Prediction of Atrial Fibrillation and Associated Risks at Primary Care Level using Longitudinal Learning Stances

PIC2 - Master in Computer Science and Engineering
Instituto Superior Técnico, Universidade de Lisboa

Henrique Anjos — 99081*
henriquemdanjos@tecnico.ulisboa.pt

Advisor: Rafael Costa
Co-advisor: Rui Henriques

Abstract Atrial fibrillation (AF) is recognized as the most prevalent cardiac arrhythmia worldwide, and its occurrence is strongly associated with a significant risk increase of stroke, heart failure, and mortality. Although several traditional methods have been developed to detect and prognostic AF from biometric data, these approaches often fail to fully capture the complexity of AF patterns, limiting their predictive accuracy. Over recent years, there has been an increasing amount of research exploring the potential of machine learning (ML) techniques for predicting AF, with many studies experimentally showing superior performance against conventional methods. However, most state-of-the-art ML models for AF prediction fail to incorporate longitudinal data, limiting their ability to account for the evolving nature of individual behaviors and cardiophysiological indicators over varying prediction horizons. The lack of a longitudinal perspective often results in a limited understanding of how the risk of AF develops and changes over time. In this study, we aim to address this gap by further investigating and developing advanced ML models specifically designed to predict AF and associated events within a longitudinal Portuguese cohort. By incorporating temporal stances, this research offers a more comprehensive analysis of the progression of AF and its related conditions. Additionally, to bridge the gap between predictive modeling and clinical application, we introduce a prototype decision-support interface that integrates these models. The interface is designed to offer clinicians a platform that provides real-time, data-driven insights into the likelihood of AF and related events. By enabling more accurate and timely clinical decision-making, this study aims to improve patient outcomes and reduce healthcare costs associated with managing AF at primary care centers.

Keywords — Atrial Fibrillation, Machine Learning, Medical Tool Interface, Clinical Longitudinal Data, Portuguese Population, Electrocardiogram, Decision Support

---

*I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa (https://nape.tecnico.ulisboa.pt/en/apoio-ao-estudante/documentos-importantes/regulamentos-da-universidade-de-lisboa/).

Contents

1   Introduction

Atrial fibrillation (AF) is the most common cardiac arrhythmia, and its occurrence associated with a significant risk increase of stroke, heart failure, and mortality [1]. Despite its prevalence and severe consequences, AF often goes undiagnosed until complications arise due to its episodic nature and the lack of consistent early-warning signs [2]. This diagnostic gap highlights the need for tools that can predict and detect AF effectively, particularly in its early stages.

A plethora of traditional methods have been proposed for predicting AF that rely on heuristic approaches or rule-based systems [3]. While functional, they often fail to capture the complexity of AF patterns [4]. In the last years, machine learning (ML) models have demonstrated superior performance compared to these traditional approaches [5]. However, the existing state-of-the-art ML methods generally suffer from two major drawbacks:

1. failing to incorporate longitudinal data, which limits their ability to analyze the progression of biometric and cardiophysiological indicators across varying prediction horizons.

2. overlooking the importance of integrating different screening methods. Many approaches do not fully leverage the predictive value embedded in electrophysiological signals and neglect other critical factors such as risk behaviors, comorbidities, and drug regimen.

This thesis aims to address these challenges by developing ML models tailored to assist in the early detection of AF and the prediction of associated risks. The models are built using data collected from a Portuguese cohort, with a focus on incorporating longitudinal data from diverse primary care screenings, including demographics, biometric indicators, clinical and physiological history, lifestyle factors and electrophysiological data. By integrating these data modalities, the proposed models aspire to enhance the early detection of AF.

In addition, this thesis will also explore the ability to prognosticate AF and related outcomes, such as hospital admissions, incopassing myocardial infarction, heart failure, or any cause, as well as mortality from any cause or cardiovascular mortality, to provide deeper insights into patient risks.

Finally, to bridge the gap between predictive modeling and clinical application, this project also introduces a prototype interface that integrates the predictive models, offering clinicians an intuitive decision-support tool.

This project is conducted in collaboration with the Unidade Local de Saúde de Matosinhos (ULSM), with external validation to be carried out in a clinical context by the partnering institution. While this approach holds significant promise, the study acknowledges key limitations, such as the need for extensive validation in real-world clinical settings and addressing potential biases inherent to the cohort data.

1.1   Work Objectives

The primary objective of this thesis is to create a machine learning-based clinical decision support tool tailored to assist in the early detection of AF and the prediction of associated risks, aiming to assist in clinical diagnosis. This goal is supported by the following specific objectives:

1. **Develop predictive models of AF:** Develop and optimize machine learning algorithms to accurately predict the occurrence of AF and related events, using routine clinical data and electrocardiographic data from a Portuguese cohort. This will involve the following key strategies:

   - Evaluating and comparing state-of-the-art approaches with emerging methodologies, such as those leveraging deep learning advancements, to determine the most effective techniques for the cohort.
   - Assessing the predictive value of longitudinal data views and exploring the potential of direct electrocardiographic signal processing for improved prediction accuracy, specifically considering the cohort's characteristics.

- Conducting a rigorous evaluation of the various predictive strategies to ensure their suitability for the target cohort and generalization capacity.

2. **Design and develop a clinical decision support interface:** Complementarily, define and implement relevant interfaces to integrate the predictive models, focusing on usability for healthcare professionals. This includes:

   - Developing both a programmatic API and a graphical user interface (GUI) to provide a robust foundation for clinical applications.
   - Establishing a solid framework that can be further refined to meet necessary usability and functional requirements.

Additionally, this research aims to leverage the predictive models to generate population-specific insights and acquire new knowledge, contributing to a deeper understanding of AF prediction and related events in the context of a well-established Portuguese cohort.

1.2 Expected Contributions

This thesis aims to make the following contributions:

1. **Scientific Contributions:** Advance the field of machine learning for AF prediction by developing one or more models that improve accuracy and usability compared to existing approaches.

2. **Practical Contributions:** Provide a prototype interface that integrates one or more predictive models, assisting healthcare professionals in clinical diagnosis and contributing to improved patient care.

3. **Population-Specific Insights:** Leverage data from a Portuguese cohort to address population-specific biases and create models tailored to this demographic.

4. **Foundation for Future Work:** Lay the groundwork for further research on AF prevention, screening policies, and the integration of machine learning tools into clinical workflows.

2 Background

This section covers key concepts and introduces the notation utilized throughout this work. We begin with an overview of heart failure and essential definitions in Section 2.1, followed by a review of foundational Machine Learning (ML) concepts in Section 2.2. Finally, Section 2.3 discusses the critical role of machine learning in predicting the risk of heart disease.

2.1 Heart Failure Essentials

Heart failure (HF) is a chronic condition in which the heart is unable to pump blood efficiently, leading to insufficient blood flow to meet the body's needs [6]. This condition arises due to a variety of underlying causes, including structural or functional abnormalities of the heart, and is often associated with other comorbidities that exacerbate its progression. HF is a significant global health concern, with rising prevalence and a substantial impact on mortality, morbidity, and healthcare costs [7–9].

This chapter will explore the critical aspects of heart failure, beginning with an overview of heart structure, followed by the role of heart disease comorbidities, the importance of anthropometric and clinical measurements, and the diagnostic contributions of electrocardiograms and echocardiogram.

Heart Structure The heart is divided into four chambers: the right atrium and right ventricle, and the left atrium and left ventricle (see Figure 1). The right atrium receives deoxygenated blood from the body through the superior and inferior vena cavae. Blood then flows through the tricuspid valve into

the right ventricle. From there, the right ventricle sends the blood to the lungs via the pulmonary valve and the pulmonary arteries, where it undergoes oxygenation [10]. In contrast, the left atrium receives oxygen-rich blood returning from the lungs through the pulmonary veins. The blood passes through the mitral valve into the left ventricle. The left ventricle, the strongest chamber, pumps the oxygenated blood through the aortic valve into the aorta, distributing it to the rest of the body.

The heart is enclosed in a protective protective double membrane called the pericardium, which contains fluid that reduces friction as the heart beats. The walls of the chambers are composed of cardiac muscle tissue, the myocardium, a contractile tissue whose contraction and relaxation are responsible for pumping blood through the heart and to the rest of the body. The coordinated opening and closing of the valves ensure unidirectional blood flow through the heart [10, 11].

In addition to these structures, the heart contains a specialized network of cells called the cardiac conduction system. This is a mini nervous system that generates and transmits electrical impulses that trigger the contraction of the heart muscles. Malfunctions in the cardiac conduction system can lead to arrhythmias, such as AF [12].
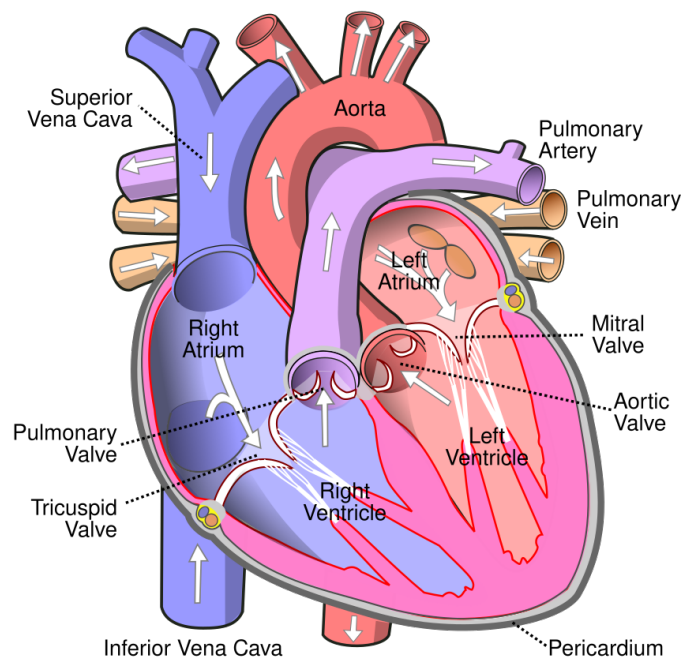

Figure 1: Diagram of the Heart and Blood Flow. Source: Wikipedia

Heart Disease comorbidities. Several pathological conditions are related to heart disease, each affecting the heart in different ways. These heart-related diseases can be interconnected, with one condition often leading to another, and all jointly contributing to the weakening of heart function. Acute myocardial infarction (AMI), commonly known as a heart attack, occurs when there is an insufficient supply of oxygen to the heart muscle (myocardium). This is often caused by a blood clot in one of the coronary arteries. If not treated promptly, the lack of oxygen can cause permanent damage to the heart tissue [13]. Another related condition is coronary artery disease (CAD), a chronic condition where the arteries supplying blood to the myocardium become narrowed or blocked due to the buildup of plaque (atherosclerosis). This reduces the amount of oxygen available to the heart muscle, impairing its function [14]. Angina, characterized by chest pain due to reduced blood flow to the heart, is often a precursor or companion to CAD [15]. Dyslipidemia, an abnormal amount of lipids in the blood, contributes to atherosclerosis and increases the likelihood of CAD and AMI [16]. Both AMI and CAD

can eventually lead to heart failure, a condition in which the heart is unable to pump blood effectively enough to meet the body's demands for oxygen and nutrients.

Chronic conditions like diabetes mellitus and chronic kidney disease (CKD) are also significant comorbidities, as they both accelerate vascular damage and increase the risk of heart disease [17, 18]. Sleep apnea, particularly obstructive sleep apnea, disrupts oxygen supply during sleep, placing strain on the heart and contributing to hypertension and arrhythmias, such as AF [19]. Additionally, structural abnormalities such as aortic valvular disease, left atrial dilation, and left ventricular dilation impair normal cardiac function and elevate the risk of heart failure [20].

Hypertension, or high blood pressure, is another major risk factor. In this condition, the pressure of the blood against the artery walls is consistently too high, which can damage the heart over time, making it weaker and more susceptible to the conditions mentioned above [21].

Anthropometric and Clinical Measurements. Assessing heart failure requires a range of clinical and anthropometric measurements that provide valuable insights into both cardiac and overall health. Key measurements include systolic and diastolic blood pressure, heart rate, weight, height, body mass index (BMI), age, gender, smoking status, alcohol consume, family history of hypertension and cardiovascular diseases, among other relevant factors [22–24]. These indicators help identify contributing risk factors such as hypertension and obesity, which are closely associated with the development and progression of heart failure [25, 26]. Monitoring these parameters allows healthcare providers to track changes over time, assess the effectiveness of treatment strategies, and adjust management plans as needed to improve patient outcomes.

Electrocardiogram. The electrocardiogram (ECG) is a crucial diagnostic tool that records the electrical activity of the heart. It provides important information regarding the heart's rhythm, rate, and conduction patterns. In the context of heart failure, the ECG can capture prodromal abnormalities, such as AF, which is often associated with heart failure [27]. Key features of the ECG include the RR interval (or heart rate), QRS complex, P-wave, and T-wave, all of which provide valuable insights into cardiac function.

The RR interval represents the time between two successive R-wave peaks, corresponding to one complete cardiac cycle [28]. It is a critical measure for determining heart rate and rhythm regularity. Irregular RR intervals are a hallmark of atrial fibrillation, indicating the erratic timing of ventricular contractions [29]. The QRS complex reflects the depolarization of the ventricles, which is the electrical activity leading to their contraction. It is typically characterized by a sharp and narrow waveform. Prolongation or abnormalities in the QRS complex can signal conduction issues, such as bundle branch blocks or ventricular hypertrophy, both of which may be linked to heart failure [30]. The P-wave represents atrial depolarization, corresponding to the contraction of the atria. Variations in the P-wave, such as prolonged duration, altered morphology, or the absence of the waveform, can indicate atrial conduction delays or arrhythmias, such as atrial fibrillation [31]. The T-wave reflects ventricular repolarization, which is the recovery phase of the ventricles after contraction. Abnormalities in the T-wave, such as flattening, inversion, or alternans (beat-to-beat variation in amplitude), may signal ventricular strain or ischemia, conditions often associated with heart failure [32].

Echocardiogram. The echocardiogram (ECHO) is a non-invasive imaging technique that uses ultrasound to create images of the heart, allowing healthcare providers to assess its structure and function. It essential to assess functional abnormalities in the heart. Key indicators of heart strain observable on an echocardiogram include left atrial dilation, left ventricular dilation, and ejection fraction [33, 34]. Left atrial dilation, which reflects elevated pressure within the heart, is a common finding in heart failure and can increase the risk of atrial fibrillation [35], often signaling worsening cardiac function. Left ventricular dilation, another critical measurement, occurs as a result of chronic pressure or volume overload, commonly due to prolonged hypertension or heart valve disease, and can eventually lead to dysfunction in the left ventricle itself, which is a major marker of heart failure [33]. With the heart measurements the ejection fraction (EF) can also be calculated, measuring the percentage of blood ejected from the left ventricle with each heartbeat. A reduced ejection fraction indicates impaired pumping ability and is

6

characteristic of systolic heart failure [36].

