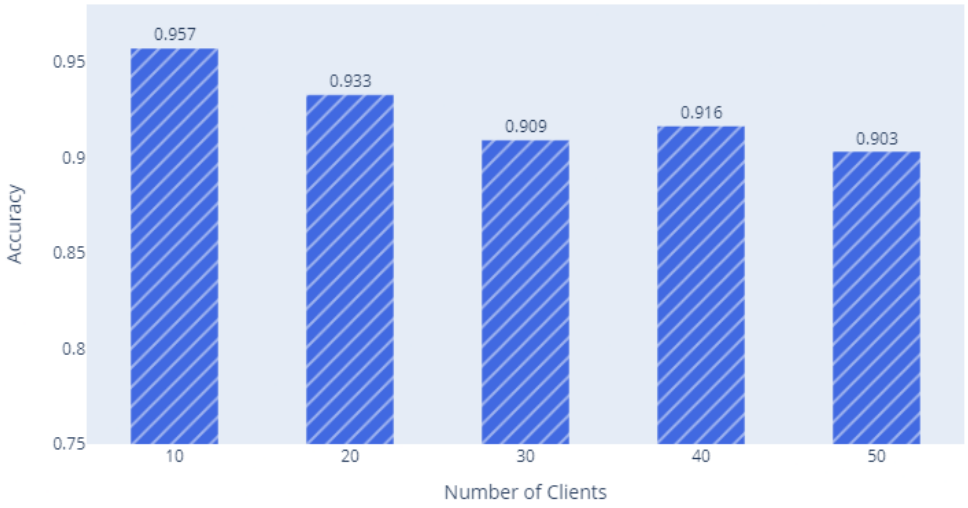


Figure 4.3: FL-LSTM clients vs accuracy

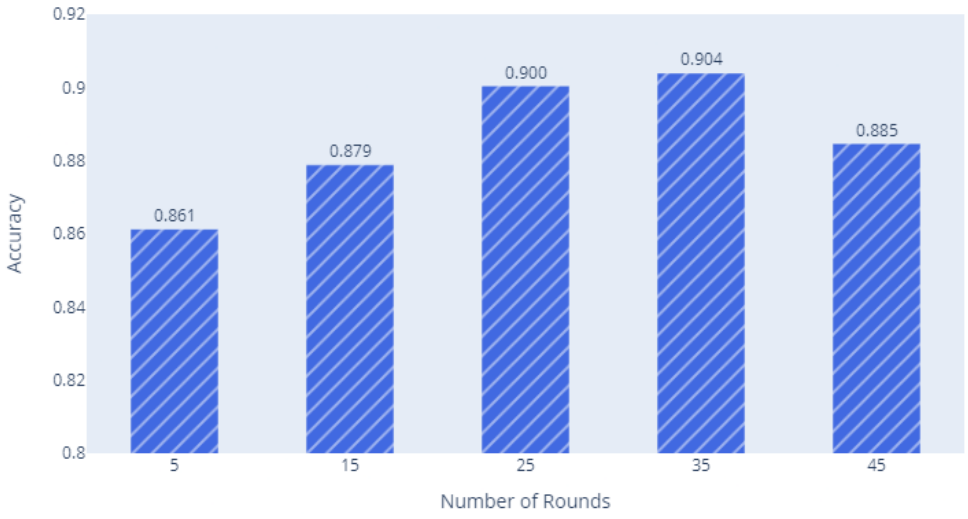


Figure 4.4: FL-ANN rounds vs accuracy

4.5 AUC-ROC curve of the models

In a ROC curve, a higher value on the X-axis indicates an increase in False Positives relative to True Negatives, while a larger value on the Y-axis signifies a higher proportion of True Positives compared to False Negatives. To strike a balance between

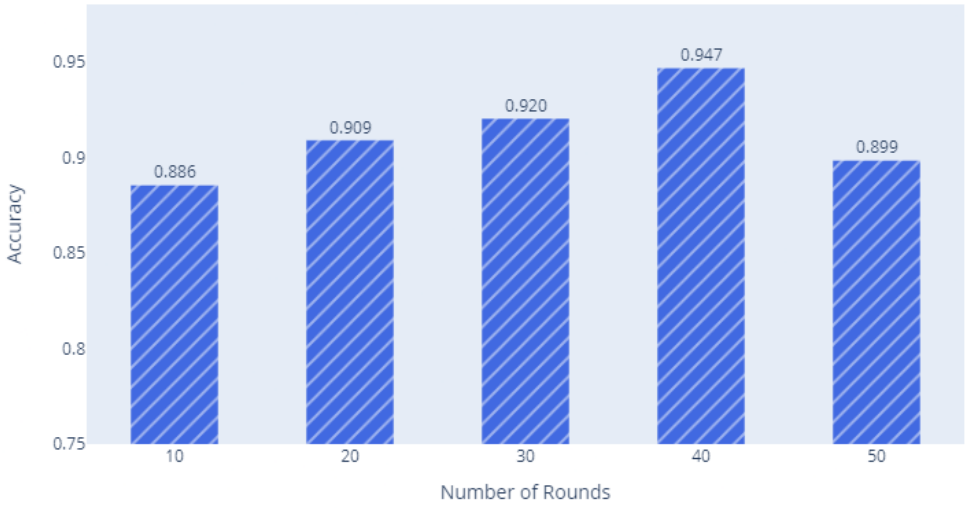
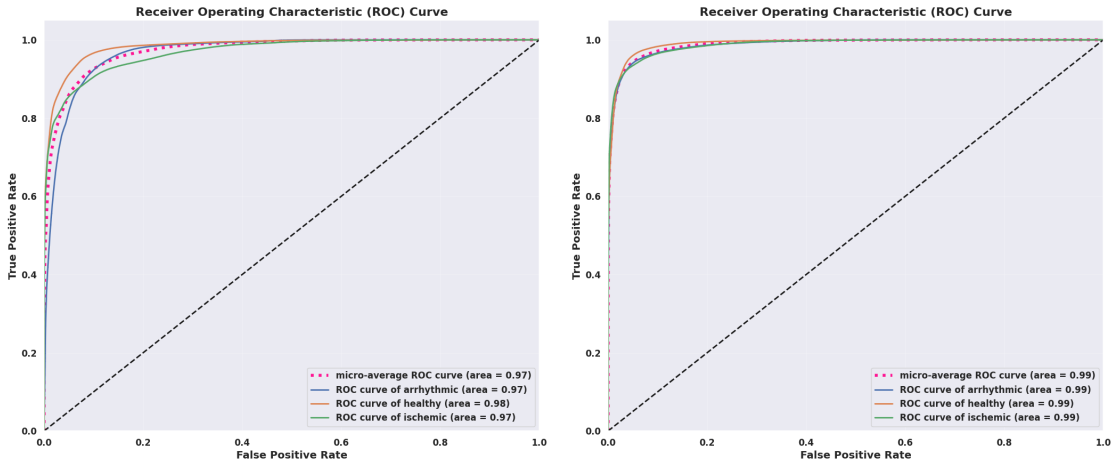


