TÉCNICO LISBOA

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

Henrique Miguel Duarte Anjos

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisors: Prof. Rui Miguel Carrasqueiro Henriques
Prof. Rafael Sousa Costa

Examination Committee

Chairperson: Prof. Helena Isabel De Jesus Galhardas
Supervisor: Prof. Rui Miguel Carrasqueiro Henriques
Member of the Committee: Prof. Daniel Rebelo dos Santos

October 2025

Declaration
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.

Acknowledgments

I want to thank my family for all their love and support. Special thanks to my parents, Carla and Rui, for making my studies possible and always being there for me. To my brother Diogo, thanks for growing up with me. My grandparents, Ofélia and Luís, my aunts and uncles—Nuno, Sónia, Jorge, Lúcia, Mafalda, Canoa, Joana, and João—and my cousins Maria, Salvador, Benedita, Lara, Tiago, Dinis, Joshua, and Simão, thank you for always being there and putting up with me.

I also want to thank my girlfriend Beatriz and her family for their love and support.

To my friends, thank you for being with me through these years. Caldas, Albino, Vasco, Francisco, and Vale, I really appreciate your company. Lima, thanks for your resilience as a friend. Tocha, Pinto, Barbosa, Batalheiro, and Inês, thanks for your friendship. And Simão, Rodrigo, Afonso, André, and Gonçalo, thanks for growing with me.

I would like to thank my dissertation supervisors, Prof. Rui Henriques and Prof. Rafael Costa and ULSM professionals Dr. Ana and Dr. Cristina, for their guidance, support and sharing of knowledge that has made this Thesis possible.

Finally, I want to thank all my friends and colleagues who helped me grow as a person and were there for me in good and bad times. And a special thanks to my grandparents, Luís and Celeste, for their love and for giving me motivation and purpose.

Abstract

Atrial fibrillation (AF) is the most prevalent cardiac arrhythmia worldwide and is strongly associated with increased risks of stroke, heart failure, and mortality. Traditional methods to detect and prognostic AF and its associated risks often fail to capture the full complexity of AF patterns, limiting their predictive accuracy. In spite of the improvements achieved by machine learning (ML) techniques, state-of-the-art AF-focused predictors do not generally incorporate longitudinal data, reducing their capacity to model the dynamic and evolving nature of individual behaviors and cardiophysiological indicators over time. The absence of a longitudinal perspective restricts understanding of how AF risk develops and changes across different prediction horizons. This study addresses these limitations by developing superior ML models tailored to predict adverse events within a longitudinal cohort of individuals with AF, while also laying the groundwork for future models that predict AF onset. Our work focuses on six critical clinical endpoints: stroke, all-cause death, cardiovascular death, heart failure hospitalizations, inpatient visits, and acute coronary syndrome. The ML models yielded an AUC of 0.65 for 1-year stroke prediction, outperforming CHA_2DS_2-VASc (0.59) and GARFIELD-AF (0.63). For all-cause mortality prediction, the models achieved an AUC of 0.78 against the 0.72 reference of GARFIELD-AF. In addition to predictive advances, the study identifies determinants of AF-related risks and introduces a prototype decision-support tool for clinical use. The work was conducted in collaboration with ULS Matosinhos, which reviewed and validated the findings.

Keywords

Atrial Fibrillation, Machine Learning, Electronic Health Records, Longitudinal Clinical Data, Medical Tool Interface, Decision Support.

Resumo

A fibrilhação auricular (FA) é a arritmia cardíaca mais prevalente a nível mundial e está fortemente associada a um aumento do risco de AVC, insuficiência cardíaca e mortalidade. Os métodos tradicionais de deteção e prognóstico da FA e dos seus riscos raramente captam toda a complexidade dos seus padrões, limitando a precisão preditiva. Apesar dos avanços com técnicas de machine learning (ML), os modelos atuais focados em FA raramente incorporam dados longitudinais, reduzindo a capacidade de representar a natureza dinâmica dos comportamentos individuais e dos indicadores cardiológicos ao longo do tempo. A ausência desta perspetiva longitudinal restringe a compreensão de como o risco de FA se desenvolve e muda em diferentes horizontes temporais. Este estudo procura ultrapassar essas limitações através do desenvolvimento de modelos de ML avançados, concebidos para prever eventos adversos numa coorte longitudinal de doentes com FA, estabelecendo também bases para futuros modelos de previsão da ocorrência de FA. O trabalho centra-se em seis desfechos clínicos críticos: AVC, morte por todas as causas, morte cardiovascular, hospitalização por insuficiência cardíaca, internamentos e síndrome coronária aguda. Os modelos obtiveram uma AUC de 0,65 para a previsão de AVC, superando o CHA_2DS_2-VASc (0,59) e o GARFIELD-AF (0,63). Para a mortalidade por todas as causas, atingiram uma AUC de 0,78 face ao 0,72 do GARFIELD-AF. Além dos ganhos preditivos, o estudo identifica determinantes do risco associado à FA e apresenta um protótipo de ferramenta de apoio à decisão clínica, desenvolvido com a ULS Matosinhos, que validou os resultados.

Palavras Chave

Fibrilhação Auricular, Aprendizagem Automática, Dados Clínicos Longitudinais, População Portuguesa, Interface Médica, Apoio à Decisão.

Contents

List of Figures

List of Tables

Acronyms

AF	Atrial Fibrillation
ML	Machine Learning
HF	Heart Failure
AMI	Acute Myocardial Infarction
CAD	Coronary Artery Disease
CKD	Chronic Kidney Disease
ECG	Electrocardiogram
ECHO	Echocardiogram
EF	Ejection Fraction
AI	Artificial Intelligence
SVM	Support Vector Machine
kNN	k-Nearest Neighbours
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
PCA	Principal Component Analysis
MLP	Multi-Layer Perceptron
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
GNN	Graph Neural Network
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
TPR	True Positive Rate

PPV	Predicted Positive Value
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
BMI	Body Mass Index
MRI	Magnetic Resonance Imaging
NNS	Number Needed to Screen
SHAP	Shapley Additive Explanations

1

Introduction

Contents

Atrial Fibrillation (AF) is the most common cardiac arrhythmia worldwide, and its occurrence associated with a significant risk increase of stroke, heart failure, and mortality [3]. Studies estimate that around 9 million individuals in the European Union (EU) were affected by AF in 2010, and the number is projected to double by 2060 [4]. In Portugal, 4070 deaths were attributable to AF in 2010, corresponding to nearly 4% of all deaths [5]. 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 [6]. This diagnostic gap highlights the need for tools that can detect AF and prognosticate its complications effectively.

A plethora of traditional methods have been proposed for predicting AF and its complications, encompassing point-based systems or traditional statistical models [7]. While these methods are functional, they often fail to capture the complexity of AF patterns [8]. In recent years, Machine Learning (ML) models have demonstrated superior performance compared to these traditional approaches [9]. 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. underestimating the significance of integrating specific data modalities. 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.

To address these challenges, this thesis lays the groundwork for developing a clinical decision-support tool designed to detect and prognosticate AF, as well as its associated clinical endpoints, using machine learning models specifically tailored for these tasks. This work is part of a broader initiative in collaboration with the Unidade Local de Saúde de Matosinhos (ULSM), which is expected to proceed in three phases:

1. Prognostication of complications in the AF patient population (focus of this thesis).
2. Prediction of AF onset in a case-control population.
3. Incorporation of advanced modalities such as cardiac imaging and electrophysiological data.

Given the data available at the time, this thesis focuses on phase 1, while also laying the groundwork for the following phases. It leverages a longitudinal cohort of AF patients, using electronic health records from ULSM collected over the past decades to derive population-specific insights. The work involves developing and optimizing machine learning algorithms to accurately predict AF-associated clinical endpoints, including stroke/systemic embolism (SE), all-cause mortality, cardiovascular death, heart failure hospitalizations, inpatient visits, and acute coronary syndrome. In parallel, a prototype clinical decision-support tool is designed, consisting of a backend API that provides programmatic access to the predictive models and a user-friendly graphical interface that allows healthcare professionals to input patient data and visualize predictions intuitively. The thesis also evaluates classical risk calculators, providing a comparative context, and establishes a foundation for future development and clinical application.