In addition to these structural measurements, an echocardiogram can reveal the presence of aortic valve disease [37]. Aortic valve disease, such as aortic stenosis, increases the workload on the left ventricle, leading to hypertrophy and, over time, contributing to heart failure [38]. Similarly, mitral valve disease can lead to elevated pressure in the left atrium, causing left atrial dilation and pulmonary congestion, which can further exacerbate symptoms of heart failure [39]. Together, these echocardiographic findings provide critical insights into the structural changes and functional impairments underlying heart failure, offering essential information for its diagnosis and management.

2.2 Machine Learning Essentials

Machine Learning (ML) is a branch of artificial intelligence (AI) focused on developing algorithms that enable computers to learn from data, supporting knowledge acquisition and decision making [40]. Instead of being explicitly programmed for specific tasks, ML systems use statistical techniques to identify patterns, build models, and improve performance as they encounter more data. This adaptability has led ML to become a fundamental tool across industries, particularly in healthcare, where it supports applications like diagnostic assistance, disease prediction, and personalized treatment planning. Broadly, ML methods fall into three main types: Supervised Learning, Unsupervised Learning and Reinforcement Learning. In this work, the later, Reinforcement Learning, will not be addressed as it is not relevant to the scope of this study.

Supervised Learning. The ML model is trained using labeled data, where input data is paired with the correct output. The algorithm learns a mapping between the input and output variables, which can be subsequently used to predict outputs for unseen data [41]. Through this process, the model identifies patterns in the data to accurately map inputs to desired outputs. Supervised learning in its simplest form can be divided into two types of problems: **classification** and **regression**.

- **Classification**: This type of problem involves predicting categorical outcomes, where the model learns to classify input data into one of several predefined categories or classes. Depending on the number of classes to predict from, classification can be further divided into three types: binary, multi-class and multi-label [42]. Binary classification tasks target output variables with only two possible categories, such as determining whether a patient has a specific disease (e.g., diabetes or no diabetes). Multi-class classifiers are inherently prepared to handle output variables with an arbitrary cardinality, such as diagnosing a patient with one of several potential diseases based on their symptoms. Multi-label classification tasks learning mapping models with multiple output variables, capturing their interdependencies between output variables; for example, classifying a patient as having both diabetes and hypertension.

- **Regression**: In regression problems, the model predicts one or more continuous numerical outputs, mapping inputs to real-valued estimates. Regression can be categorized into two types: single-output regression and multi-output regression [43]. Single-output regression focuses on predicting a single continuous value, such as estimating patient survival rates based on clinical factors. Multi-output regression extends this capability to predict multiple continuous values simultaneously, capturing relationships among the outputs; for example, estimating both a patient's life expectancy and their probability of developing diabetes.

Unsupervised Learning. Unsupervised learning deals with unlabeled data, where the goal is to uncover hidden structures or patterns in the data [44]. Unlike supervised learning, the model is not provided with correct outputs during training. Clustering, representation learning, pattern discovery, dimensionality reduction and anomaly analysis are typical tasks in unsupervised learning. In healthcare, unsupervised learning can be applied to model and segment patient populations based on clinical and molecular screening.

Machine Learning Algorithms ML encompasses a wide range of algorithms, each designed with particular strengths to address specific types of tasks. Below, we explore some of the most important algorithms in each category, highlighting their applications and unique benefits.

For classification tasks, **logistic regression** is one of the simplest and most widely used linear classifiers, predicting the probability of categorical outcomes [45]. Nonlinear variants, such as kernel logistic regression, extend its applicability to more complex datasets [46]. **Support vector machines (SVMs)** are another powerful option, finding an optimal hyperplane to separate classes and enabling nonlinear classification through the use of kernel functions [47]. Similarly, **k-Nearest Neighbors (kNN)** is a straightforward and versatile algorithm that classifies data by identifying the k closest data points based on a specified distance metric [48].

In regression tasks, **linear regression** establishes a relationship between input variables and a continuous output variable by fitting a straight line through the data points [49]. Nonlinear approaches, such as polynomial [50] and kernel regression [51], extend this framework to handle more complex relationships by fitting curved functions to the data. Analogizer algorithms, like **SVMs** and **kNN**, can also be applied to regression, offering flexible solutions for both linear and nonlinear patterns [47, 52].

Another important family of regression algorithms includes **decision trees**, which use an intuitive tree-like structure to split data into branches based on input features [53]. These can be extended into ensemble methods such as **random forests** and **gradient boosting**. Random forests improve predictive accuracy and reduce overfitting by building multiple trees on bootstrapped subsets of data and using random subsets of features for each split. Gradient boosting enhances predictions by sequentially constructing trees, where each new tree focuses on correcting errors made by the previous ones [54, 55].

Clustering algorithms, used for grouping similar data points, include **k-means clustering**, which partitions the data into k clusters by minimizing within-cluster variance. Each data point is assigned to the nearest cluster center, and the process iterates until the optimal clusters are formed [56]. Hierarchical clustering is another technique that builds a tree of clusters by either merging or splitting clusters at each step, enabling a flexible approach to understanding data structure [57]. Density-based clustering, like **DBSCAN**, is effective for handling noise in data and discovering clusters of arbitrary shapes, which can be beneficial when data is not well-separated or has outliers [58]. Dimensionality reduction techniques, such as **Principal Component Analysis (PCA)**, are crucial for simplifying complex datasets by reducing the number of features while retaining as much variance as possible. This process is especially valuable in high-dimensional spaces, where it can speed up computation and improve model performance [59].

Finally, **neural networks**, are powerful and versatile algorithms for predictive and representation learning. They are composed of interconnected layers of processing units called neurons, inspired by the structure and function of the human brain. These networks excel at capturing complex patterns and relationships in data through a process of learning that adjusts the connections between neurons based on the input they receive.

At the foundation of neural networks is the **perceptron** [60], a simple model consisting of a single neuron with weighted inputs, a bias term, and an activation function that determines its output. The perceptron serves as a building block for more complex architectures.

Expanding on the perceptron, the **Multi-Layer Perceptron** (MLP) introduces multiple layers of neurons organized into an input layer, one or more hidden layers, and an output layer. MLPs enable neural networks to learn and represent non-linear relationships by applying activation functions at each layer [61]. The learning process involves iteratively adjusting the weights of the network through an optimization method, typically using backpropagation and gradient descent to minimize a loss function [62]. This iterative optimization process improves the network's ability to map inputs to outputs effectively.

Deep Learning. Deep learning is a subset of machine learning that focuses on building and training large neural networks, known as deep neural networks, which feature multiple hidden layers [63]. Unlike traditional machine learning methods that rely on hand-crafted features, deep learning models automatically discover patterns and representations directly from raw data. This ability makes them especially powerful for complex tasks such as image recognition, natural language processing, and speech analysis.

One of the most prominent classes of architectures in deep learning is the **Convolutional Neural**

Network (CNN), which is specifically designed to exploit the spatial dependencies in data, particularly images. CNNs achieve this through specialized layers such as convolutional layers, which apply filters to detect local patterns, and pooling layers, which reduce spatial dimensions while preserving important features [64]. This hierarchical approach enables CNNs to extract increasingly abstract representations of the input data, making them highly effective for tasks like image classification and object detection.

Another key class of architectures is the **Recurrent Neural Network** (RNN), designed to handle sequential data. RNNs maintain context through internal memory mechanisms, allowing them to process sequences of information, such as time series or natural language text. This makes them well-suited for applications like language modeling, machine translation, and time-series forecasting. [65]

In recent years, **Graph Neural Networks** (GNNs) have emerged as a powerful class of models designed to process data represented as graphs. GNNs are capable of learning the relationships between entities within a graph structure By utilizing node and edge information, GNNs capture the dependencies between connected components, enabling them to make predictions based on structural information [66].

Mixed-variable models represent another important advancement in deep learning. These models are specifically designed to handle datasets with both continuous and categorical variables. By leveraging both types of data, mixed-variable models enable more flexible and accurate predictions, particularly in complex domains where the data is heterogeneous, such as healthcare, finance, and economics. These models combine the strengths of traditional neural network architectures with specialized techniques for dealing with diverse data types, making them highly effective for a wide range of applications.

Although deep learning models, particularly those using deep neural networks, often require substantial amounts of data and computational resources to perform well [67], they have revolutionized the field of artificial intelligence, achieving unprecedented accuracy across numerous domains [68].

Evaluation Metrics. In machine learning, evaluation metrics are crucial because they provide objective ways to assess model performance. While unsupervised learning lacks labeled outputs, making standard evaluation less straightforward, supervised learning benefits from well-defined metrics. With labeled data, supervised models allow for clear comparisons between predicted and actual outcomes, with distinct metrics for classification and regression.

For classification, **accuracy** serves as a fundamental metric, measuring the proportion of correct predictions out of all predictions made.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

However, accuracy alone can be misleading, especially with imbalanced datasets, a common issue in the healthcare field. It does not provide insight into the distribution of errors, such as false positives (FP) and false negatives (FN), which are crucial for understanding model performance, particularly in situations where the cost of misclassification differs across classes. Therefore, other metrics, such as **precision**, **sensitivity** also known as recall or true positive rate (TPR), **Fβ-score**, and **specificity**, are often employed.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F}\beta\text{-score} = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}} \tag{4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

Precision represents the proportion of true positives (TP) among all predicted positives, making it particularly valuable when the cost of false positives is high, such as diagnosing a disease that isn't

present. Sensitivity, on the other hand, measures the proportion of true positives out of all actual positive cases, which is critical when false negatives are costly, as in missing a serious condition in a patient. Specificity, also known as the true negative rate, evaluates the proportion of true negative (TN) cases correctly identified as negative, and it is particularly useful in contexts where avoiding false positives is important, such as in screening tests to reduce unnecessary follow-ups. The F1-score, which is the harmonic mean of precision and recall, balances these two metrics when the trade-off between false positives and false negatives is significant [69].

In multi-class contexts, these metrics can be either associated with a specific class of interest (treated as the positive class, with the remaining classes as negative) or, alternatively, extended using macro, micro, or weighted averaging. Macro averaging (equation 13) calculates the metric independently for each class and averages the results, treating all classes equally. Micro averaging (equation 14) aggregates the contributions of all classes to compute the metric, making it suitable for datasets with class imbalances. Weighted averaging (equation 15) adjusts for the proportion of instances in each class, giving more importance to metrics for larger classes [70]. In multi-label contexts, macro, micro, and weighted averaging are also used, with the addition of Hamming loss, which measures the fraction of incorrect labels in the predictions, and subset accuracy, which measures the percentage of instances that have all their labels correctly predicted [71].

Another essential metric for classification, particularly in binary classification tasks, is the Receiver Operating Characteristic Area Under the Curve (**ROC-AUC**). ROC-AUC evaluates a model's ability to distinguish between classes by measuring the area under the curve that plots the true positive rate (sensitivity) against the false positive rate at various threshold settings. A higher ROC-AUC indicates a better ability to separate positive and negative classes, making it particularly useful in applications where balancing true positives and false positives is critical, such as fraud detection or medical diagnosis. [72]

Regression problems assess model performance with metrics that evaluate the difference between predicted and actual values. Mean absolute error (**MAE**), provides a straightforward measure of average prediction accuracy, while mean squared error (**MSE**), penalizes larger errors more heavily, making it sensitive to outliers. The root mean squared error (**RMSE**), is the square root of MSE and restores the metric to the original units of the target variable, making it interpretable. Additionally, the **R-squared** value, offers insight into how well the model explains the variability of the target variable, representing the proportion of variance captured by the model [73].

2.3 Heart Failure Prediction with Machine Learning

ML has proven highly valuable for predicting heart failure, aiding in early diagnosis, risk stratification, and personalized treatment [74–76]. This section delves into the specific types of data used in heart failure prediction, the common ML models applied, and the evaluation metrics important for this domain. By understanding these components, we can better appreciate the role of ML in advancing heart disease diagnosis and management.

Types of modalities. Various types of data are used for heart failure prediction, including demographics, symptoms, laboratory tests, imaging results, and wearable sensor data. Clinical data, such as age, sex, blood pressure, and body mass index (BMI), along with patient history—including comorbidities like diabetes and hypertension—form the foundation for assessing heart disease risk. Continuous data from wearable devices, such as heart rate and ECG readings, enable real-time monitoring and early detection of cardiovascular stress or arrhythmias. Key features of the ECG, including the P-wave, T-wave, RR interval and the QRS complex, can be useful for predicting heart failure. Moreover, machine learning (ML) can be applied directly to ECG readings [77]. Imaging data from echocardiograms and cardiac MRI provide detailed insights into heart structure and function. Metrics like ejection fraction and left ventricular hypertrophy are particularly crucial for determining the degree of heart strain and dysfunction. Furthermore, applying machine learning directly to these images is also possible [78].

Machine Learning models employed. A variety of machine learning models have been applied to heart failure prediction [77], each suited to different types of data and specific prediction goals. Logistic

regression, for example, is commonly used for its interpretability, particularly in binary classification tasks where the presence or absence of a specific heart condition is being predicted [79]. Decision trees, along with ensemble methods like Random Forests and Gradient Boosting, capture complex, non-linear relationships in data and work well with diverse clinical modalities [80, 81]. Support Vector Machines (SVM) are also used, particularly in high-dimensional spaces and data contexts with a limited number of available observations, which are common in clinical research, offering strong performance where clear boundaries between classes are essential [82]. Neural networks, especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are powerful tools for handling imaging and time-series data [83, 84]. CNNs excel at identifying intricate patterns in heart imaging, while RNNs capture temporal patterns, making them ideal for longitudinal data from wearable sensors.

Feature Engineering. Feature engineering plays a critical role in developing effective machine learning models for heart failure prediction. By transforming raw data into meaningful features, it enhances the model's ability to identify patterns and relationships. Commonly engineered features include risk scores derived from patient demographics, comorbidities, and vital signs, as well as time-series features like heart rate variability or trends in blood pressure measurements. For imaging data, advanced techniques like extracting texture, shape, and structural features from echocardiograms or cardiac MRI scans provide deeper insights into heart function [85]. Additionally, signal processing techniques applied to ECG data enable the extraction of critical wave features such as the P-wave, T-wave, RR interval, and QRS complex [86]. Dimensionality reduction methods, such as principal component analysis (PCA), and feature selection techniques are often employed to handle the high dimensionality of data, ensuring that the most predictive attributes are retained. Proper feature engineering not only improves model performance but also ensures interpretability, techniques such as SHAP (Shapley Additive Explanations) help by providing insights into the importance of individual features [87], helping clinicians make data-driven decisions in the diagnosis and management of heart failure.

Spefific Metrics. In the heart failure prediction context, additional metrics are often used to evaluate the performance and clinical utility of predictive models. One such metric is the Number Needed to Screen (NNS), which represents the number of patients at risk of heart failure who need to undergo further screening examinations to identify one individual with confirmed HF. Another important metric is the Hazard Ratio (HR), which measures the relative risk of an event, such as the onset of HF, occurring in one group compared to another over a specified period. These metrics provide valuable insights into the effectiveness and practicality of screening and prediction models, supporting clinicians in decision-making and resource allocation.