Figure 4.5: FL-LSTM rounds vs accuracy

False Positives and False Negatives, an appropriate threshold must be selected. The model's performance is assessed through the ROC curve and its AUC score. A higher AUC reflects the model's ability to better distinguish between positive and negative classes. An AUC score of 1 indicates perfect differentiation between all positive and negative instances, while an AUC of 0 signifies the model is incorrectly classifying all negatives as positives and vice versa. In Figure 4.6 (a),we can observe that the three classes ; arrhythmic, healthy and ischemic have an AUC scores of 0.97, 0.98, and 0.97 respectively, for FL-ANN models. In Figure 4.6 (b), there is a noticeable change in the FL-LSTM model with an AUC scores of 0.99, 0.99, and 0.99 respectively for the three classes which implies that FL-LSTM model performs better at differentiating between the positive and negative classes than FL-ANN model.



(a) **(b)**

Figure 4.6: Receiver Operating Characteristic (ROC) curves with AUC scores of three classes for (a) FL-ANN Model (b) FL-LSTM model

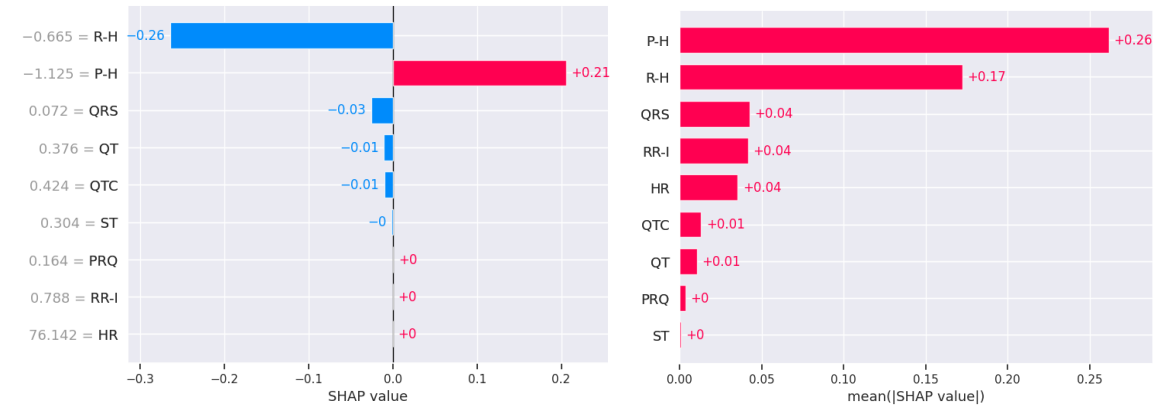
4.6 Model Explanation using SHAP

The main goal of SHAP is to calculate each feature's contribution to the model's output by taking into account all potential features and their corresponding Shapley values. In contrast to certain other methods of explanation, SHAP values consider feature interactions by default, offering a comprehensive view of how each feature influences a given prediction. In practice, SHAP values help users discover potential biases, comprehend model behavior, and improve feature engineering by allowing them to identify the precise characteristics driving a model's choice.

4.6.1 FL-ANN Model Explanation using SHAP

We applied SHAP to the test data set to identify the contribution of each feature to the model output. Figure 4.7 represents two types of shap visualization for the FL-ANN model. In figure 4.7(a), a local plot is displayed, where passing a row of SHAP values to the plot function generates a local feature importance plot. In this plot, the bars represent the SHAP values for each feature, illustrating their contributions to the specific prediction. The horizontal axis reflects the impact of each feature, with positive values driving the prediction higher and negative values pulling it lower, while the features are listed along the vertical axis. "P-H" has a notable positive effect with a SHAP value of +0.21, pushing the prediction higher. Conversely,"R-H" has the most significant negative impact with a SHAP value of -0.26, meaning it strongly pulls the prediction down for that particular feature instance. In figure 4.7(b), a global bar plot is shown, where passing a matrix of SHAP values to the bar plot function creates a global feature

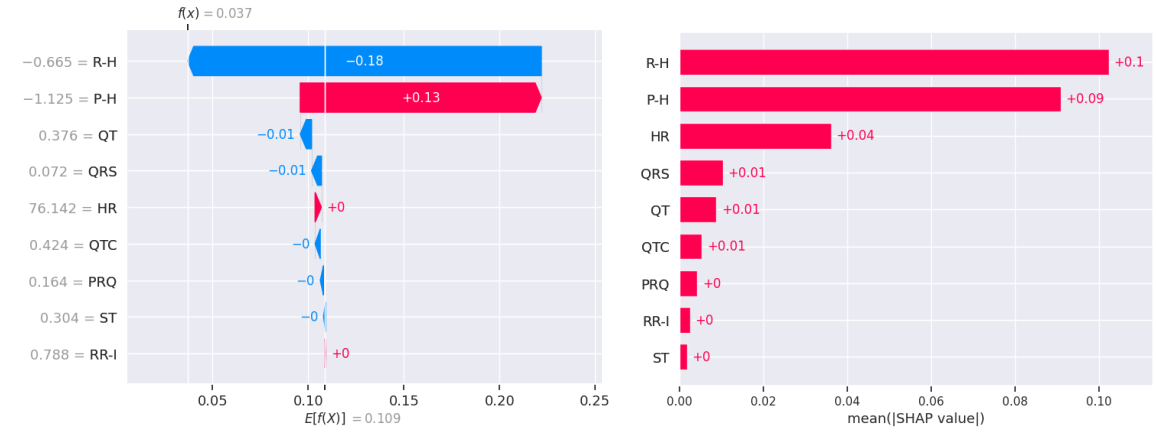
importance plot. In this plot, the global importance of each feature is determined by calculating the mean absolute SHAP value for that feature across all the samples given. Here, it is also observed that the most influential features are "P-H" and "R-H," which stand out as the key contributors. Other features like "QRS," "QT," and "QTC" have smaller contributions, either positive or negative, while "ST," "PRQ," and "RR-I" show minimal to no impact on the model's prediction.