1.1 Research Contributions

The main contributions of this thesis include the development of machine learning models for AF-related outcomes, the exploration of time-aware approaches from longitudinal data, systematic evaluation against classical risk scores, and the design of a prototype clinical decision-support tool. In summary, seven major contributions are highlighted:

I. **Optimized predictive models of six major clinical endpoints:** Developed machine learning approaches to accurately prognosticate and predict the occurrence of AF-associated clinical end-

points using routine clinical data from a longitudinal Portuguese AF cohort. The endpoints include stroke/systemic embolism (SE), all-cause mortality, cardiovascular death, heart failure hospitalizations, inpatient visits, and acute coronary syndrome. Among these, stroke/SE and acute coronary syndrome present greater challenges for short-term prediction due to their low incidence, whereas the remaining endpoints can be reliably predicted over time horizons of one month or longer.

II. **Extension of predictors to longitudinal data variants:** The machine learning models were extended to include time-aware features, such as slope-augmented features capturing the evolution of indicators over time, and temporal information, including the proximity of biometric measurements, laboratory tests, diagnoses, and prescriptions. Incorporating longitudinal data consistently improved predictions across all outcomes.

III. **Exploration of different task formulations:** The models were extended to multiple formulations, including single-output and multi-output approaches, and various strategies for handling class imbalance. In addition, different prediction time horizons were explored to assess model performance across distinct temporal intervals.

IV. **Knowledge acquisition through predictor explainability:** Predictors' explainability was analyzed using SHAP to identify the most important features and gain insights into the clinical endpoints. For example, the all-cause mortality model confirmed well-established baseline risk factors such as age, cancer, prior cardiovascular disease, body mass index, and HDL cholesterol levels, while also revealing novel time-sensitive predictors. These novel predictors encompassed the temporal proximity of creatinine and digoxin prescriptions, longitudinal changes in creatinine, glycemia, or cholesterol measurements, and patterns of healthcare utilization, such as visit timing.

V. **Monitored improvements over classical risk calculators:** The performance of the machine learning models was compared against established clinical risk scores to quantify predictive improvements. For 1-year stroke prediction, the ML models achieved an AUC of 0.65, outperforming GARFIELD-AF (0.63) and CHA_2DS_2-VASc (0.59). Similarly, for all-cause mortality prediction, the ML models reached an AUC of 0.78, compared with 0.72 for GARFIELD-AF. These results demonstrate that incorporating machine learning with longitudinal and temporal features provides a measurable improvement over traditional risk calculators, highlighting the potential of ML approaches to enhance prognostic accuracy in AF patients.

VI. **Medical validation:** The results were reviewed and validated by ULSM medical experts, who provided insights into the model predictions, their potential clinical significance, and directions for future research.

VII. **Design and development of a prototype clinical decision-support tool:** A backend API was

developed to provide access to the ML models, enabling programmatic interaction. In parallel, a user-friendly graphical user interface (GUI) was implemented, allowing clinicians to input patient data and visualize model predictions in an intuitive and informative way.

1.2 Document Structure

This thesis is organized as follows:

- **Chapter 1: Introduction** – provides background, motivation, research objectives, contributions, and an overview of the thesis structure.
- **Chapter 2: Background** – introduces key concepts related to heart failure, machine learning, and the prediction of AF-related events.
- **Chapter 3: Related Work** – reviews the history of traditional AF risk calculators and the current state of research on machine learning approaches for predicting AF with and without electrocardiograms, as well as methods for predicting AF-related events.
- **Chapter 4: Dataset** – describes the study population, data sources, and presents an exploratory data analysis of the dataset.
- **Chapter 5: Solution** – details the developed solution, including data preprocessing, feature engineering, traditional risk calculators, and machine learning methodology.
- **Chapter 6: Results and Evaluation** – presents descriptive statistics, model predictive performance, feature importance analyses, and comparisons with classical risk scores. Additionally, interprets key findings and discusses clinical implications.
- **Chapter 7: Medical Tool** – describes the development of the API and GUI, and presents the interface.
- **Chapter 8: Conclusion** – summarizes the main contributions, highlights the significance of the findings, and proposes next steps for research and system limitations.

2

Background

Contents

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 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 [10]. 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. Heart Failure is a significant global health concern, with rising prevalence and a substantial impact on mortality, morbidity, and healthcare costs [11–13].

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 echocardiograms.

Heart Structure. The heart is divided into four chambers: the right atrium and right ventricle, and the left atrium and left ventricle (see Figure 2.1). 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 [14]. 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 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 [14, 15].

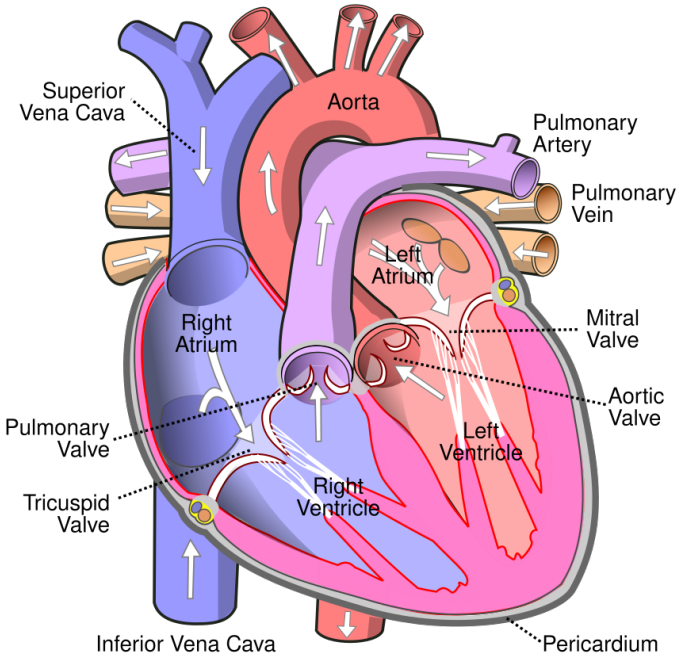

Figure 2.1: Diagram of the Heart and Blood Flow. Source: Wikipedia

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 [16].

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 [17]. 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 [18]. Angina, characterized by chest pain due to reduced blood flow to the heart, is often a precursor or companion to CAD [19]. Dyslipidemia, an abnormal amount of lipids in the blood, contributes to atherosclerosis and increases the likelihood of CAD and AMI [20]. 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 [21, 22]. 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 [23]. 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 [24].

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 [25].

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 [26–28]. These indicators help identify contributing risk factors such as hypertension and obesity, which are closely associated with the development and progression of heart failure [29, 30]. 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 [31]. 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 [32]. 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 [33]. 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 [34]. 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 [35]. 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 [36].

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 is 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 [37, 38]. 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 [39], 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 [37]. 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 characteristic of systolic heart failure [40].

In addition to these structural measurements, an echocardiogram can reveal the presence of aortic valve disease [41]. Aortic valve disease, such as aortic stenosis, increases the workload on the left ventricle, leading to hypertrophy and, over time, contributing to heart failure [42]. 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 [43]. 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 is a branch of Artificial Intelligence focused on developing algorithms that enable computers to learn from data, supporting knowledge acquisition and decision making [44]. 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 latter, Reinforcement Learning, will not be addressed as it is not relevant to the scope of this study.

Supervised Learning. Predictive models are typically trained using labeled data, where the input data is paired with the corresponding correct output. The algorithm learns a mapping between the input and output variables, which can be subsequently used to predict outputs for unseen data [45]. Through this process, the model identifies patterns in the data to map inputs to desired outputs accurately. Supervised learning in its simplest form can be divided into two major types of tasks: classification and regression.