3 Related Work

Predictive risk scores for new-onset atrial fibrillation (AF) have been developed since 2009 [88]. Section 3.1 briefly introduces the classic calculators of AF risk, while the subsequent sections 3.2 and 3.3 explore machine learning approaches developed in the absence and presence of ECG data, respectively.

3.1 Classical risk calculators of AF

Several classical risk score calculators have been developed to assess the risk of AF onset, its complications, and guide treatment decisions. These scores are heuristic approaches or rule-based systems which generally incorporate readily available clinical parameters to enable stratified risk assessment. These models are recognized for their simplicity and accessibility, as they do not require advanced computational tools or specialized biomarkers.

The CHADS$_2$ score is the primary risk stratification scheme and made part of the 2006 American College of Cardiology/American Heart Association/European Society of Cardiology (ACC/AHA/ESC) guidelines for nonvalvular AF [89]. The CHADS$_2$ score was originally developed to predict the risk of stroke in AF patients, however, it was later found to be capable of predicting new onset AF [90].

This score assigns points based on the following risk factors: C - congestive heart failure (1 point), H - hypertension (1 point), A - age (1 point), D - diabetes mellitus (1 point), and S - prior stroke or transient ischemic attack (2 points). This score's simplicity made it widely adopted, however, it did not account for certain risk factors, which led to the development of more refined scores.

The CHA_2DS_2-VASc score emerged as an improvement over $CHADS_2$ and has been widely used since 2010 [91]. It provides a more comprehensive assessment of stroke risk in AF patients by incorporating additional risk factors, such as: V - prior vascular disease, A - age between 65 and 74 years, and Sc - sex category. This refinement significantly improved risk stratification, especially for patients at lower or intermediate risk [92]. CHA_2DS_2-VASc score, like $CHADS_2$ score, was developed to predict risk of stroke in AF patients, but was also later found to be capable of predicting new onset AF [93].

In 2009, the Framingham Heart Study (FHS) score was the first model developed to actually predict risk of developing AF [88]. Using data from the FHS, a long-term population based study that tracked cardiovascular and other health conditions, FHS score predicts the 10-year risk of developing AF. The FHS score is based on the following factors: age, sex, body mass index (BMI), systolic blood pressure, treatment for hypertension, presence of significant cardiac murmur, history of heart failure, and finally, the PR interval which is derived from the ECG.

The Atherosclerosis Risk in Communities (ARIC) score is another risk prediction model that was developed to estimate the risk of developing AF in a community-based population [94]. The ARIC score also predicts the 10-year risk of developing AF, and takes into consideration other clinical available components such as race, height, smoking status and history of coronary heart disease(CHD), and other ECG features, such as left ventricular hypertrophy (LVH) and left atrial enlargement (LAE).

In 2013, Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE-AF) score was developed to predict 5-year risk of developing new onset AF [95], using data from FHS, ARIC and Cardiovascular Health Study (CHS), three large cohorts in the United States, and validated on data from the Reykjavik study (AGES) and the Rotterdam Study (RS). The CHARGE-AF score introduced diastolic blood pressure and myocardial infarction as new risk factors. Furthermore, variables from the electrocardiogram were included but did not improve overall model discrimination.

In 2016, the HATCH score, reported useful predicting factors of progression from paroxysmal AF to persistent AF [96], was also used to predict new-onset AF in a Taywan cohort [97]. Ultimately HATCH score was also capable of estimate the individual risk of new-onset AF for patients with different comorbidities.

The Maccabi Healthcare Services (MHS) score was developed in 2018, using a multivariable cox proportional hazards model, to estimate effects of risk factors in the derivation cohort, and to derive a risk equation [98]. The final models included the following variables: age, sex, BMI, history of treated hypertension, systolic blood pressure, chronic lung disease, history of myocardial infarction, history of peripheral arterial disease, heart failure and history of an inflammatory disease. The MHS score is a simple score for the prediction of 10-year risk for AF providing adequate discrimination.

In 2019, C_2HEST score was developed to predict new-onset AF using data from the Chinese Yunnan Insurance Database and validated using data from the Korean National Health Insurance Service [99]. The C_2HEST score is calculated by assigning points to the following risk factors: C_2: coronary artery disease (CAD) /chronic obstructive pulmonary disease (COPD) (1 point each); H: hypertension (1 point); E: elderly (age $\geq$ 75 years, 2 points); S: systolic HF (2 points); and T: thyroid disease (hyperthyroidism, 1 point). C_2HEST score was concluded as a simple clinical tool to assess the individual risk of developing AF in the Asian population without SHD. More recently, C_2HEST score was also concluded to be used to predict new onset AF in primary and secondary prevention patients, and in patients across different countries [100].

Several other models were developed during the 2010s such as the Women's Health Study (WSH) score [101], a model developed on a Japanese cohort [102], the Shandong multi-center health check-up study [103], the EHR-AF [104], and more recently the $HARMS_2$-AF score [105].

In 2020, a meta-analysis compared some of these classical models for incident AF risk [3], and only three models (CHARGEAF, FHS, CHA2DS2-VASc) yielded significant overall discrimination capacity for AF incidence at any follow-up duration and with any calibration despite high heterogeneity, and only CHARGE-AF showed superior discrimination with a uniform prediction window [106]. CHARGE-AF

appeared most suitable for primary screening purposes in terms of performance and applicability in older community cohorts of predominantly European descent. Similar results were reported by another systematic review that compared the efficacy of risk models to predict AF [107].

Overall these scores have shown appropriate model discrimination for the prediction of incident AF (AUC, generally ranging between 0.65-0.75) [106] and are valuable for their interpretability and ease of use in clinical settings. However, their reliance on predefined variables and linear associations can limit their ability to capture the complex and multifactorial nature of AF risk. This has driven the growing interest in ML-based approaches, which leverage large datasets and non-linear relationships to enhance prediction accuracy and identify novel risk factors [108, 109].

3.2 Non-ECG Clinical ML Approaches

A study by the University of Colorado Health Systems analyzed data from more than 2 million individuals, of whom approximately 28,000 (1.2%) developed incident atrial fibrillation (AF) during a designated 6-month period [110]. A machine learning model, using the 200 most common electronic health record features, including age and sex, and incorporating random oversampling with a fully connected single-layer neural network, achieved optimal predictive performance. The model yielded an AUC of 0.800. However, its performance was only slightly superior to a more straightforward logistic regression model based on established clinical risk factors for AF, which achieved an AUC of 0.794.

In 2019, an analysis of 2,994,837 individuals (3.2% with atrial fibrillation, AF) from the Clinical Practice Research Datalink (CPRD) identified time-varying neural networks as the most effective predictive model [79]. This model achieved an AUC of 0.827, compared to 0.725 for the CHARGE-AF model. Furthermore, it demonstrated the number of patients needed for screening (NNS) of 9 patients, compared to 13 for CHARGE-AF, at 75% sensitivity. The time-varying neural network confirmed established baseline risk factors such as age, prior cardiovascular disease, and antihypertensive medication use, while also uncovering novel time-sensitive predictors. These included the proximity of cardiovascular events, body mass index (both levels and changes), pulse pressure, and the frequency of blood pressure measurements. By integrating both known and novel predictors, this machine learning model significantly outperformed traditional AF risk models, offering enhanced predictive accuracy and a broader understanding of AF risk factors.

In 2020, this model was further validated using data from UK patients in the Whole Systems Integrated Care (WSIC) DISCOVER dataset [111]. Of nearly 2.5 million patients in the dataset, the algorithm identified around 600,000 individuals as eligible for risk assessment. Among these, 3.0% (17,880 patients) were diagnosed with atrial fibrillation (AF) by the study's end. The model achieved an area under the receiver operating characteristic curve (AUC) of 0.87 during validation, compared to 0.83 during its development phase. The number needed to screen (NNS) remained consistent with the CPRD cohort at nine patients. For patients over 30 years old, the algorithm correctly identified 99.1% of individuals without AF (negative predictive value, NPV) and 75.0% of true AF cases (sensitivity). Among those aged over 65 years (n = 117,965), the NPV was 96.7%, with a sensitivity of 91.8%, demonstrating strong predictive performance across age groups.

Additionally, the PULsE-AI trial [112], conducted from June 2019 to February 2021, further assessed the effectiveness of this former machine learning risk-prediction algorithm in conjunction with diagnostic testing for identifying undiagnosed atrial fibrillation (AF) in primary care in England. Eligible participants (aged ≥30 years without AF diagnosis; n=23 745) from six general practices in England were randomized into intervention and control arms. Intervention arm participants, identified by the algorithm as high risk of undiagnosed AF (n=944), were invited for diagnostic testing (n=256 consented); those who did not accept the invitation, and all control arm participants, were managed routinely. The primary endpoint was the proportion of AF, atrial flutter, and fast atrial tachycardia diagnoses during the trial in high-risk participants. Atrial fibrillation and related arrhythmias were diagnosed in 5.63% and 4.93% of high-risk participants in intervention and control arms, respectively. Among intervention arm participants who underwent diagnostic testing (28.1%), 9.41% received AF and related arrhythmia diagnoses [vs. 4.93% (control)]. This showed that the AF-risk-prediction algorithm was effective in identifying participants at high risk of undiagnosed AF. It concluded that the AF risk-prediction algorithm

may be an effective tool in narrowing the population at high risk of undiagnosed AF who should undergo diagnostic testing.

Recently, in 2023, another model called FIND-AF was developed to predict the risk of incident atrial fibrillation (AF) within 6 months [113]. The model was built using primary care data from over 2 million individuals in the UK Clinical Practice Research Datalink-GOLD dataset, of which approximately 7,000 developed AF within the 6-month period. In the test set, FIND-AF achieved an area under the receiver operating characteristic curve (AUC) of 0.824, outperforming CHA_2DS_2-VASc (AUC = 0.784) and C_2HEST (AUC = 0.757). The cohort with higher predicted risk had a 20-fold higher 6-month incidence rate of AF compared to the lower predicted risk cohort. Additionally, the higher risk group demonstrated a significantly greater long-term hazard for AF, with a hazard ratio (HR) of 8.75.

3.3 ML Approaches with ECG data

A systematic review published in 2020 analyzed and compared 12 studies on predicting atrial fibrillation (AF) using artificial intelligence (AI) and electrocardiograms (ECGs) [77]. The findings indicate that most studies focused solely on predicting AF cases, often overlooking other cardiovascular conditions, resulting in predominantly binary classification systems. ECG signals of 300 seconds (5 minutes) were commonly used, though extending the signal length did not consistently improve model accuracy. Key features extracted from the ECG data included the standard deviation and mean of RR intervals, low-frequency band power, and sample entropy. Noise removal and QRS complex detection are the most frequently employed preprocessing techniques, facilitating the extraction of RR interval-related features.

Support vector machines (SVMs) and convolutional neural networks (CNNs) are the most used methods, with simpler SVM approaches often outperforming deep learning models in accuracy. Models based on machine learning generally achieved higher accuracy rates and the Atrial Fibrillation Prediction Database was the primary data source for the three most accurate models. The best results obtained using a mixture of experts model, followed by SVM implementations [114] who got an overall of 0.982 accuracy, 1 sensitivity and 0.965 specificity with a data split of 47/53. The combination of features such as low-frequency band power, Standard Deviation 2, and sample entropy, alongside the use of 300-second ECG signals, contributed to superior prediction performance.

However, most of these studies rely on datasets with a limited number of participants, including the Atrial Fibrillation Prediction Database, which contains data from only 53 individuals. Out of the twelve studies analyzed, only two utilized datasets with a substantial number of participants. One notable example is a 2016 study using the China Kadoorie Biobank dataset, which included approximately 24,000 participants [115]. In this study, 10-second ECG recordings were analyzed, and support vector machines (SVMs) were employed, achieving an accuracy of 0.756 and an AUC of 0.83. The second study with a large participant cohort utilized data from the Mayo Clinic ECG Laboratory, encompassing over 125,000 individuals [116]. In this study, 10-second ECG recordings were analyzed, and convolutional neural networks (CNNs) were employed, achieving an accuracy of 0.833, a sensitivity of 0.823, and an AUC of 0.9.

A recent study integrated a clinically developed ECG-AI model, a convolutional neural network (CNN) designed to predict 5-year atrial fibrillation (AF)-free survival using input from 10 seconds 12-lead ECGs [117]. The ECG-AI model achieved an AUC of 0.823, outperforming the CHARGE-AF model, which achieved an AUC of 0.802. By combining ECG-AI and CHARGE-AF, the researchers developed the CH-AI model, which achieved an improved AUC of 0.838. Despite these advancements, the study concluded that AI analysis of 12-lead ECGs offers predictive performance comparable to clinical risk factor models for incident AF, with the two approaches being complementary.

4 Dataset

4.1 Description of the dataset

The data comes from anonymised electronic health records (EHRs) of patients followed at the Unidade Local de Saúde de Matosinhos (ULSM). ULSM is a large healthcare institution that includes 14 primary care centers and a hospital providing secondary and tertiary care services to the region of Matosinhos, reflecting the activity of over 1,000 doctors from various specialties. This study was approved by the Ethical Committee and Data Protection Officer of ULSM (translated from Comissão de Ética para a Saúde da Unidade Local de Saúde de Matosinhos). The available dataset covers a 10-year period, from 1 January 2015 to 31 December 2024, for patients aged over 40.

In Table 1, we present some key features of the dataset. In addition to the features shown in the table, there are several binary features representing comorbidities that are not listed. These include: myocardial infarction/unstable angina, stable angina, coronary surgery, cardiac surgery, percutaneous coronary intervention, carotid artery disease, carotid endarterectomy, carotid stent implantation, stroke, peripheral arterial disease, peripheral revascularization surgery, amputation, hypertension, dyslipidemia, type 1 diabetes mellitus, type 2 diabetes mellitus, valvular heart disease, chronic obstructive pulmonary disease (COPD), chronic kidney disease on renal replacement therapy, obstructive sleep apnea, moderate or severe aortic valvular disease, left atrial dilation, and left ventricular dilation.

The dataset also includes binary variables indicating whether a patient is currently on any of the following medications: loop diuretics, other diuretics, nitrates, beta blockers, angiotensin-converting enzyme inhibitors (ACE inhibitors), angiotensin receptor blockers (ARBs), mineralocorticoid receptor antagonists (MRAs), angiotensin receptor blocker-neprilysin inhibitor (ARNI), ivabradine, digoxin, antiplatelets, anticoagulants, calcium channel blockers, statins, metformin, DPP-IV inhibitors, insulin, SGLT2 inhibitors, sulfonylureas, and GLP-1 agonists.

If additional access to ECG data is available, features such as the RR interval, P-wave, T-wave, and QRS complex can be extracted to complement the existing data. Furthermore, if access to the ECG reports is also available, additional features could be extracted using natural language processing techniques.