(a) **(b)**

Figure 4.7: Interpreting ECG feature contributions in arrhythmia, ischemia, and healthy heart conditions using SHAP plots, ranked by importance. (a) Waterfall SHAP Plots for Individual Feature Importance of FL-ANN Model; (b) Bar SHAP Plots for Feature Importance of FL-ANN Model

4.6.2 FL-LSTM Model Explanation using SHAP



(a) **(b)**

Figure 4.8: Interpreting ECG feature contributions in arrhythmia, ischemia, and healthy heart conditions using SHAP plots, ranked by importance. (a) Waterfall SHAP Plots for Individual Feature Importance of FL-LSTM Models; (b) Bar SHAP Plots for Feature Importance of FL-LSTM Models

We applied SHAP to the testing dataset to identify each feature's contribution to the model's output. Figure 4.8 represents two types of shap visualization for the FL-LSTM model. In figure 4.8(a), a local plot is displayed, where passing a row of SHAP values to the plot function generates a local feature importance plot. In this plot, the bars represent the SHAP values for each feature, illustrating their contributions to the specific prediction. The horizontal axis reflects the impact of each feature, with positive values driving the prediction higher and negative values pulling it lower, while the features are listed along the vertical axis. Here, the "R-H" feature has the most significant negative impact, with a SHAP value of -0.18, substantially reducing the prediction for that particular feature instance. In contrast, 'P-H' has the highest positive influence, with a SHAP value of +0.13, increasing the prediction. In figure 4.8(b), a global bar plot is shown, where passing a matrix of SHAP values to the bar plot function creates a global feature importance plot. In this plot, the global importance of each feature is determined by calculating the mean absolute SHAP value for that feature across all the samples given. It is also observed that the most influential features are "P-H" and "R-H," which stand out as the key contributors. Other features like "QRS," "QT," and "QTC" have smaller contributions, either positive or negative, while "ST," "PRQ," and "RR-I" show minimal to no impact on the model prediction.

4.7 Model Explanation using LIME

LIME's main objective is to provide interpretable local approximations for model predictions so that users may understand the reasoning behind particular choices. The likelihood of each situation is displayed by the horizontal bars, which range from 0 to 1. The likelihood increases with the length of the bar. The status is indicated by the color of the bar: green, blue, and orange denote ischemia, arrhythmia, and healthy states, respectively. For the LIME explanation of our two models, we randomly selected row-3 for arrhythmia, row-95 for ischemia, and row-45 for healthy from the testing dataset.

4.7.1 FL-ANN Model Explanation using LIME

Figure 4.9 explains the FL-ANN model's predictions using LIME model for heart disease classification by showing prediction probabilities and feature contributions for three cases: arrhythmic, healthy, and ischemic. In the arrhythmic case of figure 4.9(a), the model predicts a 0.64 probability for being arrhythmic, 0.29 for being healthy, and 0.08 for being ischemic with contributing features P-H(-0.16),R-H(1.39),HR(57.92),QT(0.43). In the healthy case of Figure 4.9(b), the model assigns

a 0.78 probability to being healthy, 0.09 to being arrhythmic, and 0.13 to being ischemic, with significant features including P-H (0.04), R-H (0.93), HR (65.50), and QTC (0.42). In the ischemic case of Figure 4.9(c), the model predicts a 0.78 probability for being ischemic, 0.14 for being healthy, and 0.08 for being arrhythmic with contributing features P-H(0.20), HR (96.77) and R-H (0.90). These detailed contributions and feature values provide transparency into the model's decision-making process.

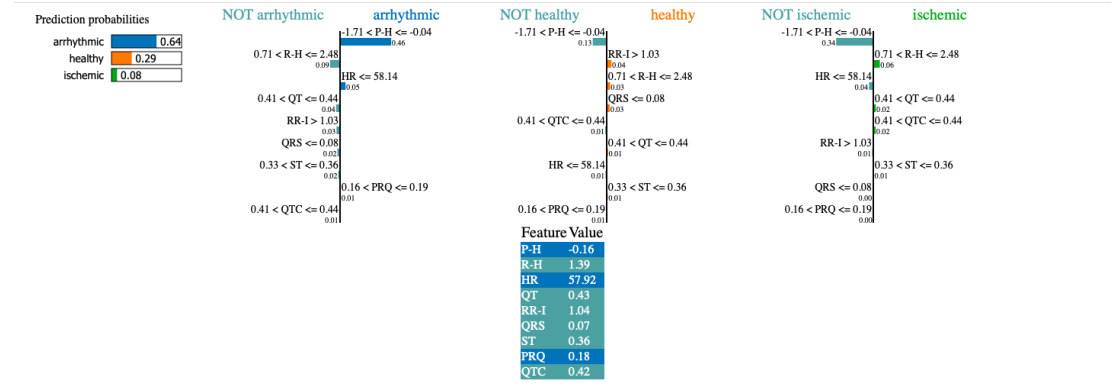
4.7.2 FL-LSTM Model Explanation using LIME

Figure 4.10 explains the FL-LSTM model's predictions using LIME model for heart disease classification by showing prediction probabilities and feature contributions for three cases: arrhythmic, healthy, and ischemic. In the arrhythmic case of Figure 4.10(a), the model predicts a 0.95 probability of arrhythmic, 0.05 for healthy, and 0.00 for ischemic, with contributing features being P-H (-0.16), R-H (1.39), HR (57.92), and QRS (0.07). In the healthy case of Figure 4.10(b) , the probabilities are 0.02 for arrhythmic, 0.82 for healthy, and 0.16 for ischemic, with significant features including P-H (0.04), R-H (0.93) and HR (65.50). In the ischemic case of Figure 4.10(c), the model predicts 0.04 for arrhythmic, 0.00 for healthy, and 0.96 for ischemic, with contributing features being P-H (-1.76), R-H (-0.21), HR (51.37), and QT (0.48). These detailed contributions and feature values provide transparency into the model's decision-making process.

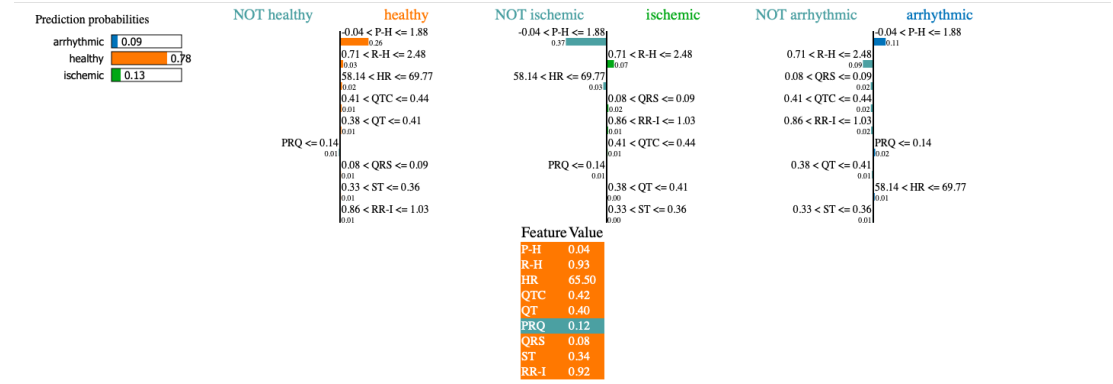
4.7.3 Interpretation of ECG Fiducial Features for Arrhythmia and Ischemia

To predict and classify ischemia and arrhythmias, it is essential to analyse ECG fiducial markers such P-height (P-H), R-wave height (R-H), the QRS complex, Heart rate(HR), QT interval and the corrected QT interval (QTc).These features provide crucial information on the electrical activity of the heart, acting as important markers for a range of cardiac disorders and improving the precision of diagnosis.