Classification tasks aim at estimating categorical outcomes. Depending on the cardinality and number of outcome variables, classification tasks can be further divided into three types: binary, multi-class, and multi-label [46]. 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.

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 [47]. 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 [48]. 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.

Logistic regression is a simple and widely used linear model [49]. Nonlinear variants, such as kernel logistic regression, extend its applicability to more complex datasets [50]. 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 [51]. Similarly, k-Nearest Neighbours (kNN) is a straightforward and versatile algorithm that classifies data by identifying the k closest data points based on a specified distance metric [52].

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 [53]. Nonlinear approaches, such as polynomial [54] and kernel regression [55], 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 [51, 56].

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 [57]. 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 [58, 59].

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 [60]. Hierarchical clustering generates a tree of clusters by either merging or splitting clusters at each step, enabling a flexible approach to understanding data structure [61]. 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 [62]. 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 [63].

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 [64], 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 [65]. 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 [66]. 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 [67]. 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 [68]. 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 Networks (RNNs), 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. [69]

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, they capture the dependen-

cies between connected components, enabling predictions based on the underlying structural information. [70].

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 [71], they have revolutionized the field of artificial intelligence, achieving unprecedented accuracy across numerous domains [72].

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 (equation 2.1) serves as a fundamental metric, measuring the proportion of correct predictions out of all predictions made.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.1}$$

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 or Predicted Positive Values (PPVs) (equation 2.2), sensitivity (equation 2.3), also known as recall or True Positive Rate (TPR), Fβ-score (equation 2.4), and specificity (equation 2.5), are often employed.

$$\text{Precision (PPV)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.2}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.3}$$

$$\text{F}_\beta\text{-score} = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}} \tag{2.4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.5}$$

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 F_β-score balances precision and recall, and it is useful when the trade-off between false positives and false negatives is significant [73].

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 extended using macro, micro, or weighted averaging. Macro averaging (equation 2.6) calculates the metric independently for each class and averages the results, treating all classes equally.

$$\text{Macro Average} = \frac{1}{n}\sum_{i=1}^{n} \text{Metric}_i \tag{2.6}$$

Micro averaging (equation 2.7) aggregates the contributions of all classes to compute the metric, making it suitable for datasets with class imbalances.

$$\text{Micro Average} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n} (\text{TP}_i + \text{FP}_i + \text{FN}_i)} \tag{2.7}$$

Weighted averaging (equation 2.8) adjusts for the proportion of instances in each class, giving more importance to metrics for larger classes [74].

$$\text{Weighted Average} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} \tag{2.8}$$

- w_i represents the weights applied to x-values and X_i represents the data values to be averaged.

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 [75].

$$\text{Hamming Loss} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{L}\sum_{j=1}^{L} \mathbf{1}(y_{ij} \neq \hat{y}_{ij}) \tag{2.9}$$

- 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, and L represents the number of labels for multi-label classification.

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 [76].

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 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 [77].

2.3 Predictive Modeling of AF-related Events

Machine Learning has proven highly valuable for predicting heart failure, aiding in early diagnosis, risk stratification, and personalized treatment [78–80]. This section delves into the specific types of data used in the prediction of cardiac complications—particularly heart failure—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 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 can be applied directly to ECG readings [81]. Imaging data from echocardiograms and cardiac Magnetic Resonance Imaging (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 [82].

Predictive approaches. A variety of machine learning models have been applied to heart failure prediction [81], 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 [83]. 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 [84, 85]. SVMs 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 [86]. Neural networks, especially CNNs and RNNs, are powerful tools for handling imaging and time-series data [87, 88]. 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 [89]. 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 [90]. Dimensionality reduction methods, such as 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 Shapley Additive Explanations (SHAP) help by providing insights into the importance of individual features [91], helping clinicians make data-driven decisions in the diagnosis and management of heart failure.

Specific 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, 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

Contents

Predictive risk scores for new-onset Atrial Fibrillation have been developed since 2009 [92]. Section 3.1 provides a brief overview of classic AF risk calculators, while Sections 3.2 and 3.3 explore machine learning approaches developed in the absence and presence of ECG data, respectively. Finally, Section 3.4 addresses the prediction of complications among patients with pre-existing AF.

3.1 Classical Risk Calculators of AF

Several classical risk score calculators have been developed to assess the risk of onset AF, its complications, and guide treatment decisions. These scores are usually either point-based systems, where each risk factor contributes a fixed number of points, or derived from Cox regression models, which estimate the relative hazard of developing AF over time. These models are recognized for their simplic-

ity and accessibility, as they rely on readily available clinical parameters and do not require advanced computational tools or specialized biomarkers, enabling straightforward stratified risk assessment.

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 [93]. 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 [94]. 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 Birmingham 2009 schema, also known as CHA_2DS_2-VASc score [2] emerged as an improvement over $CHADS_2$ and has been widely used since 2010 [95]. 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 [96]. 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 [97].

The Framingham Heart Study (FHS) score was also developed in 2009, and was the first model developed to actually predict risk of developing AF [92]. Using data from the FHS, a long-term population-based study in the United States 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, developed in 2011, to estimate the risk of developing AF in a community-based population [98]. Unlike the FHS model, which was derived primarily from a white population, ARIC was created using a prospective, community-based cohort that included both black and white participants in the United States. 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 [1], 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 in predicting factors of progression from paroxysmal AF to persistent AF [99], was also used to predict new-onset AF in a Taywan cohort [100]. Ultimately HATCH score was also capable of estimating 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 for AF [101]. 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.

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 [102]. 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 structural heart disease. 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 [103].

Several other models were developed during the 2010s such as the Women's Health Study (WSH) score [104], a model developed on a Japanese cohort [105], the Shandong multi-center health check-up study [106], the EHR-AF [107], and more recently the HARMS$_2$-AF score [108].

In 2020, a meta-analysis compared some of these classical models for incident AF risk [7], and only three models (CHARGE-AF, FHS, CHA_2DS_2-VASc) yielded significant overall discrimination capacity for AF incidence, and only CHARGE-AF showed superior discrimination with a uniform prediction window [109]. 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 [110].

Overall, these scores have shown appropriate model discrimination for the prediction of incident AF (AUC, generally ranging between 0.65-0.75) [109] 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 [111, 112].

3.2 Non-ECG Clinical ML Approaches

Salvi et al. [113] highlights the need for approaches that are able to tap into the value of clinical data and leverage longitudinal stances to promote generalization, interpretability, and applicability in real-world healthcare settings.

In 2019, 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 during a designated 6-month period [114]. A single-layer neural network, using 200 electronic health record features, achieved the best 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.

Also in 2019, an analysis of 2,994,837 individuals (3.2% with AF) from the Clinical Practice Research Datalink (CPRD), utilized longitudinal data and identified time-varying neural networks as the most effective predictive model [83]. 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 [115]. 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 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 [116], 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 in primary care in England. The algorithm proved to be a valuable tool to select primary care groups at high risk of undiagnosed AF who may benefit from diagnostic testing.

In 2020, another model, EHR-AF, was developed using data from 21 million individuals to predict 5-year incident atrial fibrillation. The model achieved an AUC of 0.808, comparable to CHARGE-AF (0.806) and superior to CHA_2DS_2-VASc (0.72) and C_2HEST (0.68) [117].

Recently, in 2023, another model called FIND-AF was developed to predict the risk of incident AF within 6 months [118]. 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 AUC of 0.824, outperforming CHA_2DS_2-VASc (AUC = 0.784) and C_2HEST (AUC = 0.757).

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) [81]. The findings show that ECG signals of 300 seconds (5 minutes) were commonly used, and 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 were obtained using a mixture of experts model, followed by SVM implementations [119], who got an overall accuracy of 0.982, 1 sensitivity, and 0.965 specificity with a data split of 47/53.

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 [120]. 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 [121]. 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)- using input from 10-second 12-lead ECGs [122]. 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.