Many of the features in the dataset are recorded with a longitudinal perspective, particularly those that vary over time, with entries taken every six months, each associated with a specific date, leading to the presence of multiple records for each feature. The targets representing atrial fibrillation and associated risks are also longitudinal, which enables a possibility of working with multiple time horizons for the target predictive problems.

Table 1: Some selected variables of the USLM dataset. Binary features representing comorbidities and dispensed medications are not listed for simplicity sake.

Feature Name	Description	Data Type	Units	Possible Values
BirthDate	date of birth	datetime	N/A	$\geq$ 01-01-1975
Sex	sex	binary	N/A	[male, female]
CardiacFreq	patient's cardiac frequency	integer	bmp	[20, 180]
Weight	patient's weight	integer	kg	N/A
Height	patient's height	integer	cm	[130, 200]
BMI	patient's body mass index	integer	kg/m^2	N/A
Obesity	whether the patient is considered obese	binary	N/A	[yes, no]
SisBP	patient's systolic blood pressure	integer	mmHg	[60, 180]
DiastBP	patient's diastolic blood pressure	integer	mmHg	[20, 120]
AbCirc	patient's abdominal circumference	integer	cm	[30, 300]
Smoker	smoking status	categorical	N/A	[non-smoker, current smoker, former smoker]
Alcohol	daily alcohol consume	categorical	N/A	[$\leq$ 1 drink, 1 < drinks $\leq$ 2, > 2 drinks]
Cancer	neoplasia	categorical	N/A	[active, not active (>5 years), not diagnosed]
QuimioRadio	whether the patient is currently doing chemotherapy or radiotherapy and the type	categorical	N/A	[chemotherapy, radiotherapy thorax, quimio and radio thorax, none]
Thyroid	whether the patient has a history of thyroid disease and what type	categorical	N/A	[hypothyroidism, hyperthyroidism, none]
EF	what is the patient's type of ejection fraction	categorical	N/A	[reduced, moderately reduced, preserved]
MitralSteno	whether the patient has a history or is currently experiencing moderate or severe mitral valve disease and what type	categorical	N/A	[no, stenosis, insufficiency]
Pacemaker	whether the patient is currently using any medical device – Pacemaker/ICD/CRT	binary	N/A	[yes, no]
FAFlutter (target)	whether the patient has a history or is currently experiencing atrial fibrillation or atrial flutter	binary	N/A	[yes, no]

4.2 Exploratory Data Analysis (EDA) of Public Heart Disease Prediction Data

Access to the target ULSM data is not yet provided. As such, a complementary EDA is presented here using a synergistic publicly available dataset, with the goal of identifying similar insights and preparing the pipeline for when the ULSM data becomes available. This dataset is sourced from the Heart Disease Prediction dataset on kaggle accessible via https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction, originally available in the University of California Irvine's Machine Learning Repository, accessible on the following url: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

4.2.1 Dataset Description and Analysis Overview

This dataset contains cardiological features and aims to support the study and prediction of heart failure. It consists of 270 observations and 14 features, with no missing values. A detailed summary of the dataset features, including their descriptions, data types, units, and possible values, is presented in Table 2.

To better understand the data distribution and relationships are conducted. The distribution of numerical and categorical features are overlaid, as shown in Figure 3, with the absence of heart failure represented in blue and the presence in orange. Features such as Max HR, Chest pain type, ST depression, Slope of ST, Number of vessels fluro, and Thallium exhibit clear separations between the classes and are likely significant predictors of heart disease. However, features like BP and Cholesterol show significant overlap, indicating they might be less predictive for distinguishing between the two classes.

Binary categorical data is filtered separately by heart failure absence in Figure 4 and presence in Figure 5. Sex and Exercise Angina appear to show clear distinctions between the two classes and may serve as strong predictors of heart disease. FBS over 120 does not seem to be a significant variable, as it is heavily skewed toward 0 in both classes. ECG Results shows some differences between the classes, with abnormal results (2) more frequent in individuals with heart disease, but the distinction is not as strong as for other variables.

The target variable's class distribution is visualized in Figure 6, confirming that the dataset exhibits fairly balanced class proportions, with with 55% absence and 45% presence of heart failure.

Finally, A correlation matrix in Figure 7 highlights the relationships between numerical and encoded categorical features. The strongest correlations with Heart Disease are observed for Chest pain type, Exercise angina, ST depression, Thallium, Number of vessels fluro, and Max HR, indicating these features are likely important predictors of heart disease. Conversely, FBS over 120 shows the weakest correlation (-0.016), suggesting it may be less relevant in this dataset. As expected, Slope of ST and ST depression exhibit a moderately high correlation (0.61), which may indicate some redundancy. Features such as BP, Cholesterol, FBS over 120, and ECG results have weaker correlations with other features, indicating they are loosely dependent in this dataset.

4.2.2 Logistic Regression Analysis

To evaluate the relationship between each feature and the target variable, logistic regression is applied. This analysis highlights the sigmoid behavior and potential inflection points for each feature in relation to the presence of heart disease, as shown in Figure 8. For example, the plot of Max HR versus Heart Disease exhibits a downward sigmoid curve, with an inflection point at 144.4 bpm. A similar pattern is observed in ST depression, indicating their importance in predicting heart disease.

Table 2: Variables of the Heart Disease Prediction dataset.

Feature Name	Description	Data Type	Units	Possible Values
Age	age	integer	years	[25, 100]
Sex	sex	binary	N/A	[male, female]
Chest pain type	level of chest pain	categorical	N/A	[none, low, medium, high]
BP	blood pressure	integer	mmHg	[20, 180]
Cholesterol	cholesterol	integer	mg/dL	N/A
FBS over 120	fasting blood suger over 120	binary	N/A	[yes, no]
ECG results	results of the electrocardiogram	binary	N/A	[normal, probability of left ventricular hypertrophy]
Max HR	maximum heart rate	integer	bpm	N/A
Exercise Angina	angina/chest pain while doing exercise	binary	years	[yes, no]
ST depression	depression of the ST segment in the electrocardiogram	numeric	mm	[0,10]
Slope of ST	slope of the ST segment	categorical	N/A	[upsloping, horizontal, downsloping]
Number of vessels fluro	number of vessels with blockages or abnormalities	integer	N/A	[1, 4]
Thallium	thallium uptake in the thallium stress test	numeric	N/A	[1, 10]
Heart Disease (target)	heart disease	binary	N/A	[yes, no]

4.2.3 Model Training and Evaluation

The dataset was split into training (70%) and testing (30%) subsets. A random forest model was trained on the dataset to predict heart failure presence. The following key results are observed:

1. The model achieved an accuracy of 0.81 (SD = 0.09), a recall of 0.73 (SD = 0.08), a precision of 0.80 (SD = 0.12), and an F1 score of 0.76 (SD = 0.16). These metrics indicate that the model performs well overall, with some variability observed across evaluation folds.

2. The ROC Curve in Figure 9 demonstrates the model's performance at various classification thresholds, with an AUC score of 0.91. This shows the model has excellent discriminatory power, effectively distinguishing between positive and negative classes.

3. The feature importance plot in Figure 10 highlights the relative importance of each feature in the model, identifying Thallium, Max HR and Chest pain type as the most influential predictors.

5   Solution

To define a solution to the target problem, we need to keep in mind our two initial objectives, namely, (1) Develop a predictive model capable of predicting new onset AF and related events, and (2) Design an interface to integrate the predictive model, assisting healthcare professionals with a well-situated clinical diagnosis. The solution identifies the major steps undertaken at three stages: preprocessing (section 5.1), model development (section 5.2), and medical interfaces for decision support (section 5.3).

5.1   Data Preprocessing

Effective data preprocessing is a critical step in the development of a robust machine learning model. The preprocessing phase ensures that the input data is clean, consistent, and suitable for analysis. This section outlines the key preprocessing steps that will be applied to the available data, including feature engineering, handling missing values, normalization, outlier detection and treatment, variable encoding, class imbalance, and dimensionality reduction.

Sampling and segmentation. In longitudinal studies, careful sampling is crucial to ensure that temporal patterns are well-represented and predictions are accurate. The individuals in this study are sampled at regular intervals, specifically every 6 months. This sampling frequency ensures that sufficient data points are captured to observe and analyze the evolution of AF and related conditions over time, while avoiding the problem of sparse or overly dense data. Additionally, the sampling criteria are considered along two distinct axes: the transversal axis, which addresses strata imbalances, and the longitudinal axis, which ensures adequate user-specific observability.

Sampling is done based on the presence of AF or specific conditions, allowing for targeted analysis of those affected by AF and its comorbidities. This approach facilitates a deeper understanding of how AF progresses and interacts with other factors over time.

Furthermore, stratified sampling may be employed to ensure that different population subgroups are adequately represented. This approach helps to mitigate any potential biases in the dataset, allowing for more robust and generalizable findings.

Segmentation in longitudinal data involves breaking the data into windows of specific durations to analyze temporal trends and capture the evolution of key variables. The choice of window size is an important factor in determining the level of granularity for both feature extraction and prediction. Shorter windows provide higher temporal resolution but may lead to increased noise, while longer windows may smooth out relevant fluctuations but capture broader trends.

For this study, we have access to 10 years of data, which provides a rich temporal context for analyzing the progression of atrial fibrillation (AF) and its related comorbidities. This extensive dataset allows for the testing of different historical window sizes, ranging from 1 to 5 years. These historical windows will serve as the historical context for each individual, providing information about their condition over varying periods of time. By experimenting with different window lengths, we can assess how the duration of historical data impacts the predictive performance of the models and the relevance of long-term trends in AF progression.

The prediction horizon is another crucial aspect, which defines the time period for which we aim to forecast future outcomes. The prediction windows range from 1 year to 5 years, as most known models predict 5 year risk of developing AF. These horizons align with typical clinical follow-up schedules and represent the practical time frame within which clinicians are interested in making decisions.

In addition to these considerations, special attention is given to temporal alignment of the data, ensuring that the timestamps of different screening events or tests are consistent across individuals. This alignment is crucial for extracting reliable features, such as annual slopes of biometric indicators, which can provide insight into the trajectory of AF and its associated conditions.

Feature Engineering. Feature engineering plays a pivotal role in improving model accuracy by transforming raw data into meaningful inputs for machine learning algorithms. Derived variables, such as

cumulative risk factors, will be created to better capture complex relationships. Additionally, interactions between variables will be explored to uncover potentially significant predictors.

Special attention will be given to temporal alignment challenges and longitudinal data analysis. For instance, the slope of cardiological features will be calculated by analyzing their changes across different time intervals, providing insights into trends over time. Moving averages within specified time windows will also be employed to smooth short-term fluctuations and emphasize long-term patterns.

In the context of ECG data, feature engineering is an essential first step, involving the extraction of features such as RR intervals, QRS complexes, P-waves, and T-waves. These features are critical for understanding the heart's electrical activity and detecting potential abnormalities.

Handling Missing Data. The dataset may contain missing or incomplete values for certain clinical variables. These missing values can be categorized into three levels:

1. **Missing exams at specific time points:** Some clinical examinations may be missing for certain individuals at specific time points. In these cases, imputation can be performed using the individual's historical data, leveraging available information from previous visits or time intervals.

2. **Missing critical exams:** Certain crucial examinations, such as ECGs, may be missing entirely. The absence of these critical exams could necessitate the development of specialized models to predict outcomes in the presence or absence of such exams.

3. **Missing clinical history:** In some cases, the dataset may lack a full clinical history for an individual, with only recent screenings or limited data available. This could require training models that account for varying lengths of historical data.

Outlier Detection and Treatment. Clinical datasets often contain outliers that may skew the analysis and affect model performance. Outlier detection will be addressed at two levels:

1. **Individual observations:** Some individuals may exhibit multiple deviant characteristics, such as those with several comorbidities. These cases can disproportionately influence the learning process and might require exclusion, specialized treatment, or the development of predictive models tailored to this specific population stratum.

2. **Deviant features:** Outliers in specific features will be identified using methods such as interquartile range (IQR) analysis and z-scores. Depending on the context, these outliers will either be removed or replaced using robust statistical techniques to minimize their impact on the predictive model.

This dual-level approach ensures a more nuanced handling of outliers, preserving the integrity of the data and the performance of the predictive model.

Encoding Categorical Variables. Categorical features such as diagnoses will be encoded using domain knowledge (e.g., numeric encodings of ordinal scales), depending on the nature of the variable.

Normalization. The application of transformations, such as normalization, will be explored primarily for approaches that are sensitive to scale. To ensure that the model treats all features equitably, continuous variables will be normalized or standardized. Techniques such as min-max scaling and z-score normalization will be applied and tested to evaluate their impact on model performance.

Addressing Class Imbalance. Class imbalance is a common challenge in AF prediction, where the number of cases with AF may be significantly smaller than the number of cases without it. To address this, oversampling techniques such as neural-inspired data augmentation, SMOTE (Synthetic Minority Over-sampling Technique), and undersampling methods will be tested. Additionally, class-weighting strategies will be applied to ensure that the model does not bias predictions toward the majority class.

Dimensionality Reduction. Given the high dimensionality of the dataset, techniques such as Principal Component Analysis (PCA) and mutual information-based feature selection will be explored to reduce the feature set while retaining the most informative variables. This step is crucial for improving model efficiency and interpretability, particularly when working with clinical series data with high temporal dimensionality and multivariate order datasets.

5.2 Model Development

The incremental development of predictors can be described in four sequential approaches, each focused on a different aspect of the problem. These stages are:

5.2.1 Static Approach

In the static approach, we focus on evaluating and comparing the performance of classical machine learning models and advanced neural network architectures without explicitly accounting for the longitudinal nature of the data. Instead, each observation will be treated as independent, allowing for a straightforward analysis of model performance.

1. **Classical Models:** Initially, we will test classical machine learning models, including random forests and gradient boosting methods. These models are well-suited for for point-wise clinical data, generally showing robust performance in many prediction tasks. Additionally, simple neural networks will be included in this phase as a baseline for comparison with more advanced neural architectures.

2. **Advanced Neural Networks:** Following the evaluation of classical models, we will test more sophisticated neural network architectures tailored to the complexities of our dataset. This includes:

   - **Mixed-variable models:** Neural networks designed to integrate and process different types of variables, such as categorical and continuous data, within the same framework.
   - **Multi-Layer Perceptrons (MLPs):** Fully connected feedforward neural networks that can model non-linear relationships in the data effectively.
   - **Graph Neural Networks (GNNs):** These architectures will be explored to leverage relational data or potential network structures within the dataset, providing a tailored approach to capture rich interactions among features.

This approach serves as a baseline analysis, disregarding temporal dependencies or trends in the data. The results from this phase will provide valuable insights into the feasibility and performance of various model types in predicting atrial fibrillation and related events based solely on static features.

5.2.2 Longitudinal Approach

In the longitudinal stance, the focus shifts to leveraging the temporal dimension of the dataset to capture trends, patterns, and dependencies over time. This approach aims to extract meaningful insights from the data's sequential nature, allowing the models to make predictions based on temporal evolution.

To map the longitudinal view of the data, the feature extraction techniques described in the previous section are essential to represent temporal information effectively.