P-height (P-H) is a measure of atrial depolarisation; variations in P-H frequently indicate atrial enlargement or conduction abnormalities linked to arrhythmias such as atrial fibrillation. These conditions can lead to irregular heart rhythms, significantly increasing the risk of severe cardiovascular events. Therefore, P-H analysis is essential for early diagnosis and treatment of arrhythmia. Using R-wave height (R-H) to measure ventricular function is essential. A drop in R-H may indicate myocardial ischemia where damaged myocardial tissue lowers electrical activity. A rise in R-H, on the other hand, might be a sign of left ventricular hypertrophy, which is a risk factor for ventricular arrhythmias. Finding patients who are at risk for these potentially life-threatening conditions is made easier with accurate R-H interpretation.



(a)

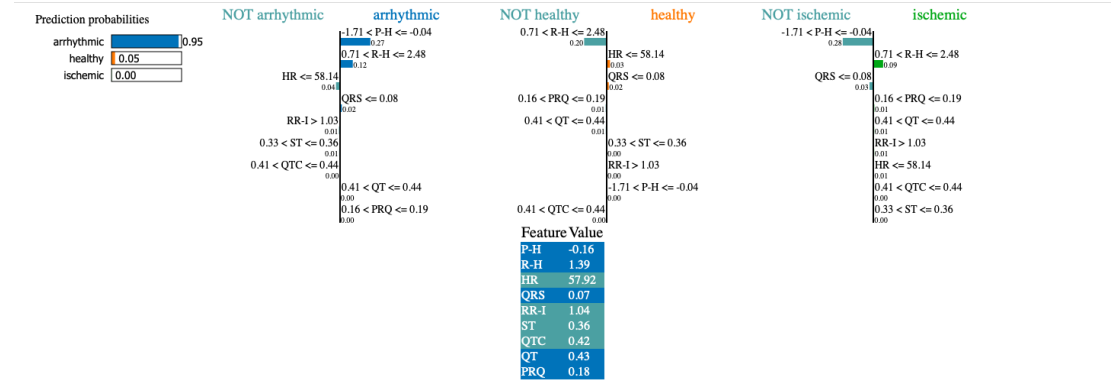


(b)

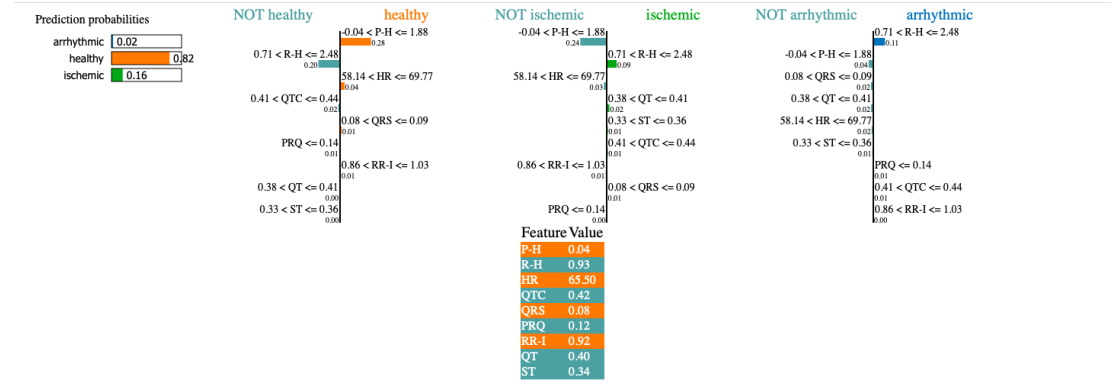
(c)

Figure 4.9: Individualized Interpretation of ECG-based heart disease prediction using the LIME Approach with FL-ANN model: (a) The cells highlighted in blue indicate the features that had the greatest influence on predicting arrhythmia risk. (b) The cells highlighted in orange indicate the features that had the greatest influence on predicting healthy individuals. (c) The cells highlighted in green indicate the features that had the greatest influence on predicting ischemia risk.

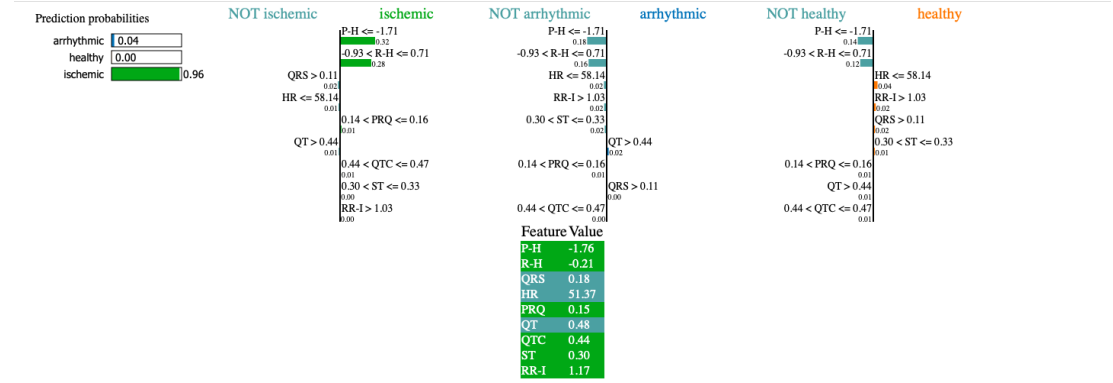
Ventricular depolarisation, or the QRS complex, is a key indicator for ischemia and arrhythmia diagnosis. The QRS complex's prolongation or aberrant morphology may indicate intraventricular conduction delays, which can result in inefficient cardiac activity and an increased risk of arrhythmic events. These delays can be observed in



(a)



(b)



(c)

Figure 4.10: Individualized Interpretation of ECG-based heart disease prediction using the LIME Approach with FL-LSTM model: (a) The cells highlighted in blue represent the features that had the most significant impact on predicting arrhythmia risk. (b) The cells highlighted in orange represent the features that had the most significant impact on predicting healthy individuals. (c) The cells highlighted in green represent the features that had the most significant impact on predicting ischemia risk.

circumstances such as bundle branch blockages. The time between ventricular depolarisation and repolarisation is represented by the QT interval and its corrected form, QTc. It is generally known that prolongation of these intervals indicate a higher risk of ventricular arrhythmias, such as torsades de pointes, which can cause sudden car-

diac death. In ischemic situations, where repolarization is frequently compromised, it is particularly important to monitor these intervals to identify patients who are at increased risk.