3.4 Prediction of AF Related Outcomes in AF Patients

As mentioned earlier, $CHADS_2$ and CHA_2DS_2-VASc are the primary risk stratification schemes developed in 2009 to predict stroke risk in patients with atrial fibrillation. Among them, CHA_2DS_2-VASc offers a more comprehensive assessment of stroke risk and has been widely adopted in clinical practice since 2010.

In 2010, a user-friendly HAS-BLED score was developed to assess 1-year risk of major bleeding in patients with AF, using data from the Euro Heart Survey on AF. The risk score gives points based on the following comorbidities: Hypertension, Abnormal renal/liver function, Stroke, Bleeding history or predisposition, Labile international normalized ratio, Elderly ($>$ 65 years), Drugs/alcohol concomitantly. The bleeding score achieved an AUC of 0.72 in the derivation cohort of AF patients with extensive use of oral anticoagulation and was consistent when applied in several subgroups. The score gave similar AUCs when patients were receiving antiplatelet agents alone, and was concluded to be a practical tool to assess the individual bleeding risk of real-world patients with AF.

In 2012, HAS-BLED bleeding risk score was experimented with in predicting cardiovascular events and mortality in anticoagulated patients with AF, and was shown to be useful in the prediction of these, even though the score was designed for bleeding risk. However, a multivariate analysis was slightly better at predicting cardiovascular events and mortality, showing that HAS-BLED is not the most accurate score to predict death or cardiovascular events but that it can be related to both bleeding and thrombotic events [123].

In 2016, ABC (age, biomarkers, clinical history) stroke risk score was developed to predict 6-year stroke risk in AF patients with a Cox regression model. The ABC-Stroke score achieved higher AUCs than CHA_2DS_2-VASc in both the derivation cohort (0.68 vs. 0.62) and the external validation cohort (0.66 vs. 0.58). The score used biomarkers such as Troponin I and NT-proBNP [124]. Later in 2016, the ABC score was also developed to predict bleeding risk score in patients with AF, and also achieved similar results [125]. Moreover, in 2017, ABC was also developed to predict death in anticoagulated patients with AF, including both clinical information and biomarkers, in the ARISTOTLE cohort. The model was well-calibrated and yielded higher AUC than a model based on all clinical variables, both in

the derivation (0.74 vs. 0.68) and validation cohorts (0.74 vs. 0.67). The derivation cohort was RE-LY trial [126]. Later in 2021, ABC-AF was also validated in patients not receiving oral anticoagulation in the ACTIVE A, and AVERROES trials, which yielded similar positive results [127].

In 2020, GARFIELD-AF risk model was developed to predict stroke, major bleeding or mortality in AF patients. The model is described as developed using sophisticated statistical modeling techniques, and can be found on https://af.garfieldregistry.org/garfield-af-risk-calculator. It achieved a higher score in stroke and mortality prediction compared with CHA_2DS_2-VASc and a higher score than has bled score in predicting major bleed events across all risk groups [128]. The model was then externally validated in ORBIT-AF cohort, and the discriminatory value was still superior to CHA_2DS_2-VASc score.

In 2022, the Fushumi AF registry — a cohort from Fushumi, Japan — was used to learn predictors of ischemic events in patients with AF. Apart from biological and past medical and treatment history, the model also used medication, blood test, and echocardiogram data. The model was able to have an increased performance compared to CHA_2DS_2-VASc with an AUC of 0.72 compared to CHA_2DS_2-VASc AUC of 0.62 [129].

Also in 2022, using the same Fushimi AF registry, another machine learning risk model was developed to predict incident heart failure in patients with AF. The model outperformed Framingham risk score with an AUC of 0.75 vs 0.67 [130]. However, it is to note that Framingham risk score is not specifically tailored to predict heart failure events in an AF specific population.

In 2023, machine learning models were trained to predict 1-year mortality after AF diagnosis in a french cohort. The best model achieved an AUC of 0.785, and the score was superior to CHA_2DS_2-VASc and HAS-BLED risk scores [131].

Also in 2023, machine learning models were trained to predict all-cause death, cardiovascular death, major bleeding and stroke in anticoagulated patients with AF. The best model achieved an AUC of 0.78 predicting all-cause death and cardiovascular death surpassing CHA_2DS_2-VASc and HAS-BLED risk scores. Bleeding prediction achieved an AUC of 0.71 also surpassing HAS-BLED, and the ischemic stroke prediction was described as suboptimal with an AUC of 0.606 due to the low number of events [132].

In 2024, another model was developed to predict 1-year stroke risk in a South Asian, Indian population, and achieved an AUC of 0.82 compared with CHA_2DS_2-VASc 0.67. However, when externally validated in another Asian cohort, the model only achieved an AUC of 0.67 while CHA_2DS_2-VASc achieved 0.62 [133].

In 2024, a meta-analysis evaluated 13 studies on stroke prediction in patients with atrial fibrillation using machine learning methods. The mean AUC across studies was 0.73, with models such as XGBoost and logistic regression achieving higher performance, while neural networks showed compar-

atively lower AUC values [134].

Within the prediction of AF-related outcomes, stroke is the most extensively studied, given the substantial increase in stroke risk associated with atrial fibrillation. Other predictive models have focused on outcomes such as bleeding or all-cause mortality, while less attention has been given to events like acute coronary syndrome, cardiovascular death, and hospitalizations.

4

Dataset

Contents

This chapter provides a detailed description of the dataset acquired for this study. 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, Portugal, reflecting the activity of over 1,000 doctors from various specialties. The dataset was specifically designed within the scope of preventive assessment initiatives led by ULSM, with a focus on prevention of AF at primary care level. Importantly, the cohort study was purposefully built by ULSM to address the specific objectives of this work.

The chapter also provides an overview of the monitored variables and an exploratory data analysis, highlighting descriptive statistics and cohort-level patterns. This analysis forms the basis for subsequent modeling and predictive tasks, ensuring a clear understanding of cohort characteristics and data quality.

4.1 Data Description

The available dataset spans approximately 25 years and includes patients over 40 years old who were diagnosed with AF between 1 January 2012 and 31 December 2021. The dataset comprises a population of 7,203, and 167 features encompassing demographics, clinical records (pertaining to laboratory, pharmaceutical, and surgical acts), and the temporal context of the previous electronic registry.

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 original data was de-identified according to the HIPAA Safe Harbour Method with noise added to all variables.

In Table 4.1, 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: chronic obstructive pulmonary disease, myocardial infarction or unstable angina, type 1 diabetes mellitus, type 2 diabetes mellitus, and valvular heart disease. Additionally, there are several outcomes: acute coronary syndrome, arterial embolism, stroke, inpatient visit, heart failure hospitalization, cardiovascular death, and all-cause death. The dataset also includes binary variables indicating whether a patient is currently on any of the following medications: antiarrhythmics, anticoagulants, angiotensin-aonverting anzyme anhibitor (ACEi), angiotensin receptor brlockers (ARBs), angiotensin receptor-neprilysin inhibitor (ARNi), antiplatelets, beta blockers, calcium channel blockers, digoxin, dipeptidyl peptidase-4 inhibitor (DPP4i), GLP-1 agonists, insulin, ivabradine, loop diuretics, other diuretics, metformin, mineralocorticoid receptor antagonist (MRA), nitrates, sodium-glucose cotransporter-2 inhibitors (SGLT2i), statins, and sulfonylurea. Moreover, the data contains additional binary variables that represent whether a patient had the following interventions before: cardiac surgery, cardiac device, coronary surgery, and percutaneous coronary intervention. To represent smoking, the dataset has a set of binary features: current smoker, former smoker, never smoked, and no information smoker. Some laboratory tests also include an additional feature representing the patient's initial measurement for that specific exam. These tests are: HbA1c, creatinine, estimated glomerular filtration rate, glycemia, HDL cholesterol, LDL cholesterol, total cholesterol, triglycerides, thyroid simulating hormone, urine albumin-to-creatinine ratio, and international normalized ratio (INR).