Depending on the extent and richness of the temporal data available, neural networks designed for either static or longitudinal variables will be tested. Models capable of handling temporal sequences, such as recurrent neural networks (RNNs) or temporal convolutional networks (TCNs), may be explored for their ability to capture sequential dependencies.

5.2.3 ECG Approach

The rich content contained in ECG data will be explored using different modeling techniques:

1. **Classical Characteristics:** The first approach involves making predictions using the clinical interpretable and well-established features extracted from the ECG signal during the data preprocessing phase. This approach is particularly useful for generating results that are both explainable and aligned with existing clinical understanding.

2. **Representation Learning:** Another approach to tap into the rich ECG content is to leverage representation learning methods, such as autoencoders, to automatically learn latent representations from the ECG data. This allows the model to capture complex, high-dimensional patterns in the signal that may not be apparent through manual feature extraction.

3. **Direct Signal Processing:** A more data-driven approach involves designing architectures that process the raw ECG signal directly. This strategy eliminates the need for feature engineering, enabling the models to learn patterns and dependencies directly from the waveform.

   Each of these approaches provides a unique perspective on the ECG data, offering complementary strategies to explore and utilize the signal's potential for predicting AF and related events.

5.2.4 Multi-output approach

Atrial fibrillation is often accompanied by other cardiovascular and systemic conditions, forming a complex network of comorbidities that interact with its progression and clinical manifestation. Therefore, the prediction of AF alone may not fully capture the broader clinical context required for effective diagnosis and intervention. To address this, a multi-output prediction approach is proposed, enabling the simultaneous prediction of AF and related conditions or events. Here, we are going to test different target groups to explore the relationships between AF and its comorbidities. These target groups will include combinations of AF with common cardiovascular diseases, such as myocardial infarction and angina, as well as systemic conditions like diabetes and hypertension. By tailoring models to predict specific sets of related conditions, we aim to evaluate how including additional targets influences the model's performance and its ability to capture shared patterns and interactions across these conditions.

5.3 Medical tool interface

This chapter outlines the development of the medical tool interface with a focus on usability and adaptability for healthcare professionals. The primary goal is to integrate the predictive models into clinical workflows, enabling efficient decision-making while maintaining simplicity and interpretability of the targets.

5.3.1 Key Design Principles

To ensure the interface meets the practical needs of its users, the following design principles have been prioritized:

1. **Usability Requirements:** Although the usability aspect is not the primary focus, the interface is designed to be intuitive and accessible. This involves minimizing complexity, streamlining navigation, and ensuring healthcare professionals can easily interpret outcomes.

2. **Ease of Use:** The interface is structured to display only the most critical variables required for generating predictions, reducing cognitive load for users. Simplifying interactions helps improve adoption and effectiveness in real-world scenarios.

3. **Multiple Versions:** To cater to varying user needs, two versions of the interface are provided:

- **Basic Version:** A more basic interface designed for quich and straightforward predictions, with less variables, suitable for environments where time efficiency is critical or there is only access to the most common features.
- **Extended Version:** A more comprehensive interface with additional input fields. This version offers a deeper analysis for patients with a more complete medical history.

5.3.2 Features and components

The medical tool is composed by two main components: a programmatic API and a user-friendly graphical user interface (GUI).

1. **Programmatic API.** A robust programmatic API is developed to facilitate integration into existing clinical systems and workflows. The API allows seamless interaction with the predictive models, ensuring that the functionalities can be extended and adapted as needed.

2. **Graphical User Interface (GUI).** A user-friendly GUI is implemented to provide direct access to predictions and insights. The GUI includes an input section with a minimal set of input fields to capture patient data in the basic version, with additional inputs available in the extended version for a more detailed analysis; and an output section with clear visualization of prediction outcomes, including probabilities of AF events and related conditions. Results are presented in a manner that is both informative and non-technical, ensuring that users can quickly understand and act upon the predictions. This may include graphical aids such as charts or indicators, to enhance interpretability.

6 Evaluation

Evaluation Methodology

Due to the longitudinal nature of the data, model evaluation will be conducted using cross-validation on a rolling basis, also referred to as time-series cross-validation. In this approach, the dataset is split sequentially to ensure that future data is not used to predict the past. The data is divided into multiple train-test splits, where the training set starts with the earliest observations and expands with each split, while the test set moves forward in time. This method can follow either an expanding window strategy, where the training set grows with additional data, or a sliding window strategy, where both the training and testing sets shift forward in fixed steps. For each split, the model is trained on the training set and evaluated on the test set, maintaining the temporal order.

Evaluation Metrics

Performance metrics such as accuracy (1), sensitivity (3), AUC and NNS are calculated for each split and aggregated to provide an overall assessment of the model's performance over time.

$$\text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{6}$$

$$\text{AUC} = \int_0^1 \text{TPR}(FPR) \, d(\text{FPR}) \tag{7}$$

$$\text{NNS} = \frac{1}{\text{CER}^* - \text{EER}^*} \tag{8}$$

* CER is the control event rate, which is the rate of the event occurring in the control group.
* EER is the experimental event rate, which is the rate of the event occurring in the experimental group.

For regression tasks, performance metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R^2 are also calculated for each split and aggregated to provide a comprehensive evaluation of the model's predictive accuracy over time.

$$\text{MAE (Mean Absolute Error)} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{9}$$

$$\text{MSE (Mean Squared Error)} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{10}$$

$$\text{RMSE (Root Mean Squared Error)} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{11}$$

$$\text{R-squared} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{12}$$

- y: Represents the actual or true value of the dependent variable in the dataset.

- $\hat{y}$: Represents the predicted value of the dependent variable, as generated by the model.

- $\bar{y}$: Represents the mean of the actual values of the dependent variable

In multi-class scenarios, macro, micro, or weighted averaging metrics will be taken into account, and in multi-label contexts, metrics such as hamming loss, subset accuracy and also macro, micro, or weighted averaging will be employed.

$$\text{Macro Average} = \frac{1}{n} \sum_{i=1}^{n} \text{Metric}_i \tag{13}$$

$$\text{Micro Average} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n} (\text{TP}_i + \text{FP}_i + \text{FN}_i)} \tag{14}$$

$$\text{Weighted Average} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} \tag{15}$$

$$\text{Hamming Loss} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{L} \sum_{j=1}^{L} \mathbf{1}(y_{ij} \neq \hat{y}_{ij}) \tag{16}$$

- w_i: Represents the weights applied to x-values.

- X_i: Represents the data values to be averaged.

- y_{ij}: Represents the true labels for the i-th sample and j-th class

- $\hat{y}_{ij}$: Represents predicted labels for the i-th sample and j-th class

- L: Represents the number of labels for multi-label classification

Statistical Tests

Statistical tests will be applied to compare the performance of different predictors and assess the significance of their differences.

The paired t-test is applied to compare the mean performance of two models over multiple runs or datasets. This test assumes that the differences between paired observations follow a normal distribution. It is particularly useful for evaluating whether the performance improvements observed for one model over another are statistically significant.

For cases where the assumption of normality is not met, the Wilcoxon signed-rank test serves as a non-parametric alternative to the paired t-test. This test evaluates whether the differences between paired samples are symmetrically distributed around zero. The Wilcoxon test is applied in the context of model comparisons to verify whether one predictor consistently outperforms another across multiple trials. Similar to the t-test, it tests if there is no significant difference between the models.

The sign test is another non-parametric method used to assess the directionality of differences between paired observations. It focuses solely on the number of positive and negative differences between the performance of two models, ignoring the magnitude of those differences.

In this work, these statistical tests are applied to determine whether the differences observed between predictors are statistically significant. By assessing both parametric and non-parametric stances, we ensure that the evaluation accounts for the nature of the data distribution and any potential violations of assumptions.

Baselines

To assess the performance of the proposed models, the results will be compared to the following baselines: Pair logistic regression and naïve Bayes. These models are selected as a baseline due to their simplicity, interpretability, and widespread use in predictive modeling, making them a suitable comparison for evaluating the predictive power of more complex models. By comparing the new models to logistic regression and naïve Bayes, we can better understand whether the added complexity leads to significant improvements in prediction accuracy. Additionally, the models will be compared to Charge-AF, a specialized model designed for atrial fibrillation prediction [95] (section 3.1). This comparison will provide valuable insights into the competitive performance of our models against a state-of-the-art approach specifically tailored for the same healthcare domain. By evaluating against both logistic regression and Charge-AF, we aim to provide a comprehensive understanding of the relative predictive capabilities of our models.

Knowlege Aquisition

By identifying the most important predictors, we can focus on examining the variables that contribute most significantly to the outcomes. Analyzing the relationships between these key variables and the target conditions enables us to uncover the underlying factors driving the development and progression of AF and its associated comorbidities. This process provides valuable insights into the clinical pathways involved, helping to identify potential biomarkers, risk factors, and mechanisms that influence AF onset and exacerbation. Furthermore, it opens the door to refining risk stratification models and targeting specific areas for clinical intervention, which could ultimately lead to improved patient management and personalized treatment strategies.

7 Work Schedule

In accordance with the proposed research goals, the suggested work planning is provided in Figure 2.
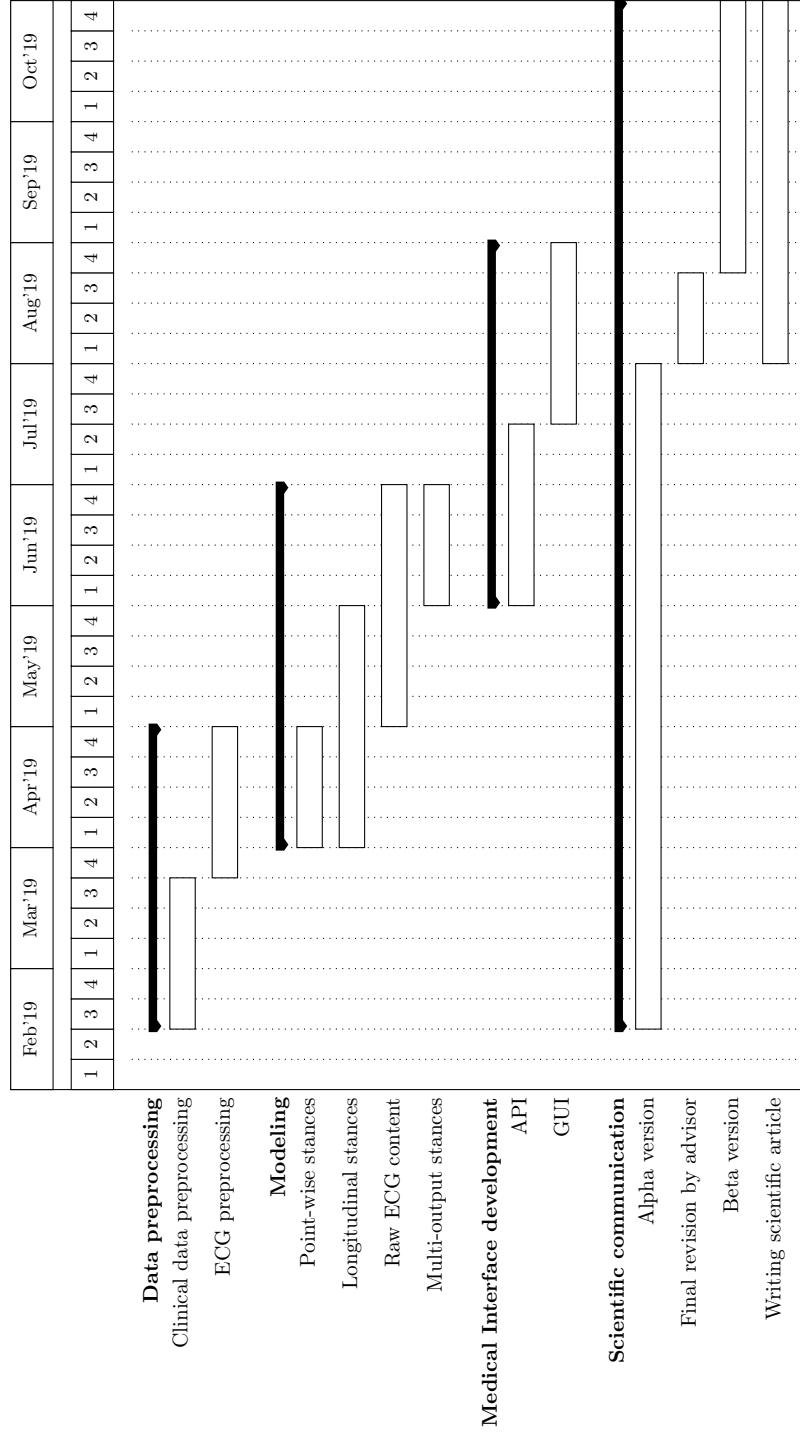

Figure 2: Planned Schedule

Bibliography

[1] V. Fuster, L. E. Rydén, D. S. Cannom, H. J. Crijns, A. B. Curtis, K. A. Ellenbogen, J. L. Halperin, G. N. Kay, J.-Y. Le Huezey, J. E. Lowe *et al.*, "2011 accf/aha/hrs focused updates incorporated into the acc/aha/esc 2006 guidelines for the management of patients with atrial fibrillation: a report of the american college of cardiology foundation/american heart association task force on practice guidelines developed in partnership with the european society of cardiology and in collaboration with the european heart rhythm association and the heart rhythm society," *Journal of the American College of Cardiology*, vol. 57, no. 11, pp. e101–e198, 2011.

[2] E. Davidson, Z. Rotenberg, I. Weinberger, J. Fuchs, and J. Agmon, "Diagnosis and characteristics of lone atrial fibrillation," *Chest*, vol. 95, no. 5, pp. 1048–1050, 1989.

[3] J. C. Himmelreich, L. Veelers, W. A. Lucassen, R. B. Schnabel, M. Rienstra, H. C. van Weert, and R. E. Harskamp, "Prediction models for atrial fibrillation applicable in the community: a systematic review and meta-analysis," *EP Europace*, vol. 22, no. 5, pp. 684–694, 2020.

[4] P. S. Jagadish and R. Kabra, "Stroke risk in atrial fibrillation: Beyond the cha 2 ds 2-vasc score," *Current cardiology reports*, vol. 21, pp. 1–9, 2019.

[5] A. S. Tseng and P. A. Noseworthy, "Prediction of atrial fibrillation using machine learning: a review," *Frontiers in Physiology*, vol. 12, p. 752317, 2021.

[6] M. Arrigo, M. Jessup, W. Mullens, N. Reza, A. M. Shah, K. Sliwa, and A. Mebazaa, "Acute heart failure," *Nature Reviews Disease Primers*, vol. 6, no. 1, p. 16, 2020.

[7] L. Wilhelmsen, H. Eriksson, K. Svärdsudd, and K. Caidahl, "Improving the detection and diagnosis of congestive heart failure," *European heart journal*, vol. 10, no. suppl_C, pp. 13–18, 1989.