The application of explainable artificial intelligence (XAI) techniques, such as SHAP and LIME, further increases the clinical utility of these models by providing transparency in decision-making. These XAI techniques provide an understanding of how each model uses particular parameters to predict ischemia, arrhythmia, or general heart conditions by elucidating the intricate links between the models and important ECG characteristics. In the context of diagnosing and treating heart disease, this transparency is essential because it helps researchers and clinicians understand and trust the prediction of the models. Our findings demonstrate the significant impact of these ECG fiducial features on all classification models used in this study. These models are more applicable in clinical settings when they can be interpreted and understood, which promotes a more precise and reliable diagnosis of cardiovascular disease and, eventually, improves patient outcomes.

4.8 Comparison with State-of-the-Art Methods

Table 4.3 represents a comparison of our proposed architecture with state-of-the-art models of heart disease classification. Our proposed models achieved remarkable accuracy along with data collaboration while preserving privacy using federated learning approach and provide explanations of the results using explainable artificial intelligence(XAI).

It is observed in table 4.3 that most of the research works are based on different machine learning or deep learning methods for which data must be present in a centralized location to automate the diagnosis of heart diseases. In the classification of arrhythmia, traditional machine learning algorithms such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF) were utilized in [36] using MIT-BIH dataset and Convolutional Neural Network(CNN) was applied in [78] using sinus rhythm ECGS from 1,80,922 patients. Additionally, various models, including Artificial Neural Networks (ANN), SVM, and KNN, were applied to the European ST-T database to predict ischemia [28]. Finally, federated learning was explored for arrhythmia classification with the MIT-BIH dataset in [79], showcasing a collaborative approach in model training. However, all these approaches lacked interpretability in their outcomes.

So the majority of contemporary research encounters challenges stemming from insufficient collaboration, concerns regarding data privacy, and the restricted availabil-

ity of clinical datasets. Moreover, these existing classification models were typically designed to focus on a single disease, either arrhythmia or ischemia heart disease. Despite employing more complex and deeper models, it is evident that some studies struggled to achieve comparable performance and overlooked the interpretability or explainability of their results. This further justifies the usefulness of our proposed model.

Comparing it to the state-of-the-art models, it is noticed that our proposed FL-ANN and FL-LSTM models have attained accuracies of 87% and 90%, respectively, in classifying three distinct heart conditions, while also ensuring data collaboration, privacy protection, and interpretability of the outcome.

Table 4.3: Comparison with state-of-the-art methods for heart disease classification

Reference	Method	Dataset used	Class Label	Primary Results
Sraitih*et al*. [36]	SVM, KNN, RF	MIT-BIH	Arrhythmia	(Accuracy: 83%)
Attia *et al*. [78]	CNN	Sinus Rhythm ECGS from 1,80,922 patients	Arrhythmia	(AUC: 0.90, Accuracy: 83.3%)
Murthy *et al*. [28]	ANN, SVM, KNN	European ST-T	Ischemia	(Accuracy: 96.62%)
Couto *et al*. [79]	FL	MIT-BIH	Arrhythmia	(Recall: 0.83, Precision: 0.58)
Proposed Method	FL-LSTM with Shap and LIME	MIT-BIH, Europian ST-T, Fantasia	Arrhthmia Ischemia Healthy	(AUC: 0.99, Accuracy:90%)

4.9 Discussion

In this study, we introduce an explainable federated learning model designed to classify individuals into three categories: healthy, ischaemia, and arrhythmia. The model uses ECG feature data as input and generates corresponding class labels. By using a federated learning approach, we ensure heart disease is diagnosed through key ECG features, known as fiducial points, which capture important signals in heart activity. These features were collected and combined from three well-known ECG databases to create a strong and diverse dataset for accurate classification. To maintain balance and fairness, we applied data pre-processing and used SMOTE to augment underrepresented classes, improving the model's performance for all users.

Federated learning enables multiple decentralized devices to collaboratively train a shared model while keeping their raw data local and private. A central server first sends out a global model to each client device. Clients then train the model locally using their own ECG data with deep learning techniques like Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM) networks. Once training is done, the clients send back updates (like changes to model parameters), which the server combines using the Federated Averaging (FedAvg) method. This improved model is then sent back to the clients, and the cycle repeats until the model is fully trained.

To make the model results easier to understand, we used two explanation models SHAP and LIME. These models help show how much each ECG feature (like P-wave height, R-wave height, QRS complex, heart rate, QT interval, and corrected QT interval) contributed to the final prediction. This makes the model more transparent and supports doctors in making informed decisions.

Our proposed models, FL-ANN and FL-LSTM, achieved strong classification accuracies of 87% and 90%, respectively. By combining privacy-preserving learning, accurate classification, and understandable results, our approach offers a helpful way to diagnose heart conditions and support clinical decision making.

Chapter 5

Conclusion

5.1 Summary

In this study, we present an explainable federated learning model designed to accurately classify participants into three categories: healthy, ischaemia, and arrhythmia. The model generates class labels by processing ECG feature data as input. By leveraging the federated learning approach, heart disease is diagnosed through the extraction of fiducial features from ECG recordings, each of which contains key reference points. These features were then extracted and combined from three well-known ECG databases, creating a robust dataset for the classification of the three cardiac conditions. To ensure balance and fairness across the dataset, pre-processing steps were applied, and SMOTE was utilized for data augmentation, optimizing the model's performance across all clients. We adopted federated learning, a collaborative framework where multiple decentralized devices work together to train a unified model for heart disease classification. The process starts with a central server initializing a global model and distributing it to decentralized clients. Deep learning architectures, including Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM) networks, are then used to train individual client models locally on their respective datasets. After local training, each client sends model update such as adjustments to gradients or parameters back to the central server. These updates were then aggregated by the Federated Averaging (FedAvg) algorithm, typically by averaging the received updates, resulting in an improved global model. This refined model is then returned to the clients, and the process repeats iteratively until convergence is achieved.