Every feature—except for index date, age (HIPAA), and sex (female)—has an associated integer day count indicating when the event occurred relative to the AF diagnosis. For comorbidities, the index date defines the diagnosis data. For laboratory tests, it indicates the date of the measurement. For biometric features, it corresponds to the date the information was extracted, while those for medications refer to the last prescription.

Table 4.1: Some selected variables of the USLM dataset. Binary features representing comorbidities and dispensed medications are not listed for simplicity's sake.

Feature Name	Description	Data Type	Units	Possible Values
index_date	index date (random)	integer	N/A	[-100,000; 100,000]
age_hipaa	age (HIPAA compliant)	categorical	years	[40-49, 50-59, 60-69, 70-79, 80-89, 90+]
female	female	binary	N/A	[0, 1]
wgt	weight	float	kg	> 0
hgt	height	integer	cm	> 0
bmi	body mass index	float	kg/m^2	> 0
sbp	systolic blood pressure	float	mmHg	> 0
dbp	diastolic blood pressure	float	mmHg	> 0
tc	total cholesterol	float	mg/dL	> 0
ldl	LDL cholesterol	float	mg/dL	> 0
hdl	HDL cholesterol	float	mg/dL	> 0
glc	glycemia	float	mg/dL	> 0
tg	triglycerides	float	mg/dL	> 0
crt	creatinine	float	mg/dL	> 0
egfr	estimated glomerular filtration rate	float	mL/min/1.73m²	> 0
tsh	thyroid stimulating hormone	float	mIU/L	> 0
uacr	urine albumin-to-creatinine ratio	float	mg/g	> 0
a1c	Hba1c (glycated hemoglobin)	float	mmol/mol	> 0
cancer	cancer	binary	N/A	[0, 1]
carotd	carotid disease	binary	N/A	[0, 1]
cd	coronary disease	binary	N/A	[0, 1]
dlp	dyslipidemia	binary	N/A	[0, 1]
esrd	end stage renal disease	binary	N/A	[0, 1]
flt	flutter	binary	N/A	[0, 1]
hf	heart failure	binary	N/A	[0, 1]
hta	hypertension	binary	N/A	[0, 1]
slap	sleep apnea	binary	N/A	[0, 1]
stk	stroke	binary	N/A	[0, 1]
tyrd	thyroid disease	binary	N/A	[0, 1]

4.2 Exploratory Data Analysis (EDA) of the AF Cohort

To provide a clearer understanding of the dataset and its key variables, descriptive statistics for selected numerical variables—including mean, standard deviation, and proportion of missing values—are presented in Table 4.2. As shown, LDL cholesterol, HbA1c, TSH, UACR, and eGFR have a relatively high proportion of missing values ($>$20%), which may impact subsequent analyses. In contrast, variables age, weight, systolic, and diastolic blood pressure have negligible missingness ($<$1%). Overall,

the mean and standard deviation values indicate moderate variability for most anthropometric and biochemical measures. Triglycerides, glycemia, and UACR display relatively large standard deviations, suggesting substantial variability across participants.

Table 4.2: Descriptive statistics of selected numerical variables: mean and missing rate (after outlier removal)

Variable	Value (mean $\pm$ SD)	Missing Values (%)
Age*	77.1 ± 11.2	0.0%
Weight	74.4 ± 15.4	15.0%
Height	162.1 ± 9.6	16.5%
Body mass index (BMI)	28.3 ± 5.5	17.0%
Systolic blood pressure	136.6 ± 18.9	12.2%
Diastolic blood pressure	77.4 ± 12.0	12.2%
Total cholesterol	176.7 ± 41.9	14.3%
LDL cholesterol	108.7 ± 35.6	31.2%
HDL cholesterol	45.5 ± 13.6	15.0%
Triglycerides	116.9 ± 59.1	14.6%
Creatinine	1.1 ± 0.6	4.8%
Glycemia	127.2 ± 55.9	5.0%
Hba1c	6.3 ± 1.2	45.6%
TSH	1.9 ± 1.4	33.4%
UACR	33.0 ± 57.1	55.7%
eGFR	63.7 ± 21.1	35.3%

* Age is estimated from categorical age groups using midpoints.

Table 4.3 presents the distribution of the most relevant variables, including demographic and clinical characteristics. For each variable, the table shows the number of subjects in each category, the corresponding percentage of the total population, and the number of cardiovascular death cases observed within that category. The majority of the population is between 70 and 89 years old (64.4%), with a slightly higher proportion of women than men (53.3% women). BMI and weight are higher than reference values from healthy populations. Hypertension is highly prevalent, affecting 81% of the population; however, systolic and diastolic blood pressure values appear to be within normal ranges, likely due to the widespread use of antihypertensive medication (91.5%). Total cholesterol and triglycerides are elevated in only a subset of subjects (24% and 17.9%, respectively). Nearly 40% of the population is diabetic (37.4%), and most have never smoked (75.8%). Several comorbidities are common: dyslipidemia (64.4%), heart failure (30.3%), valvular heart disease (26.2%), and cancer (22.5%). Other conditions with notable prevalence include myocardial infarction or unstable angina (17.1%), chronic obstructive pulmonary disease (12.4%), thyroid disease (11.3%), coronary heart disease (10.8%), peripheral artery disease (9%), and carotid disease (6.3%). The distribution of cardiovascular deaths across population groups is consistent with the underlying group sizes.

Table 4.3: Descriptive statistics of the study cohort

Risk Factor	N	Percentage	CV Death
Age (years)			
40-49	130	1.8%	3
50-59	435	6.0%	24
60-69	1,178	16.4%	107
70-79	2,184	30.3%	322
80-89	2,453	34.1%	620
$\geq$ 90	774	10.7%	204
Gender			
Male	3,366	46.7%	582
Female	3,837	53.3%	698
Body mass index (kg/m^2)			
$<$20	261	3.6%	55
20–$<$25	1513	21.0%	298
25–$<$30	2296	31.9%	386
$\geq$ 30	2087	29.0%	332
Height (cm)			
$<$156	1653	22.9%	322
156–$<$164	1816	25.2%	332
164–$<$173	1773	24.6%	291
$\geq$ 170	947	13.1%	138
Weight (kg)			
$<$55	567	7.9%	122
55–$<$70	1919	26.6%	368
70–$<$85	2361	32.8%	394
85–$<$100	1079	15.0%	169
$\geq$ 100	361	5.0%	41
Systolic blood pressure (mmHg)			
$<$100	152	2.1%	25
100–$<$120	935	13.0%	186
120–$<$140	2682	37.2%	416
140–$<$160	2023	28.1%	355
$\geq$ 160	662	9.2%	149
Cancer	1622	22.5%	324
Carotid disease	454	6.3%	110
Coronary heart disease	777	10.8%	192
Dyslipidemia	4635	64.4%	821

Table 4.3 (*continued*)

Risk Factor	N	Percentage	CV Death
Diastolic blood pressure (mmHg)			
$<$70	1612	22.4%	359
70–$<$80	2047	28.4%	342
80–$<$90	1909	26.5%	300
90–$<$100	693	9.6%	100
$\geq$ 100	194	2.7%	31
Total cholesterol (mg/dL)			
$<$200	4580	63.6%	852
200–$<$240	1292	17.9%	208
$\geq$ 240	441	6.1%	70
Triglycerides (mg/dL)			
$<$150	5008	69.5%	913
150–$<$200	770	10.7%	125
$\geq$ 200	520	7.2%	89
Diabetes mellitus (mg/dL)			
Type 1	323	4.5%	72
Type 2	2368	32.9%	466
Hypertension medication use	6593	91.5%	1202
Smoking status			
Never	5456	75.8%	969
Former	426	5.9%	62
Current	428	5.9%	59
Chronic obstructive pulmonary disease	894	12.4%	204
Flutter	602	8.4%	125
Heart failure	2184	30.3%	529
Myocardial infarction or unstable angina	1234	17.1%	276
Peripheral artery disease	651	9.0%	159
Stroke	197	2.73%	29
Thyroid disease	815	11.3%	151
Valvular heart disease	1885	26.2%	409
Hypertension	5859	81.3%	1085
Cardiac device	157	2.2%	28
Anticoagulants	3192	44.3%	473

Figure 4.1 presents the distribution of several key indicators in the dataset, stratified by cardiovascular death and normalized to account for class imbalance. The results suggest that patients aged 80–89 years or older have a higher likelihood of cardiovascular death. Similarly, weight below 70 kg or a BMI under 35 appear associated with increased risk.