[8] K. Takenaka, T. Sakamoto, K. Amano, J. Oku, K. Fujinami, T. Murakami, I. Toda, K. Kawakubo, and T. Sugimoto, "Left ventricular filling determined by doppler echocardiography in diabetes mellitus," *The American journal of cardiology*, vol. 61, no. 13, pp. 1140–1143, 1988.

[9] S. Lévy, "Factors predisposing to the development of atrial fibrillation," *Pacing and clinical electrophysiology*, vol. 20, no. 10, pp. 2670–2674, 1997.

[10] V. Mahadevan, "Anatomy of the heart," *Surgery (Oxford)*, vol. 36, no. 2, pp. 43–47, 2018.

[11] R. H. Whitaker, "Anatomy of the heart," *Medicine*, vol. 38, no. 7, pp. 333–335, 2010.

[12] D. S. Park and G. I. Fishman, "The cardiac conduction system," *Circulation*, vol. 123, no. 8, pp. 904–915, 2011.

[13] E. Boersma, N. Mercado, D. Poldermans, M. Gardien, J. Vos, and M. L. Simoons, "Acute myocardial infarction," *The Lancet*, vol. 361, no. 9360, pp. 847–858, 2003.

[14] G. K. Hansson, "Inflammation, atherosclerosis, and coronary artery disease," *New England journal of medicine*, vol. 352, no. 16, pp. 1685–1695, 2005.

[15] K. W. Schef, P. Tornvall, J. Alfredsson, E. Hagström, A. Ravn-Fischer, S. Soderberg, T. Yndigegn, and T. Jernberg, "Prevalence of angina pectoris and association with coronary atherosclerosis in a general population," *Heart*, vol. 109, no. 19, pp. 1450–1459, 2023.

[16] S. Koba and T. Hirano, "Dyslipidemia and atherosclerosis," *Nihon rinsho. Japanese journal of clinical medicine*, vol. 69, no. 1, pp. 138–143, 2011.

[17] P. Wilson, "Diabetes mellitus and coronary heart disease," *American Journal of Kidney Diseases*, vol. 32, no. 5, pp. S89–S100, 1998.

[18] L. Di Lullo, A. House, A. Gorini, A. Santoboni, D. Russo, and C. Ronco, "Chronic kidney disease and cardiovascular complications," *Heart failure reviews*, vol. 20, pp. 259–272, 2015.

[19] A. S. Hersi, "Obstructive sleep apnea and cardiac arrhythmias," *Annals of thoracic medicine*, vol. 5, no. 1, pp. 10–17, 2010.

[20] M. Kurt, J. Wang, G. Torre-Amione, and S. F. Nagueh, "Left atrial function in diastolic heart failure," *Circulation: Cardiovascular Imaging*, vol. 2, no. 1, pp. 10–15, 2009.

[21] S. Oparil, M. C. Acelajado, G. L. Bakris, D. R. Berlowitz, R. Cífková, A. F. Dominiczak, G. Grassi, J. Jordan, N. R. Poulter, A. Rodgers *et al.*, "Hypertension," *Nature reviews. Disease primers*, vol. 4, p. 18014, 2018.

[22] M. J. Klag, J. He, L. A. Mead, D. E. Ford, T. A. Pearson, and D. M. Levine, "Validity of physicians' self-reports of cardiovascular disease risk factors," *Annals of epidemiology*, vol. 3, no. 4, pp. 442–447, 1993.

[23] A. W. Haider, M. G. Larson, S. S. Franklin, and D. Levy, "Systolic blood pressure, diastolic blood pressure, and pulse pressure as predictors of risk for congestive heart failure in the framingham heart study," *Annals of internal medicine*, vol. 138, no. 1, pp. 10–16, 2003.

[24] H. H. Alhawari, S. Al-Shelleh, H. H. Alhawari, A. Al-Saudi, D. Aljbour Al-Majali, L. Al-Faris, and S. A. AlRyalat, "Blood pressure and its association with gender, body mass index, smoking, and family history among university students," *International journal of hypertension*, vol. 2018, no. 1, p. 4186496, 2018.

[25] J. K. Alexander, "Obesity and coronary heart disease," *The American journal of the medical sciences*, vol. 321, no. 4, pp. 215–224, 2001.

[26] Y. Kokubo and C. Matsumoto, "Hypertension is a risk factor for several types of heart disease: review of prospective studies," *Hypertension: from basic research to clinical practice*, pp. 419–426, 2017.

[27] A. J. Camm, G. Corbucci, and L. Padeletti, "Usefulness of continuous electrocardiographic monitoring for atrial fibrillation," *The American journal of cardiology*, vol. 110, no. 2, pp. 270–276, 2012.

[28] J. Forester, H. Bo, J. Sleigh, and J. Henderson, "Variability of rr, p wave-to-r wave, and r wave-to-t wave intervals," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 273, no. 6, pp. H2857–H2860, 1997.

[29] J. Lian, L. Wang, and D. Muessig, "A simple method to detect atrial fibrillation using rr intervals," *The American journal of cardiology*, vol. 107, no. 10, pp. 1494–1497, 2011.

[30] M. B. Simson, "Use of signals in the terminal qrs complex to identify patients with ventricular tachycardia after myocardial infarction." *Circulation*, vol. 64, no. 2, pp. 235–242, 1981.

[31] J. J. MORRIS JR, E. H. ESTES JR, R. E. Whalen, H. K. THOMPSON JR, and H. D. MCINTOSH, "P-wave analysis in valvular heart disease," *Circulation*, vol. 29, no. 2, pp. 242–252, 1964.

[32] S. M. Narayan, "T-wave alternans and the susceptibility to ventricular arrhythmias," *Journal of the American College of Cardiology*, vol. 47, no. 2, pp. 269–281, 2006.

[33] A. J. Sanfilippo, V. M. Abascal, M. Sheehan, L. B. Oertel, P. Harrigan, R. A. Hughes, and A. E. Weyman, "Atrial enlargement as a consequence of atrial fibrillation. a prospective echocardiographic study." *Circulation*, vol. 82, no. 3, pp. 792–797, 1990.

[34] J. F. Pombo, B. L. Troy, and R. O. RUSSELL JR, "Left ventricular volumes and ejection fraction by echocardiography," *Circulation*, vol. 43, no. 4, pp. 480–490, 1971.

[35] C. Fornengo, M. Antolini, S. Frea, C. Gallo, W. Grosso Marra, M. Morello, and F. Gaita, "Prediction of atrial fibrillation recurrence after cardioversion in patients with left-atrial dilation," *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 3, pp. 335–341, 2015.

[36] M. W. Bloom, B. Greenberg, T. Jaarsma, J. L. Januzzi, C. S. Lam, A. P. Maggioni, J.-N. Trochu, and J. Butler, "Heart failure with reduced ejection fraction," *Nature reviews Disease primers*, vol. 3, no. 1, pp. 1–19, 2017.

[37] R. Gramiak and P. M. Shah, "Echocardiography of the normal and diseased aortic valve," *Radiology*, vol. 96, no. 1, pp. 1–8, 1970.

[38] I. J. Amat-Santos, J. Rodés-Cabau, M. Urena, R. DeLarochellière, D. Doyle, R. Bagur, J. Villeneuve, M. Côté, L. Nombela-Franco, F. Philippon *et al.*, "Incidence, predictive factors, and prognostic value of new-onset atrial fibrillation following transcatheter aortic valve implantation," *Journal of the American College of Cardiology*, vol. 59, no. 2, pp. 178–188, 2012.

[39] E. BRAUNWALD and W. C. AWE, "The syndrome of severe mitral regurgitation with normal left atrial pressure," *Circulation*, vol. 27, no. 1, pp. 29–35, 1963.

[40] Z.-H. Zhou, *Machine learning.* Springer nature, 2021.

[41] P. Cunningham, M. Cord, and S. J. Delany, "Supervised learning," in *Machine learning techniques for multimedia: case studies on organization and retrieval.* Springer, 2008, pp. 21–49.

[42] F. Herrera, F. Charte, A. J. Rivera, M. J. Del Jesus, F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, *Multilabel classification.* Springer, 2016.

[43] H. Borchani, G. Varando, C. Bielza, and P. Larranaga, "A survey on multi-output regression," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 5, pp. 216–233, 2015.

[44] J. J. Oliver, R. A. Baxter, and C. S. Wallace, "Unsupervised learning using mml," in *ICML*, 1996, pp. 364–372.

[45] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic regression.* Springer, 2002.

[46] C. Ngufor and J. Wojtusiak, "Extreme logistic regression," *Advances in Data Analysis and Classification*, vol. 10, pp. 27–52, 2016.

[47] R. G. Brereton and G. R. Lloyd, "Support vector machines for classification and regression," *Analyst*, vol. 135, no. 2, pp. 230–267, 2010.

[48] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[49] X. Su, X. Yan, and C.-L. Tsai, "Linear regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 3, pp. 275–294, 2012.

[50] E. Ostertagová, "Modelling using polynomial regression," *Procedia engineering*, vol. 48, pp. 500–506, 2012.

[51] K. Q. Weinberger and G. Tesauro, "Metric learning for kernel regression," in *Artificial intelligence and statistics.* PMLR, 2007, pp. 612–619.

[52] S. Kohli, G. T. Godwin, and S. Urolagin, "Sales prediction using linear and knn regression," in *Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019.* Springer, 2020, pp. 321–329.

[53] Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015.

[54] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 5, p. 272, 2012.

[55] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[56] J. A. Hartigan, M. A. Wong *et al.*, "A k-means clustering algorithm," *Applied statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[57] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[58] D. Deng, "Dbscan clustering algorithm based on density," in *2020 7th international forum on electrical engineering and automation (IFEEA)*. IEEE, 2020, pp. 949–953.

[59] B. M. S. Hasan and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 20–30, 2021.

[60] T. Khanna, *Foundations of neural networks*. Addison-Wesley Longman Publishing Co., Inc., 1990.

[61] R. Kruse, S. Mostaghim, C. Borgelt, C. Braune, and M. Steinbrecher, "Multi-layer perceptrons," in *Computational intelligence: a methodological introduction*. Springer, 2022, pp. 53–124.

[62] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Transactions on Circuits and Systems*, vol. 8, no. 7, pp. 579–588, 2009.

[63] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[64] J. Wu, "Introduction to convolutional neural networks," *National Key Lab for Novel Software Technology. Nanjing University. China*, vol. 5, no. 23, p. 495, 2017.

[65] L. R. Medsker, L. Jain *et al.*, "Recurrent neural networks," *Design and Applications*, vol. 5, no. 64-67, p. 2, 2001.

[66] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[67] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "The computational limits of deep learning," *arXiv preprint arXiv:2007.05558*, vol. 10, 2020.

[68] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: Cnn, rnn, lstm, gru," *arXiv preprint arXiv:2305.17473*, 2023.

[69] Ž. Vujović *et al.*, "Classification model evaluation metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599–606, 2021.

[70] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *arXiv preprint arXiv:2008.05756*, 2020.

[71] N. A. Muhammad, A. Rehman, and U. Shoaib, "Accuracy based feature ranking metric for multi-label text classification," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, 2017.

[72] S. Narkhede, "Understanding auc-roc curve," *Towards data science*, vol. 26, no. 1, pp. 220–227, 2018.

[73] A. V. Tatachar, "Comparative assessment of regression models based on model evaluation metrics," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 9, pp. 853–860, 2021.

[74] D. K. Plati, E. E. Tripoliti, A. Bechlioulis, A. Rammos, I. Dimou, L. Lakkas, C. Watson, K. McDonald, M. Ledwidge, R. Pharithi *et al.*, "A machine learning approach for chronic heart failure diagnosis," *Diagnostics*, vol. 11, no. 10, p. 1863, 2021.

[75] D. R. Sax, D. G. Mark, J. Huang, O. Sofrygin, J. S. Rana, S. P. Collins, A. B. Storrow, D. Liu, and M. E. Reed, "Use of machine learning to develop a risk-stratification tool for emergency department patients with acute heart failure," *Annals of Emergency Medicine*, vol. 77, no. 2, pp. 237–248, 2021.

[76] D. Bertsimas, A. Orfanoudaki, and R. B. Weiner, "Personalized treatment for coronary artery disease patients: a machine learning approach," *Health Care Management Science*, vol. 23, no. 4, pp. 482–506, 2020.

[77] I. Matias, N. Garcia, S. Pirbhulal, V. Felizardo, N. Pombo, H. Zacarias, M. Sousa, and E. Zdravevski, "Prediction of atrial fibrillation using artificial intelligence on electrocardiograms: A systematic review," *Computer Science Review*, vol. 39, p. 100334, 2021.

[78] L. D. Liastuti, B. B. Siswanto, R. Sukmawan, W. Jatmiko, Y. Nursakina, R. Y. I. Putri, G. Jati, and A. A. Nur, "Detecting left heart failure in echocardiography through machine learning: A systematic review," *Reviews in Cardiovascular Medicine*, vol. 23, no. 12, p. 402, 2022.

[79] N. R. Hill, D. Ayoubkhani, P. McEwan, D. M. Sugrue, U. Farooqui, S. Lister, M. Lumley, A. Bakhai, A. T. Cohen, M. O'Neill *et al.*, "Predicting atrial fibrillation in primary care using machine learning," *PloS one*, vol. 14, no. 11, p. e0224582, 2019.

[80] M. Zabihi, A. B. Rad, A. K. Katsaggelos, S. Kiranyaz, S. Narkilahti, and M. Gabbouj, "Detection of atrial fibrillation in ecg hand-held devices using a random forest classifier," in *2017 computing in cardiology (cinc)*. IEEE, 2017, pp. 1–4.

[81] S. D. Goodfellow, A. Goodwin, R. Greer, P. C. Laussen, M. Mazwi, and D. Eytan, "Classification of atrial fibrillation using multidisciplinary features and gradient boosting," in *2017 Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.

[82] V. Gliner and Y. Yaniv, "An svm approach for identifying atrial fibrillation," *Physiological Measurement*, vol. 39, no. 9, p. 094007, 2018.

[83] M. Limam and F. Precioso, "Atrial fibrillation detection and ecg classification based on convolutional recurrent neural network," in *2017 computing in cardiology (CinC)*. IEEE, 2017, pp. 1–4.

[84] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *Journal of the American Medical Informatics Association*, vol. 24, no. 2, pp. 361–370, 2017.

[85] S. Batool, I. A. Taj, and M. Ghafoor, "Ejection fraction estimation from echocardiograms using optimal left ventricle feature extraction based on clinical methods," *Diagnostics*, vol. 13, no. 13, p. 2155, 2023.

[86] C. Bhyri, S. Hamde, and L. Waghmare, "Ecg feature extraction and disease diagnosis," *Journal of medical engineering & technology*, vol. 35, no. 6-7, pp. 354–361, 2011.

[87] C. Guan, A. Gong, Y. Zhao, C. Yin, L. Geng, L. Liu, X. Yang, J. Lu, and B. Xiao, "Interpretable machine learning model for new-onset atrial fibrillation prediction in critically ill patients: a multi-center study," *Critical Care*, vol. 28, no. 1, p. 349, 2024.