Finally, we employed SHAP and LIME, two powerful explainable models, to provide clear visual interpretations of the model outcomes. These models demonstrated the significance and contribution of each ECG fiducial feature, including P-wave height (P-H), R-wave height (R-H), the QRS complex, heart rate (HR), QT interval, and corrected QT interval (QTc), providing crucial insights into the heart's electrical activity

and serving as key markers for cardiac disorders. By ensuring data collaboration, privacy preservation and the interpretability of the results, our proposed FL-ANN and FL-LSTM models achieved impressive accuracies of 87% and 90%, respectively. The insights generated by these explainable models are expected to support clinical decision-making, facilitating patient diagnosis and treatment while empowering physicians with transparent and understandable explanations for their diagnoses.

5.2 Future Works

There are several areas for improvement. One key challenge is the absence of a single publicly available dataset containing ischemia, arrhythmia, and healthy control groups. As a result, the three datasets used in this analysis come from distinct populations, which may introduce bias stemming from differences in demographics, genetics, and lifestyle factors. Moreover, Federated learning is a new field in healthcare which comes with its own challenges, such as ensuring data consistency across devices, handling device heterogeneity, and addressing issues related to communication latency and bandwidth constraints. Further experiments can be performed with signal data of heart diseases and the effectiveness of the proposed pipelines can also be explored in other domains. An explainable federated learning model for heart disease classification along with ensuring state-of-the-art performance can be an interesting research domain to consider in future endeavors.

REFERENCES

[1] (2024) Harley Street Heart and Vascular Centre. [Online]. Available: https://www.harleystreet.sg/kb/about-heart-disease/arrhythmia/

[2] (2024) Cleveland Clinic. [Online]. Available: https://my.clevelandclinic.org/health/diseases/ischemia

[3] A. Rawshani, "The ECG Leads: Electrodes, Limb Leads, Chest (Precordial) Leads, 12-Lead ECG (EKG)," *ECG Learning*, vol. 29, 2017.

[4] J. Park, J. An, J. Kim, S. Jung, Y. Gil, Y. Jang, K. Lee, and I.-y. Oh, "Study on the Use of Standard 12-Lead ECG Data for Rhythm-type ECG Classification Problems," *Computer Methods and Programs in Biomedicine*, vol. 214, p. 106521, 2022.

[5] R. J. Oweis and B. O. Al-Tabbaa, "QRS Detection and Heart Rate Variability Analysis: A Survey," *Biomed. Sci. Eng*, vol. 2, no. 1, p. 13, 2014.

[6] T. A. Sathi, R. Jany, R. Z. Ela, A. Azad, S. A. Alyami, M. A. Hossain, and I. Hussain, "An Interpretable Electrocardiogram-based Model for predicting Arrhythmia and Ischemia in Cardiovascular Disease," *Results in Engineering*, vol. 24, p. 103381, 2024.

[7] A. Raza, K. P. Tran, L. Koehl, and S. Li, "Designing ECG Monitoring Healthcare System with Federated Transfer Learning and Explainable AI," *Knowledge-Based Systems*, vol. 236, p. 107763, 2022.

[8] (2023) WORLD HEART VISION 2030: DRIVING POLICY CHANGE. [Online]. Available: https://world-heart-federation.org/world-heart-vision-2030/

[9] S. Don, D. Chung, D. Min, and E. Choi, "Analysis of Electrocardiogram Signals of Arrhythmia and Ischemia using Fractal and Statistical Features," *Journal of Mechanics in Medicine and Biology*, vol. 13, no. 01, p. 1350008, 2013.

[10] Y. M. Ayano, F. Schwenker, B. D. Dufera, and T. G. Debelee, "Interpretable Machine Learning Techniques in ECG-based Heart Disease Classification: A Systematic Review," *Diagnostics*, vol. 13, no. 1, p. 111, 2022.

[11] T. W. Bae and K. K. Kwon, "ECG PQRST Complex Detector and Heart Rate Variability Analysis using Temporal Characteristics of Fiducial Points," *Biomedical Signal Processing and Control*, vol. 66, p. 102291, 2021.

[12] I. Hussain and S. J. Park, "Big-ECG: Cardiographic Predictive Cyber-Physical System for Stroke Management," *IEEE Access*, vol. 9, p. 123146, 2021.

[13] M. S. Islam, I. Hussain, M. M. Rahman, S. J. Park, and M. A. Hossain, "Explainable Artificial Intelligence Model for Stroke Prediction using EEG Signal," *Sensors*, vol. 22, no. 24, p. 9859, 2022.

[14] I. Hussain and S. J. Park, "HealthSOS: Real-Time Health Monitoring System for Stroke Prognostics," *IEEE Access*, vol. 8, p. 213574, 2020.

[15] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. J. M. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boer, "Efficient Prediction of Cardiovascular Disease using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques," *IEEE Access*, vol. 9, p. 19304, 2021.

[16] I. Hussain, R. Jany, R. Boyer, A. Azad, S. A. Alyami, S. J. Park, M. M. Hasan, and M. A. Hossain, "An Explainable EEG-Based Human Activity Recognition Model using Machine-Learning Approach and LIME," *Sensors*, vol. 23, no. 17, p. 7452, 2023.

[17] A. Anand, T. Kadian, M. K. Shetty, and A. Gupta, "Explainable AI Decision Model for ECG data of Cardiac Disorders," *Biomedical Signal Processing and Control*, vol. 75, p. 103584, 2022.

[18] G. B. Moody and R. G. Mark, "The Impact of the MIT-BIH Arrhythmia Database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, p. 45, 2001.

[19] A. Taddei, G. Distante, M. Emdin, P. Pisani, G. Moody, C. Zeelenberg, and C. Marchesi, "The European ST-T database: Standard for Evaluating Systems for the Analysis of ST-T Changes in Ambulatory Electrocardiography," *European Heart Journal*, vol. 13, no. 9, p. 1164, 1992.

[20] N. Iyengar, C. Peng, R. Morin, A. L. Goldberger, and L. A. Lipsitz, "Age-Related Alterations in the Fractal Scaling of Cardiac Interbeat Interval Dynamics," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 271, no. 4, p. R1078, 1996.

[21] S. Lundberg and S. Lee, "A Unified Approach to interpreting Model Predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, p. 4768, 2017.

[22] M. Ribeiro, S. Singh, and C. Guestrin, “"Why Should I Trust You?" Explaining the Predictions of any Classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, p. 1135.

[23] (2024) Cleveland Clinic. [Online]. Available: https://my.clevelandclinic.org/health/diseases/16749-arrhythmia

[24] S. Dey, R. Pal, and S. Biswas, “Deep Learning Algorithms for Efficient Analysis of ECG Signals to Detect Heart Disorders,” in *Biosignal Processing*, 2022.

[25] Y. M. Ayano, F. Schwenker, B. D. Dufera, and T. G. Debelee, “Interpretable Machine Learning Techniques in ECG-based Heart Disease Classification: A Systematic Review,” *Diagnostics*, vol. 13, no. 1, p. 111, 2022.