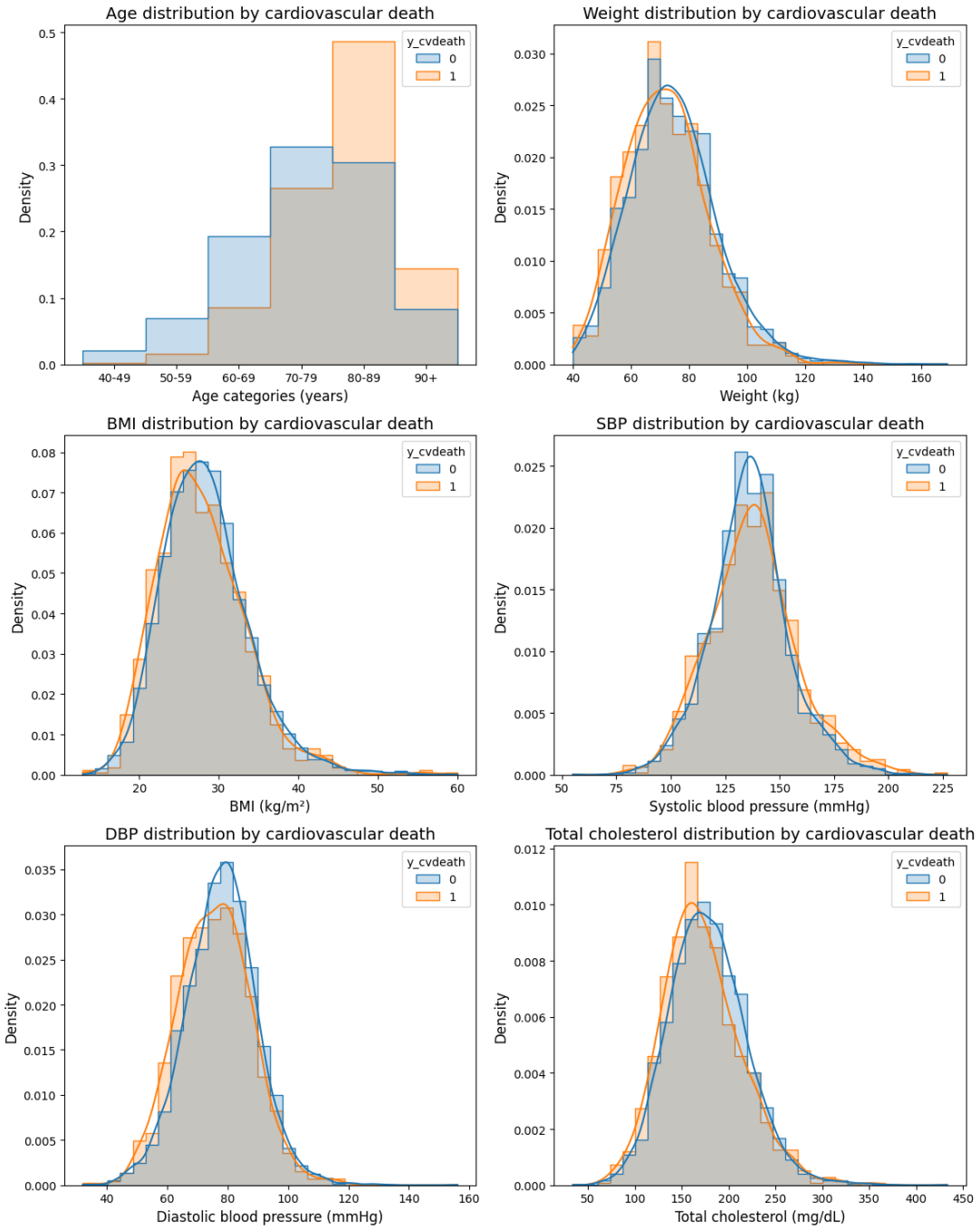

Figure 4.1: Histograms of selected variables stratified by cardiovascular death

For systolic blood pressure, values above 160 mmHg or below 120 mmHg are linked to a higher risk, while for diastolic blood pressure, levels below 70 mmHg show a similar trend. In addition, total cholesterol levels under approximately 170 mg/dL are also associated with increased likelihood of cardiovascular death.

Figures A.1, A.2, and A.3 in the appendix present additional variables stratified and normalized by cardiovascular death, showing comparable trends. Specifically, height under 160 cm, LDL cholesterol below 100 mg/dL, HDL cholesterol under 40 mg/dL, glucose levels above 125 mg/dL (the diagnostic threshold for diabetes), HbA1c greater than 6.5% (also diagnostic for diabetes), creatinine levels above 1.3 mg/dL (indicating impaired kidney function), eGFR below 60 (the threshold for chronic kidney disease), TSH above 3 (with values above 4 indicating hypothyroidism), and UACR above 25 (with values above 30 indicating kidney damage) are all associated with increased cardiovascular mortality. In addition, both the presence of heart failure and valvular heart disease are strong predictors of cardiovascular death. Other noteworthy comorbidities and treatments associated with increased risk include cancer, coronary artery disease, chronic obstructive pulmonary disease, myocardial infarction/unstable angina, peripheral artery disease, type 2 diabetes, hypertension, and the use of anticoagulants, ACE inhibitors, antiplatelets, beta-blockers, calcium channel blockers, loop diuretics, other diuretics, and nitrates.

Figure 4.2 (left) illustrates the distribution of patient follow-up across the dataset. A spike is observed near 0 years, although only 28 patients had no follow-up. Beyond this, the distribution approximates normality, with the longest follow-up reaching 25 years. Figure 4.2 (right) presents the distribution of all timestamp records in the cohort. Most events cluster around the AF diagnosis, with negative values representing patient history and positive values representing outcomes. Despite this concentration at the diagnosis date, only about 42,000 events occurred precisely on that day.

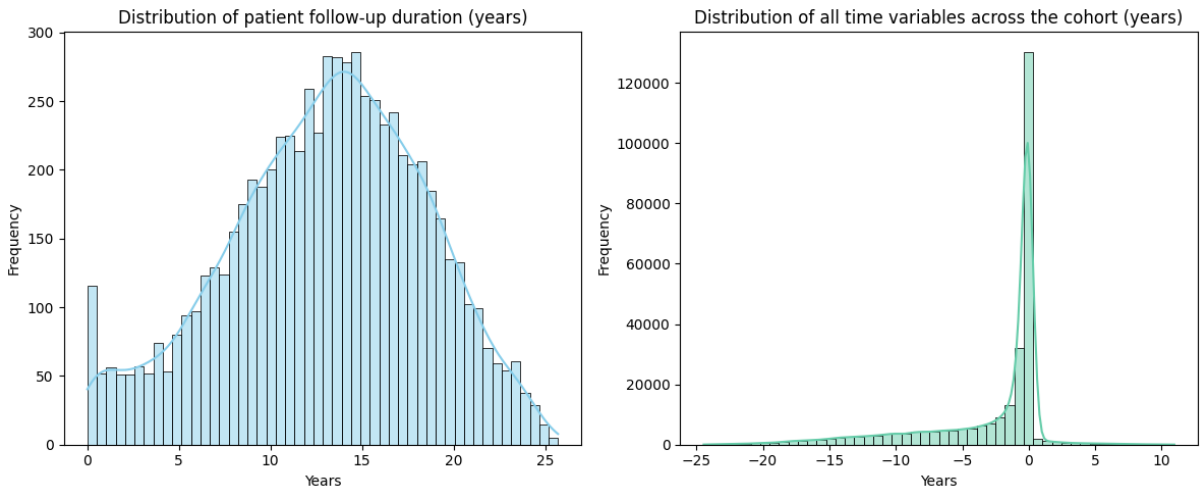

Figure 4.2: Distribution of patient follow-up duration across the cohort (left), and distribution of all time variables across the cohort (right)

To examine the longitudinal patterns of the outcome variables, Figure 4.3 presents the distributions, measured in days, of each outcome following the AF diagnosis, up to a two-year period. Most outcomes show a peak in frequency shortly after the diagnosis, which then gradually decreases over time.