[88] R. B. Schnabel, L. M. Sullivan, D. Levy, M. J. Pencina, J. M. Massaro, R. B. D'Agostino, C. Newton-Cheh, J. F. Yamamoto, J. W. Magnani, T. M. Tadros *et al.*, "Development of a risk score for atrial fibrillation (framingham heart study): a community-based cohort study," *The Lancet*, vol. 373, no. 9665, pp. 739–745, 2009.

[89] V. Fuster, L. E. Rydén, D. S. Cannom, H. J. Crijns, A. B. Curtis, K. A. Ellenbogen, J. L. Halperin, J.-Y. Le Heuzey, G. N. Kay, J. E. Lowe *et al.*, "Acc/aha/esc 2006 guidelines for the management of patients with atrial fibrillation—executive summary: a report of the american college of cardiology/american heart association task force on practice guidelines and the european society of cardiology committee for practice guidelines (writing committee to revise the 2001 guidelines for the management of patients with atrial fibrillation) developed in collaboration with the european heart rhythm association and the heart rhythm society," *Journal of the American College of Cardiology*, vol. 48, no. 4, pp. 854–906, 2006.

[90] T.-F. Chao, C.-J. Liu, S.-J. Chen, K.-L. Wang, Y.-J. Lin, S.-L. Chang, L.-W. Lo, Y.-F. Hu, T.-C. Tuan, T.-J. Wu *et al.*, "Chads2 score and risk of new-onset atrial fibrillation: a nationwide cohort study in taiwan," *International journal of cardiology*, vol. 168, no. 2, pp. 1360–1363, 2013.

[91] D. with the Special Contribution of the European Heart Rhythm Association (EHRA), E. by the European Association for Cardio-Thoracic Surgery (EACTS), A. F. Members, A. J. Camm, P. Kirchhof, G. Y. Lip, U. Schotten, I. Savelieva, S. Ernst, I. C. Van Gelder *et al.*, "Guidelines for the management of atrial fibrillation: the task force for the management of atrial fibrillation of the european society of cardiology (esc)," *European heart journal*, vol. 31, no. 19, pp. 2369–2429, 2010.

[92] D. F. Katz, T. M. Maddox, M. Turakhia, A. Gehi, E. C. O'Brien, S. A. Lubitz, A. Turchin, G. Doros, L. Lei, P. Varosy *et al.*, "Contemporary trends in oral anticoagulant prescription in atrial fibrillation patients at low to moderate risk of stroke after guideline-recommended change in use of the chads2 to the cha2ds2-vasc score for thromboembolic risk assessment: analysis from the national cardiovascular data registry's outpatient practice innovation and clinical excellence atrial fibrillation registry," *Circulation: Cardiovascular Quality and Outcomes*, vol. 10, no. 5, p. e003476, 2017.

[93] W. Saliba, N. Gronich, O. Barnett-Griness, and G. Rennert, "Usefulness of chads2 and cha2ds2-vasc scores in the prediction of new-onset atrial fibrillation: a population-based study," *The American journal of medicine*, vol. 129, no. 8, pp. 843–849, 2016.

[94] A. M. Chamberlain, S. K. Agarwal, A. R. Folsom, E. Z. Soliman, L. E. Chambless, R. Crow, M. Ambrose, and A. Alonso, "A clinical risk score for atrial fibrillation in a biracial prospective cohort (from the atherosclerosis risk in communities [aric] study)," *The American journal of cardiology*, vol. 107, no. 1, pp. 85–91, 2011.

[95] A. Alonso, B. P. Krijthe, T. Aspelund, K. A. Stepas, M. J. Pencina, C. B. Moser, M. F. Sinner, N. Sotoodehnia, J. D. Fontes, A. C. J. Janssens *et al.*, "Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the charge-af consortium," *Journal of the American Heart Association*, vol. 2, no. 2, p. e000102, 2013.

[96] C. B. De Vos, R. Pisters, R. Nieuwlaat, M. H. Prins, R. G. Tieleman, R.-J. S. Coelen, A. C. van den Heijkant, M. A. Allessie, and H. J. Crijns, "Progression from paroxysmal to persistent atrial fibrillation: clinical correlates and prognosis," *Journal of the American College of Cardiology*, vol. 55, no. 8, pp. 725–731, 2010.

[97] K. Suenari, T.-F. Chao, C.-J. Liu, Y. Kihara, T.-J. Chen, and S.-A. Chen, "Usefulness of hatch score in the prediction of new-onset atrial fibrillation for asians," *Medicine*, vol. 96, no. 1, p. e5597, 2017.

[98] D. Aronson, V. Shalev, R. Katz, G. Chodick, and D. Mutlak, "Risk score for prediction of 10-year atrial fibrillation: a community-based study," *Thrombosis and Haemostasis*, vol. 118, no. 09, pp. 1556–1563, 2018.

[99] Y.-G. Li, D. Pastori, A. Farcomeni, P.-S. Yang, E. Jang, B. Joung, Y.-T. Wang, Y.-T. Guo, and G. Y. Lip, "A simple clinical risk score (c2hest) for predicting incident atrial fibrillation in asian subjects: derivation in 471,446 chinese subjects, with internal validation and external application in 451,199 korean subjects," *Chest*, vol. 155, no. 3, pp. 510–518, 2019.

[100] D. Pastori, D. Menichelli, Y.-G. Li, T. Brogi, F. G. Biccirè, P. Pignatelli, A. Farcomeni, and G. Y. Lip, "Usefulness of the c2hest score to predict new onset atrial fibrillation. a systematic review and meta-analysis on> 11 million subjects," *European Journal of Clinical Investigation*, p. e14293, 2024.

[101] B. M. Everett, N. R. Cook, D. Conen, D. I. Chasman, P. M. Ridker, and C. M. Albert, "Novel genetic markers improve measures of atrial fibrillation risk prediction," *European heart journal*, vol. 34, no. 29, pp. 2243–2251, 2013.

[102] R. Hamada and S. Muto, "Simple risk model and score for predicting of incident atrial fibrillation in japanese," *Journal of cardiology*, vol. 73, no. 1, pp. 65–72, 2019.

[103] L. Ding, J. Li, C. Wang, X. Li, Q. Su, G. Zhang, and F. Xue, "Incidence of atrial fibrillation and its risk prediction model based on a prospective urban han chinese cohort," *Journal of Human Hypertension*, vol. 31, no. 9, pp. 574–579, 2017.

[104] O. L. Hulme, S. Khurshid, L.-C. Weng, C. D. Anderson, E. Y. Wang, J. M. Ashburner, D. Ko, D. D. McManus, E. J. Benjamin, P. T. Ellinor *et al.*, "Development and validation of a prediction model for atrial fibrillation using electronic health records," *JACC: Clinical Electrophysiology*, vol. 5, no. 11, pp. 1331–1341, 2019.

[105] L. Segan, R. Canovas, S. Nanayakkara, D. Chieng, S. Prabhu, A. Voskoboinik, H. Sugumar, L.-H. Ling, G. Lee, J. Morton *et al.*, "New-onset atrial fibrillation prediction: the harms2-af risk score," *European heart journal*, vol. 44, no. 36, pp. 3443–3452, 2023.

[106] C. Goudis, S. Daios, F. Dimitriadis, and T. Liu, "Charge-af: a useful score for atrial fibrillation prediction?" *Current Cardiology Reviews*, vol. 19, no. 2, pp. 5–10, 2023.

[107] M. H. Poorthuis, N. R. Jones, P. Sherliker, R. Clack, G. J. de Borst, R. Clarke, S. Lewington, A. Halliday, and R. Bulbulia, "Utility of risk prediction models to detect atrial fibrillation in screened participants," *European journal of preventive cardiology*, vol. 28, no. 6, pp. 586–595, 2021.

[108] K. C. Siontis, X. Yao, J. P. Pirruccello, A. A. Philippakis, and P. A. Noseworthy, "How will machine learning inform the clinical care of atrial fibrillation?" *Circulation research*, vol. 127, no. 1, pp. 155–169, 2020.

[109] F. K. Wegner, L. Plagwitz, F. Doldi, C. Ellermann, K. Willy, J. Wolfes, S. Sandmann, J. Varghese, and L. Eckardt, "Machine learning in the detection and management of atrial fibrillation," *Clinical Research in Cardiology*, vol. 111, no. 9, pp. 1010–1017, 2022.

[110] P. Tiwari, K. L. Colborn, D. E. Smith, F. Xing, D. Ghosh, and M. A. Rosenberg, "Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation," *JAMA network open*, vol. 3, no. 1, pp. e1 919 396–e1 919 396, 2020.

[111] S. Sekelj, B. Sandler, E. Johnston, K. G. Pollock, N. R. Hill, J. Gordon, C. Tsang, S. Khan, F. S. Ng, and U. Farooqui, "Detecting undiagnosed atrial fibrillation in uk primary care: validation of a machine learning prediction algorithm in a retrospective cohort study," *European journal of preventive cardiology*, vol. 28, no. 6, pp. 598–605, 2021.

[112] N. R. Hill, L. Groves, C. Dickerson, A. Ochs, D. Pang, S. Lawton, M. Hurst, K. G. Pollock, D. M. Sugrue, C. Tsang *et al.*, "Identification of undiagnosed atrial fibrillation using a machine learning risk-prediction algorithm and diagnostic testing (pulse-ai) in primary care: a multi-centre randomized controlled trial in england," *European Heart Journal-Digital Health*, vol. 3, no. 2, pp. 195–204, 2022.

[113] R. Nadarajah, J. Wu, D. Hogg, K. Raveendra, Y. M. Nakao, K. Nakao, R. Arbel, M. Haim, D. Zahger, J. Parry *et al.*, "Prediction of short-term atrial fibrillation risk using primary care electronic health records," *Heart*, vol. 109, no. 14, pp. 1072–1079, 2023.

[114] E. Ebrahimzadeh, M. Kalantari, M. Joulani, R. S. Shahraki, F. Fayaz, and F. Ahmadi, "Prediction of paroxysmal atrial fibrillation: A machine learning based approach using combined feature vector and mixture of expert classification on hrv signal," *Computer methods and programs in biomedicine*, vol. 165, pp. 53–67, 2018.

[115] Y. Shen, Y. Yang, S. Parish, Z. Chen, R. Clarke, and D. A. Clifton, "Risk prediction for cardiovascular disease using ecg data in the china kadoorie biobank," in *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2016, pp. 2419–2422.

[116] 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.

[117] S. Khurshid, S. Friedman, C. Reeder, P. Di Achille, N. Diamant, P. Singh, L. X. Harrington, X. Wang, M. A. Al-Alusi, G. Sarma *et al.*, "Ecg-based deep learning and clinical risk factors to predict atrial fibrillation," *Circulation*, vol. 145, no. 2, pp. 122–133, 2022.

A   Complementary Images of the EDA of the Heart Disease Dataset
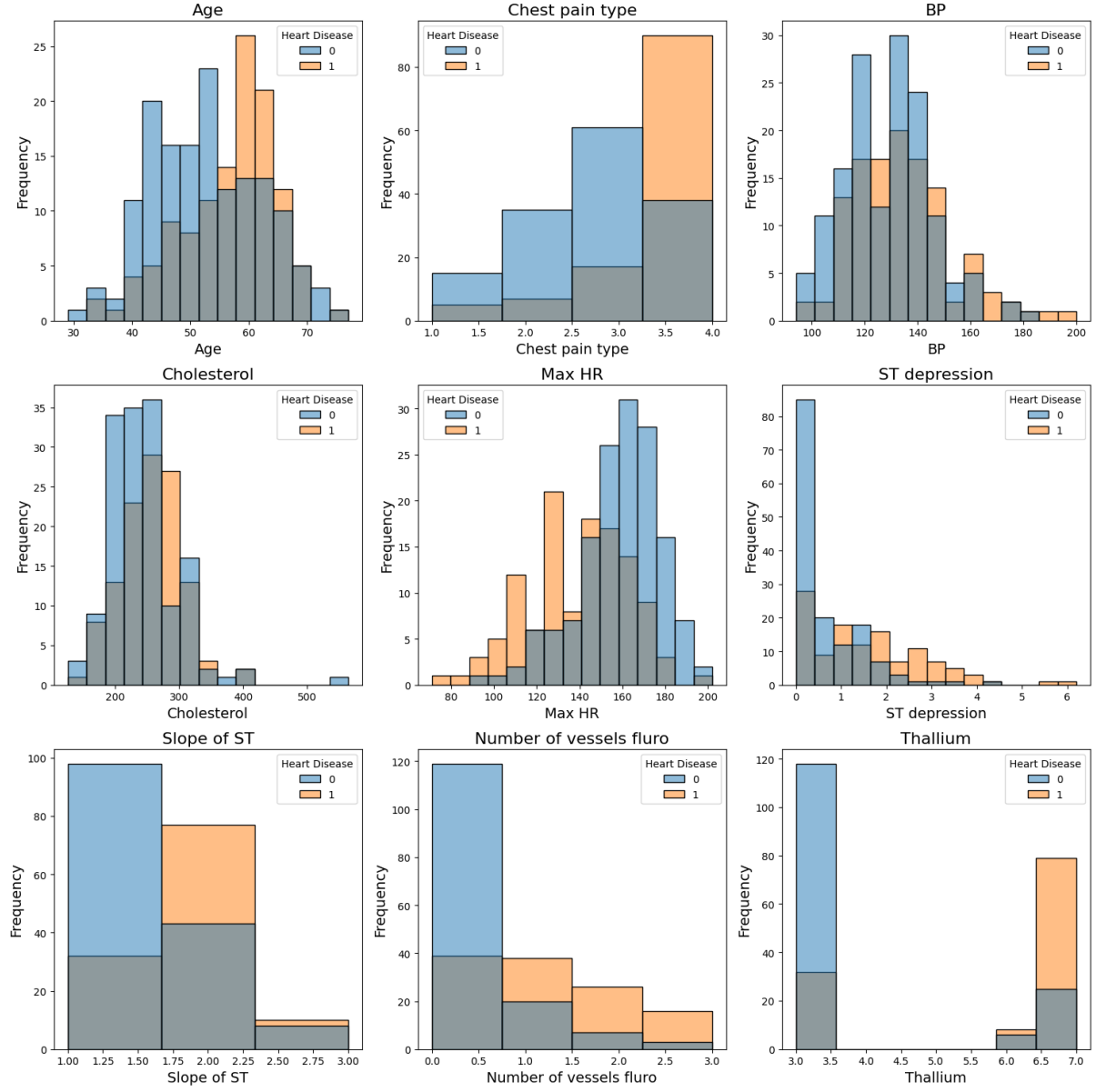
This is an appendix.