[26] P. Kligfield, L. S. Gettes, J. J. Bailey, R. Childers, B. J. Deal, E. W. Hancock, G. Van Herpen, J. A. Kors, P. Macfarlane, D. M. Mirvis *et al.*, “Recommendations for the Standardization and Interpretation of the Electrocardiogram: Part I: the Electrocardiogram and its Technology: A Scientific Statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society endorsed by the International Society for Computerized Electrocardiology,” *Circulation*, vol. 115, no. 10, p. 1306, 2007.

[27] R. Rouhi, M. Clausel, J. Oster, and F. Lauer, “An Interpretable Hand-Crafted Feature-Based Model for Atrial Fibrillation Detection,” *Frontiers in Physiology*, vol. 12, 2021.

[28] H. N. Murthy and M. Meenakshi, “ANN, SVM and KNN Classifiers for Prognosis of Cardiac Ischemia-A Comparison,” *Bonfring International Journal of Research in Communication Engineering*, vol. 5, no. 2, p. 7, 2015.

[29] J. Park, W. Pedrycz, and M. Jeon, “Ischemia Episode Detection in ECG using Kernel Density Estimation, Support Vector Machine and Feature Selection,” *Biomedical Engineering Online*, vol. 11, p. 1, 2012.

[30] H. Lee and M. Shin, “Learning Explainable Time-Morphology Patterns for Automatic Arrhythmia Classification from Short Single-Lead ECGs,” *Sensors*, vol. 21, no. 13, p. 4331, 2021.

[31] M. Athif, P. C. Yasawardene, and C. Daluwatte, “Detecting Atrial Fibrillation from Short Single Lead ECGs using Statistical and Morphological Features,” *Physiological Measurement*, vol. 39, no. 6, p. 064002, 2018.

[32] R. Czabanski, K. Horoba, J. Wrobel, A. Matonia, R. Martinek, T. Kupka, M. Jezewski, R. Kahankova, J. Jezewski, and J. M. Leski, "Detection of Atrial Fibrillation Episodes in Long-Term Heart Rhythm Signals using a Support Vector Machine," *Sensors*, vol. 20, no. 3, p. 765, 2020.

[33] A. Rizwan, A. Zoha, I. B. Mabrouk, H. M. Sabbour, A. S. Al-Sumaiti, A. Alomainy, M. A. Imran, and Q. H. Abbasi, "A Review on the State of the art in Atrial Fibrillation Detection enabled by Machine Learning," *IEEE Reviews in Biomedical Engineering*, vol. 14, p. 219, 2020.

[34] S. J. M. Yazdi, H.-J. Park, C.-S. Son, and J.-H. Lee, "A Novel Machine Learning Approach to classify and detect Atrial Fibrillation using Optimized Implantable Electrocardiogram Sensor," *IEEE Access*, vol. 9, p. 149250, 2021.

[35] S. Kuila, N. Dhanda, and S. Joardar, "Feature Extraction and Classification of MIT-BIH Arrhythmia Database," in *Proceedings of the 2nd International Conference on Communication, Devices and Computing: ICCDC 2019*, 2020, p. 417.

[36] M. Sraitih, Y. Jabrane, and A. Hajjam El Hassani, "An Automated System for ECG Arrhythmia Detection using Machine Learning Techniques," *Journal of Clinical Medicine*, vol. 10, no. 22, p. 5450, 2021.

[37] A. H. Khan, M. Hussain, and M. K. Malik, "Arrhythmia Classification Techniques using Deep Neural Network," *Complexity*, vol. 2021, p. 1, 2021.

[38] Q. Xiao, K. Lee, S. A. Mokhtar, I. Ismail, A. L. b. M. Pauzi, Q. Zhang, and P. Y. Lim, "Deep Learning-Based ECG Arrhythmia Classification: A Systematic Review," *Applied Sciences*, vol. 13, 2023.

[39] I. Tougui, A. Jilbab, and J. El Mhamdi, "Heart Disease Classification using Data Mining Tools and Machine Learning Techniques," *Health and Technology*, vol. 10, p. 1137, 2020.

[40] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "ECG Heartbeat Classification: A Deep Transferable Representation," in *2018 IEEE international conference on healthcare informatics (ICHI)*, 2018, p. 443.

[41] S. Irfan, N. Anjum, T. Althobaiti, A. A. Alotaibi, A. B. Siddiqui, and N. Ramzan, "Heartbeat Classification and Arrhythmia Detection using a Multi-Model Deep Learning Technique," *Sensors*, vol. 22, no. 15, p. 5606, 2022.

[42] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-Lead Electrocardiogram Database for Arrhythmia Research covering more than 10,000 Patients," *Scientific Data*, vol. 7, no. 1, p. 48, 2020.

[43] Z. Yao, Z. Zhu, and Y. Chen, "Atrial Fibrillation Detection by Multi-Scale Convolutional Neural Networks," in *2017 20th International Conference on Information Fusion (Fusion)*. IEEE, 2017, pp. 1–6.

[44] S. Liaqat, K. Dashtipour, A. Zahid, K. Assaleh, K. Arshad, and N. Ramzan, "Detection of Atrial Fibrillation using a Machine Learning Approach," *Information*, vol. 11, no. 12, p. 549, 2020.

[45] U. R. Acharya, H. Fujita, O. S. Lih, Y. Hagiwara, J. H. Tan, and M. Adam, "Automated Detection of Arrhythmias using Different Intervals of Tachycardia ECG Segments with Convolutional Neural Network," *Information sciences*, vol. 405, p. 81, 2017.

[46] J. D. Peshave and R. Shastri, "Feature Extraction of ECG signal," in *2014 International Conference on Communication and Signal Processing*. IEEE, 2014, pp. 1864–1868.

[47] S. L. Oh, E. Y. Ng, R. San Tan, and U. R. Acharya, "Automated Diagnosis of Arrhythmia using Combination of CNN and LSTM Techniques with Variable Length Heart Beats," *Computers in Biology and Medicine*, vol. 102, p. 278, 2018.

[48] S. Sattar, R. Mumtaz, M. Qadir, S. Mumtaz, M. A. Khan, T. De Waele, E. De Poorter, I. Moerman, and A. Shahid, "Cardiac Arrhythmia Classification using Advanced Deep Learning Techniques on Digitized ECG Datasets," *Sensors*, vol. 24, no. 8, p. 2484, 2024.

[49] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," *ArXiv Preprint ArXiv:2006.11371*, 2020.

[50] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—Explainable Artificial Intelligence," *Science Robotics*, vol. 4, no. 37, p. eaay7120, 2019.