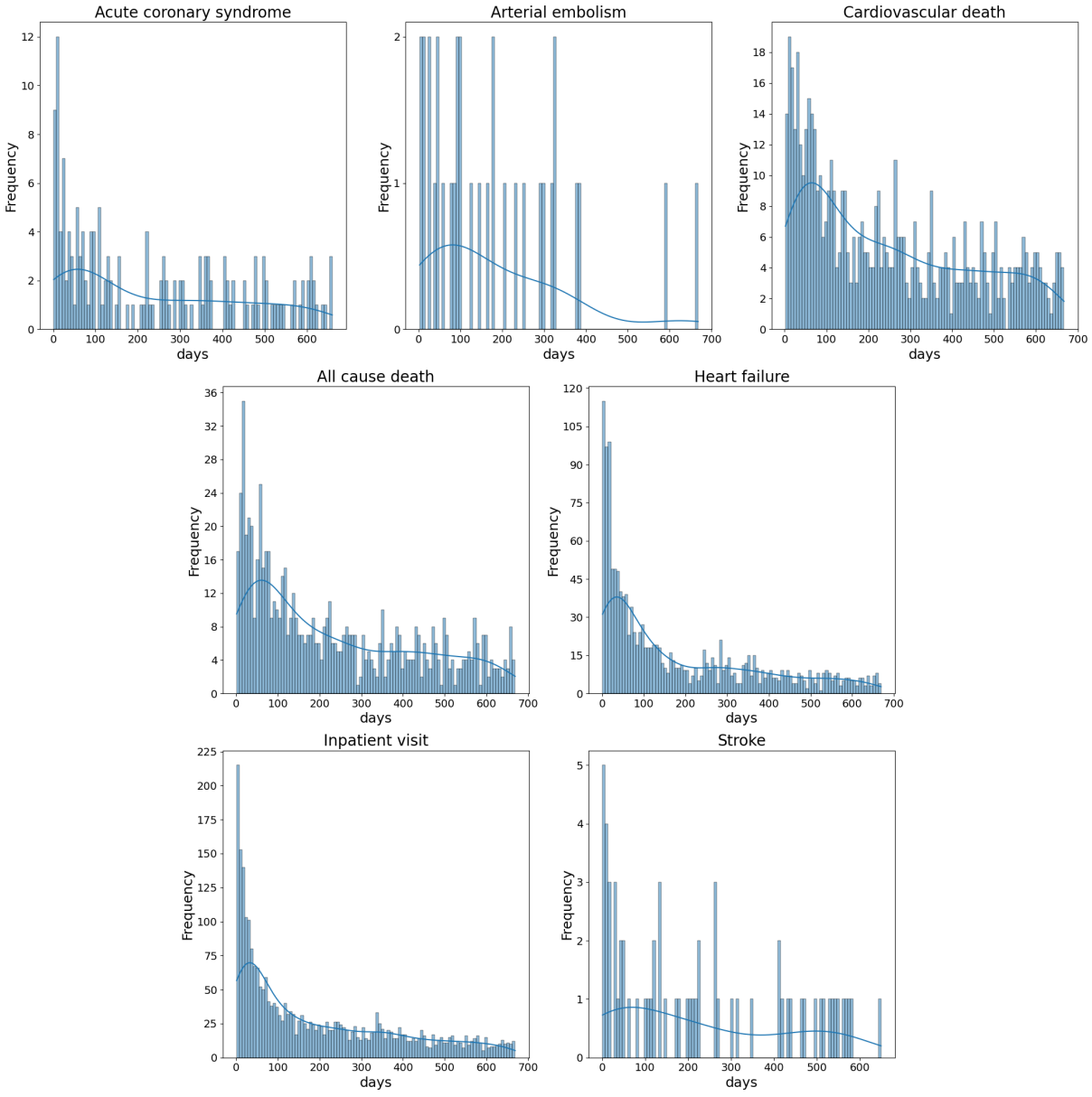

Figure 4.3: Distribution of elapsed time (in days) from AF diagnosis to each outcome, shown separately by outcome.

Complementing these visual distributions, Table 4.4 summarizes the outcome variables in terms of the number of occurrences and their corresponding percentages across predefined time intervals following AF diagnosis. Most outcomes are imbalanced, particularly within the 6-month interval. In addition, arterial embolism, stroke, and acute coronary syndrome are especially imbalanced.

Table 4.4: Distribution of outcome variables across different time intervals

Variable	1 month	3 months	6 months	1 year	2 years	Total
Acute coronary syndrome	34(0.5%)	57(0.8%)	83(1.2%)	116(1.6%)	176(2.5%)	306(4.3%)
Arterial embolism	6(0.1%)	13(0.2%)	21(0.3%)	29(0.4%)	35(0.5%)	52(0.7%)
Cardiovascular death	66(0.9%)	179(2.5%)	269(3.8%)	408(5.7%)	614(8.6%)	1280(17.8%)
All cause death	100(1.4%)	248(3.5%)	374(5.2%)	533(7.5%)	790 (11.0%)	1546(21.5%)
Heart failure hospitalization	382(5.3%)	680(9.5%)	904(12.6%)	1170(16.4%)	1500(21.0%)	2285(31.7%)
Inpatient visit	640(8.9%)	1175(16.4%)	1585(22.2%)	2149(30.0%)	2806(39.2%)	4188(58.1%)
Stroke	14(0.2%)	22(0.3%)	33(0.5%)	46(0.6%)	67(0.9%)	119(1.7%)

To complete the analysis, Figure 4.4 presents the correlations and corresponding p-values between a set of common risk indicators and the outcome variables, as well as their time-associated measurements. The figure shows that when an outcome exhibits a correlation, its time-associated counterpart generally also shows a correlation.