Figure 3: Overlayed histograms of numerical and categorical variables in the dataset.
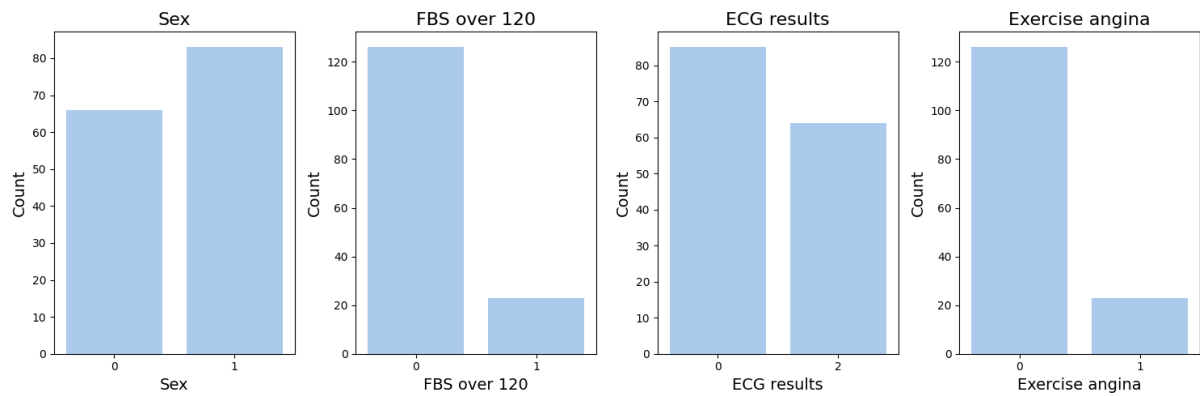
Figure 4: Bar charts of the binary variables on the dataset, filtered by class 0 (no heart disease).
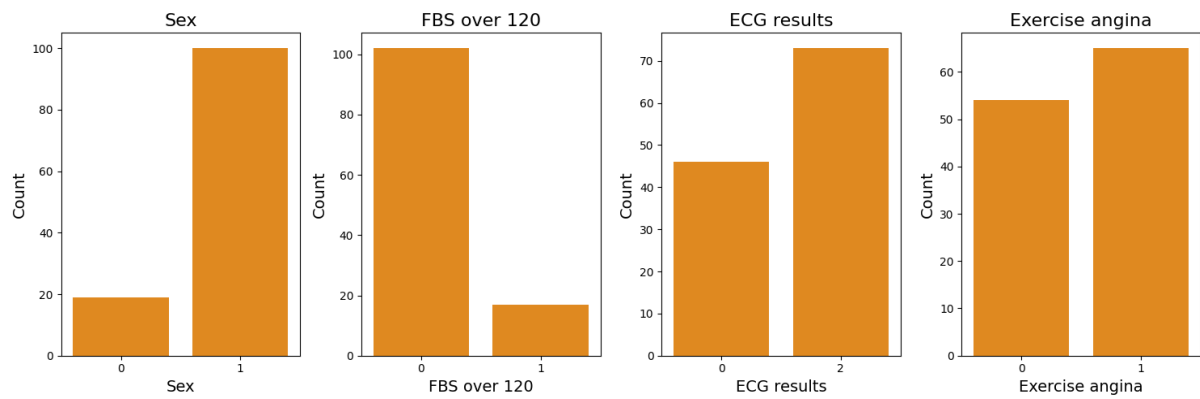

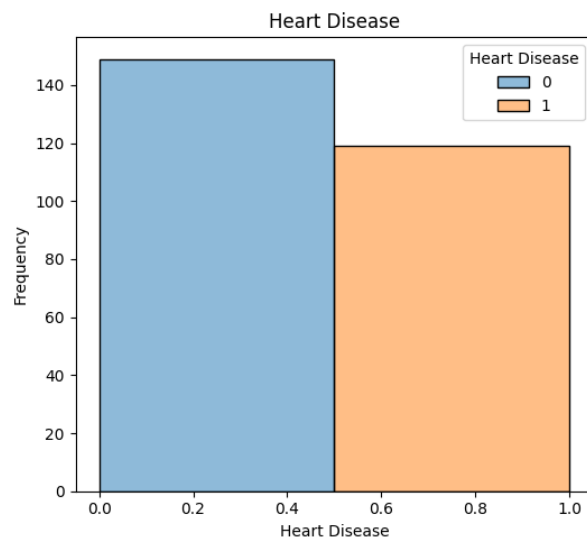**Figure 5:** Bar charts of the binary variables on the dataset, filtered by class 1 (heart disease).

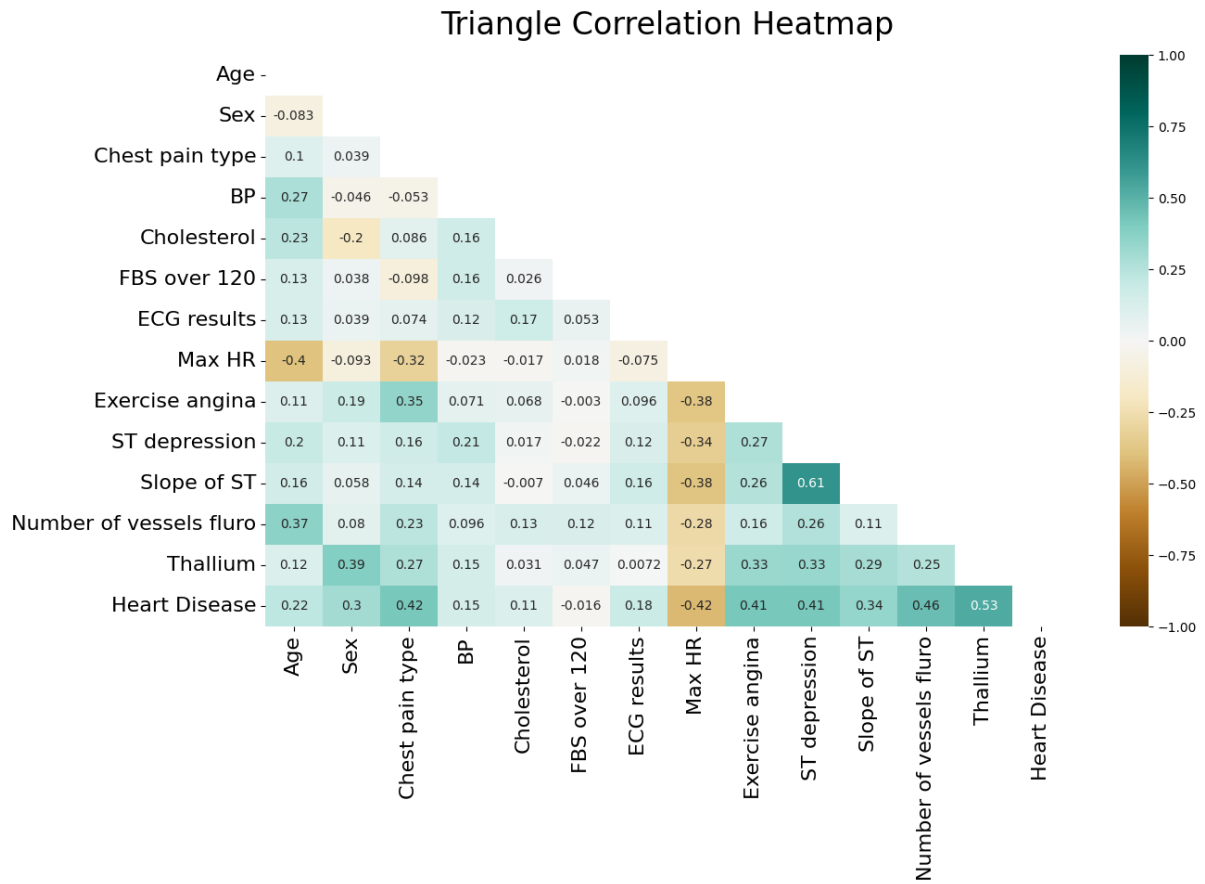**Figure 6:** Distribution of the target variable (Heart Disease).
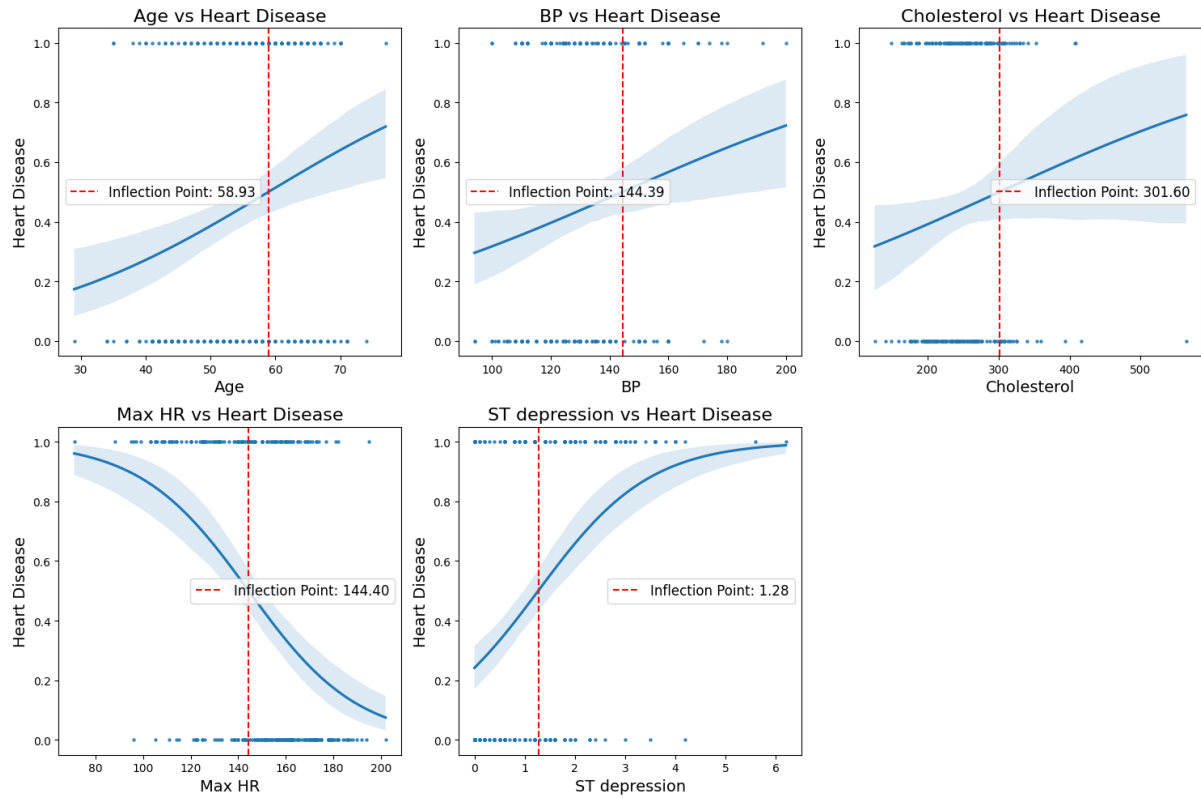
Figure 7: Correlation matrix.

Figure 8: Logistic Regression of the numerical variables with inflexion point.

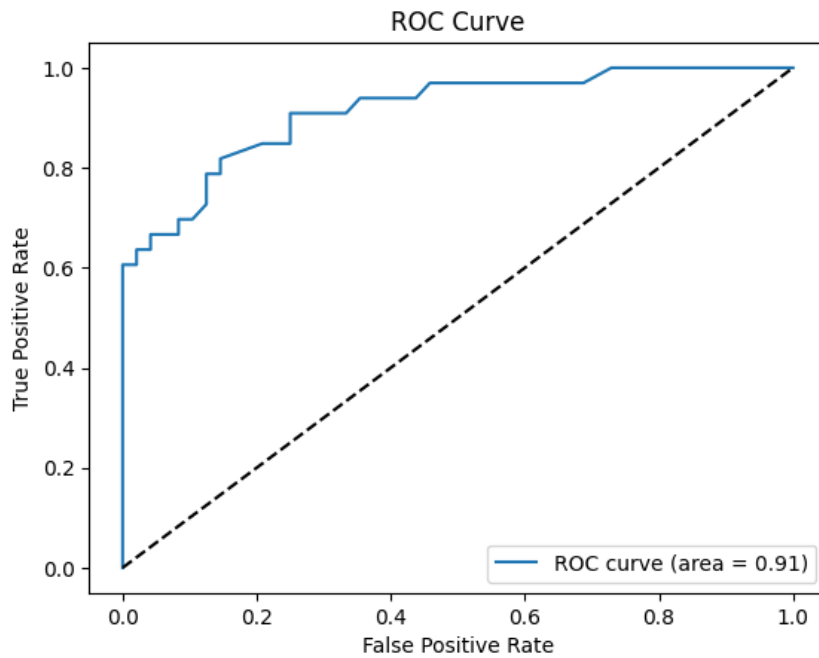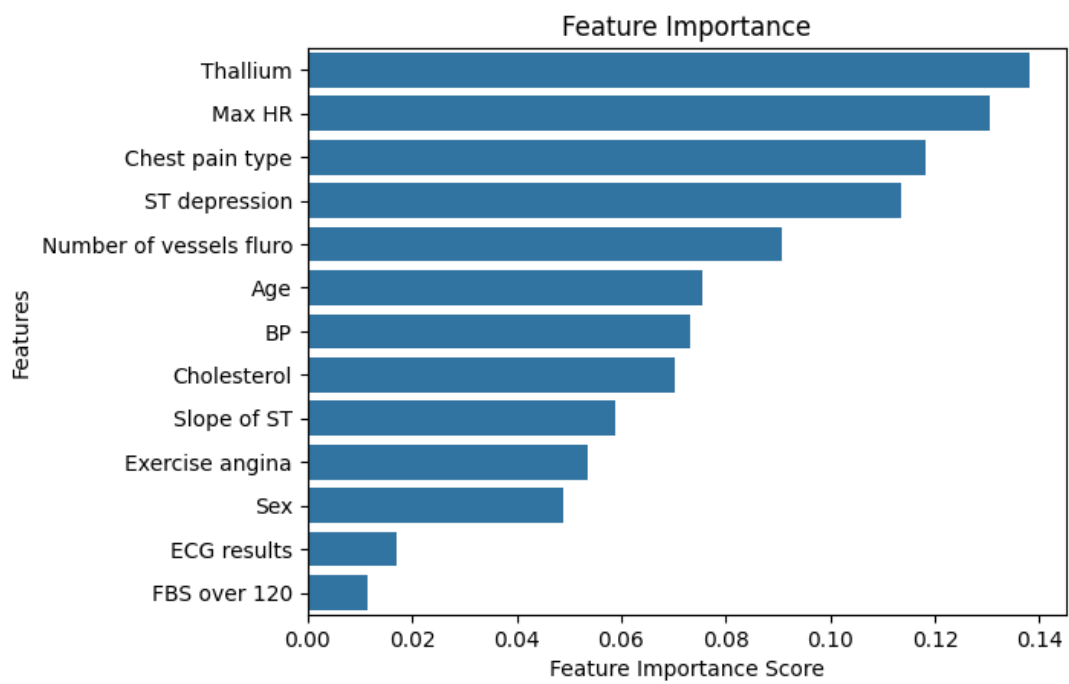**Figure 9:** ROC curve of random forest model.

Figure 10: Feature importance of random forest model.