[51] D. Dave, H. Naik, S. Singhal, and P. Patel, "Explainable AI meets Healthcare: A study on Heart Disease Dataset," *ArXiv Preprint ArXiv:2011.03195*, 2020.

[52] T. A. Abdullah, M. S. M. Zahid, and W. Ali, "A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions," *Symmetry*, vol. 13, no. 12, p. 2439, 2021.

[53] Q. Hua, Y. Yaqin, B. Wan, B. Chen, Y. Zhong, and J. Pan, "An Interpretable Model for ECG Data based on Bayesian Neural Networks," *IEEE Access*, vol. 9, p. 57001, 2021.

[54] V. Krasteva, I. Christov, S. Naydenov, T. Stoyanov, and I. Jekova, "Application of Dense Neural Networks for Detection of Atrial Fibrillation and Ranking of Augmented ECG Feature Set," *Sensors*, vol. 21, no. 20, p. 6848, 2021.

[55] P. Guleria, P. Naga Srinivasu, S. Ahmed, N. Almusallam, and F. K. Alarfaj, "XAI Framework for Cardiovascular Disease Prediction using Classification Techniques," *Electronics*, vol. 11, no. 24, p. 4086, 2022.

[56] I. Neves, D. Folgado, S. Santos, M. Barandas, A. Campagner, L. Ronzio, F. Cabitza, and H. Gamboa, "Interpretable Heartbeat Classification using Local Model-Agnostic Explanations on ECGs," *Computers in Biology and Medicine*, vol. 133, p. 104393, 2021.

[57] M. Bodini, M. W. Rivolta, and R. Sassi, "Interpretability Analysis of Machine Learning Algorithms in the Detection of ST-Elevation Myocardial Infarction," in *2020 Computing in Cardiology*, 2020, p. 1.

[58] A. Rahman, M. S. Hossain, G. Muhammad, D. Kundu, T. Debnath, M. Rahman, M. S. I. Khan, P. Tiwari, and S. S. Band, "Federated Learning-based AI Approaches in Smart Healthcare: Concepts, Taxonomies, Challenges and Open Issues," *Cluster computing*, vol. 26, no. 4, p. 2271, 2023.

[59] A. Nilsson, S. Smith, G. Ulm, E. Gustavsson, and M. Jirstrand, "A Performance Evaluation of Federated Learning Algorithms," in *Proceedings of the Second Workshop on Distributed Infrastructures for Deep Learning*, 2018, p. 1.

[60] M. M. Yaqoob, M. Nazir, A. Yousafzai, M. A. Khan, A. A. Shaikh, A. D. Algarni, and H. Elmannai, "Modified Artificial Bee Colony Based Feature Optimized Federated Learning for Heart Disease Diagnosis in Healthcare," *Applied Sciences*, vol. 12, no. 23, p. 12080, 2022.

[61] M. M. Yaqoob, M. Nazir, M. A. Khan, S. Qureshi, and A. Al-Rasheed, "Hybrid classifier-based federated learning in health service providers for cardiovascular disease prediction," *Applied Sciences*, vol. 13, no. 3, p. 1911, 2023.

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

[63] AcqKnowledge Software. [Online]. Available: https://www.biopac.com/product/acqknowledge-software/

[64] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Transactions on Biomedical Engineering*, no. 3, p. 230, 1985.

[65] S. De Raedt, A. De Vos, and J. De Keyser, "Autonomic Dysfunction in Acute Ischemic Stroke: An Underexplored Therapeutic Area," *Journal of the Neurological Sciences*, vol. 348, no. 1-2, p. 24, 2015.

[66] I. Ogrezeanu, D. Stoian, A. Turcea, and L. M. Itu, "Deep Learning based Myocardial Ischemia Detection in ECG Signals," in *2020 24th International Conference on System Theory, Control and Computing (ICSTCC)*, 2020, p. 250.

[67] E. Pueyo, L. Sornmo, and P. Laguna, "QRS Slopes for Detection and Characterization of Myocardial Ischemia," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 2, p. 468, 2008.

[68] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. de Gusmão *et al.*, "Flower: A friendly federated learning framework," 2022.

[69] S. Hong, W. Zhang, C. Sun, Y. Zhou, and H. Li, "Practical Lessons on 12-Lead ECG Classification: Meta-Analysis of Methods from PhysioNet computing in Cardiology Challenge 2020," *Frontiers in Physiology*, vol. 12, p. 811661, 2022.

[70] F. Murat, O. Yildirim, M. Talo, U. B. Baloglu, Y. Demir, and U. R. Acharya, "Application of Deep Learning Techniques for Heartbeats Detection using ECG Signals Analysis and Review," *Computers in Biology and Medicine*, vol. 120, p. 103726, 2020.

[71] M. R. Kiladze, U. A. Lyakhova, P. A. Lyakhov, N. N. Nagornov, and M. Vahabi, "Multimodal Neural Network for Recognition of Cardiac Arrhythmias Based on 12-Lead Electrocardiogram Signals," *IEEE Access*, vol. 11, p. 133744, 2023.

[72] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP," in *Proceedings of the AAAI/ACM Conference on AI, Ethics and Society*, 2020, p. 180.

[73] Z. Zhou, G. Hooker, and F. Wang, "S-LIME: Stabilized-LIME for Model Explanation," in *Proceedings of the 2021 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2021, p. 2429.

[74] G. Visani, E. Bagli, and F. Chesani, "OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms," *ArXiv*, 2020.

[75] M. Zafar and N. Khan, "Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability," *Mach. Learn. Knowl. Extr.*, vol. 3, p. 525, 2021.

[76] S. Shankaranarayana and D. Runje, "ALIME: Autoencoder Based Approach for Local Interpretability," in *Intelligent Data Engineering and Automated Learning—IDEAL 2019*, 2019, p. 454.

[77] P. K. Diederik, "Adam: A Method for Stochastic Optimization," 2014.

[78] Z. I. Attia, P. A. Noseworthy, F. Lopez-Jimenez, S. J. Asirvatham, A. J. Deshmukh, B. J. Gersh, R. E. Carter, X. Yao, A. A. Rabinstein, B. J. Erickson *et al.*, "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*, vol. 394, no. 10201, pp. 861–867, 2019.

[79] R. S. Couto and L. C. Favaro, "Federated Learning Applied to Arrhythmia Detection on Electrocardiograms," in *2023 IEEE Virtual Conference on Communications (VCC)*, 2023, p. 305.

List of Publications

Journal

1. Tanjila *et al*., "An Interpretable Electrocardiogram-based Model for predicting Arrhythmia and Ischemia in Cardiovascular Disease", in *Results in engineering*, Elsevier, vol. 24, p.103381, 2024.(Q1)
   https://doi.org/10.1016/j.rineng.2024.103381