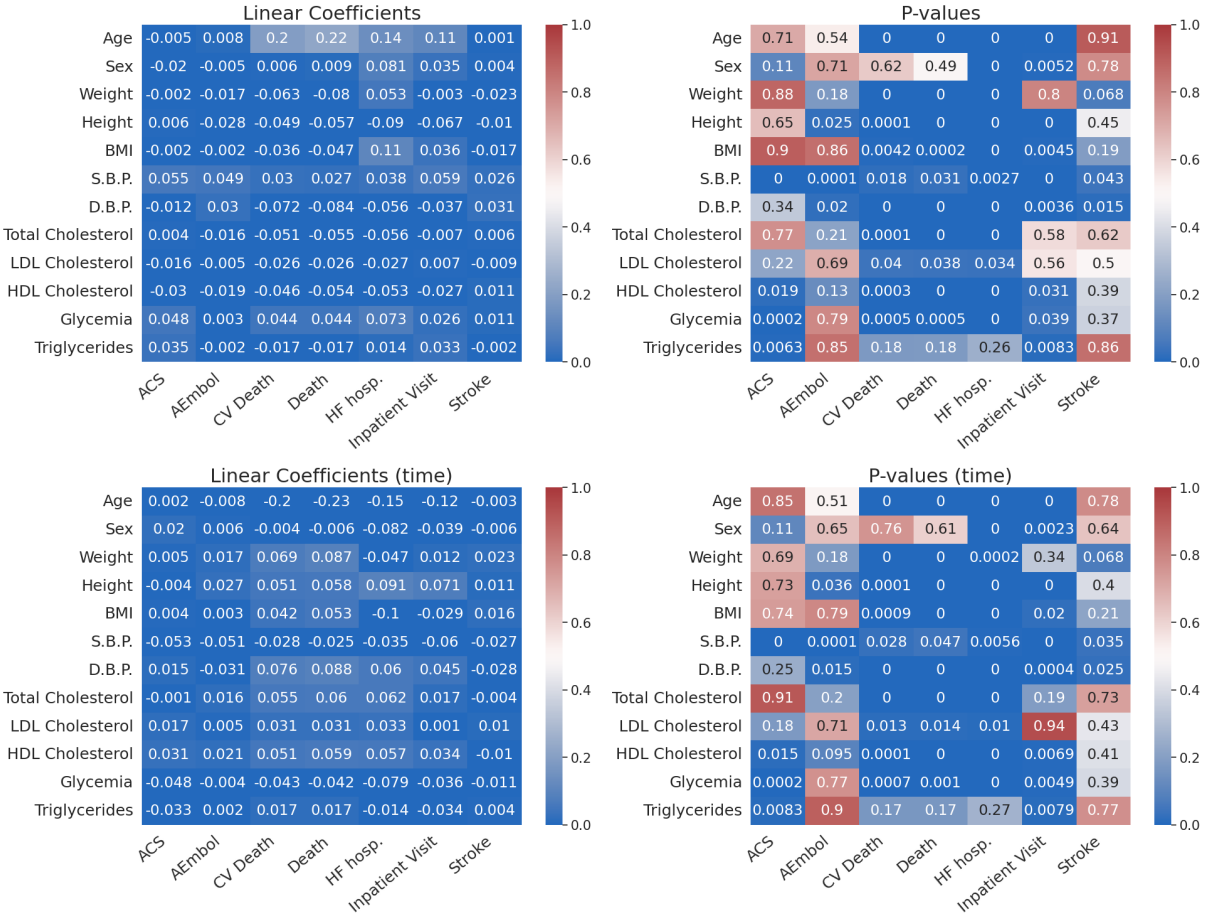

Figure 4.4: Correlation analysis of selected variables and outcomes. The top row shows the linear correlation coefficients (left) and corresponding p-values (right) for the selected variables and outcomes. The bottom row shows the correlations and p-values for the time variables associated with each outcome. Abbreviations used for the outcomes are: ACS = Acute Coronary Syndrome, AEmbol = Arterial Embolism, CV Death = Cardiovascular Death, HF hosp. = Heart Failure Hospitalization.

Overall, the variables tend to have stronger correlations with death or hospitalization outcomes. Very few

variables display meaningful correlations or statistically significant p-values for acute coronary syndrome, arterial embolism, or stroke, and the same pattern is observed for their associated time measurements.

5

Solution

Contents

This chapter outlines the methodology employed in this study. It begins with a description of the data preprocessing steps, including cleaning, transformation, and preparation for analysis. We then detail the implementation of the predictive models, ranging from classical calculators of AF and related events in AF cohorts, to machine learning predictors involving longitudinal stances.

5.1 Data Preprocessing

The data preprocessing workflow comprised the following sequence of steps: feature engineering, outlier detection and treatment, handling of missing values, encoding of categorical variables, addressing class imbalance, and feature selection.

To start the preprocessing, all time-related variables were recalculated relative to each patient's individual index date, so that each value represents the number of days before or after the AF diagnosis.

Additionally, the study cohort was intended to include only individuals aged over 40 years; however, 49 patients were identified as being younger than 40 and were consequently excluded, reducing the dataset from 7,203 to 7,154 observations.

Subsequently, outlier detection and treatment were performed. Time-related variables were capped within the range of -9,000 to 4,500 days relative to the AF diagnosis. This approach preserved the overall temporal distribution while eliminating 17 extreme outlier values. For other numerical variables, negative values were set to zero, since no variable could theoretically take on a negative score, and right-skewed variables were log-transformed using the log(1+x) function to approximate a normal distribution. Values falling outside the interval [Q_1–3IQR, Q_3+3IQR] were capped at the corresponding bounds; this capping strategy, rather than replacing outliers with null values, was found to optimize model performance.

The dataset contained a considerable amount of missing data of three main types: binary variables that were coded as 1 for the positive class and null for the negative class, numerical exam variables where the test was not prescribed or performed, and ordinary missing values. An initial inspection showed 873 observations without weight, height, or BMI, of which 680 were also missing systolic and diastolic blood pressure. Since these are key features, those 873 observations were removed. A further analysis of the 163 observations with missing BMI, weight or height revealed inconsistencies that made reliable imputation impossible, which led to the removal of these records as well, resulting in a total of 1,040 observations removed. For binary variables, nulls corresponding to the negative class were imputed as 0, meaning that the absence of data implied a negative case. The associated time variables were set to 10,000 when the binary variable was 0. After this preprocessing and the removal of the mentioned observations, most variables had less than 5% missing values, many with none at all. A few variables had between 5% and 20% missing values, including HDL cholesterol (8.5%), total cholesterol (8%), and triglycerides (8.2%), while others ranged from 20% to 40%, such as A1C, eGFR, LDL cholesterol, TSH, UACR, and several time variables linked to numeric features. No variable exceeded a 40% missing rate. Each of these was examined using ANOVA testing, considering both p-values and F-scores to assess significance before imputation. Although some variables appeared to be strong candidates for removal, all were retained due to the low dimensionality of the dataset and the possibility that variables deemed insignificant in univariate analysis could still contribute meaningfully in a multivariate context. Then, to impute the remaining values, KNN imputation was initially tested; however, it appeared to introduce unusual patterns in the data. To avoid distorting the medical information, all remaining missing values, including those of the time variables, were imputed using the mean. As an exception, missing values for the first measurement of numeric variables were imputed using the mean of the corresponding main numerical variable, and the same approach was applied to their corresponding time variables.

The only variable that required encoding was `age_hipaa`, which was ordinally encoded. Several variables were removed due to being constant, including `ep_pc` (primary care visit), `px_cas` (cardiac

surgery), px_cas_t cardiac surgery time, px_cos (coronary surgery), px_cs_t (coronary surgery time), and esrd_t (end stage renal disease time). The index_date variable was excluded because it was not relevant to the analysis, and the variables smk_nev (never smoked) and smk_nan (no information smoking) were also removed, as their information was already captured by the other smoking features. The variable ep_pc_sc (both primary and secondary visit) was removed because its information was also already captured by other visit features. The variables smk_nev_t (never smoked time), smk_nan_t (no information smoking time), and ep_pc_sc_t (both primary and secondary care time) were removed as they had no meaning. The feature History of Vascular Disease was created through coronary disease, myocardial infarction or unstable angina, and peripheral artery disease. Antihypertensive medication was created using the following prescriptions: angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, angiotensin receptor–neprilysin inhibitors, beta blockers, calcium channel blockers, loop diuretics, other diuretics, or mineralocorticoid receptor antagonists. Diabetes Mellitus was created from Type 1 Diabetes and Type 2 Diabetes. Diabetes medication was created using the following features: Metformin, SGLT2i, GLP1a, DPP4i, Insulin, and Sulfonylurea. Abnormal Kidney Function variable was created through having end-stage renal disease, or having abnormal levels of creatinine (> 1.3mg/dL), eGFR ($<$ mL/min/1.73 m^2), or uACR ($>$ 30 mg/g). From the temporal variables, each outcome was then transformed into a set of binary indicators representing whether the outcome occurred within 1, 3, 6, 12, or 24 months. Moreover, longitudinal features were derived to capture the variation over time in the following measurements: HbA1c, creatinine, eGFR, glycemia, HDL cholesterol, LDL cholesterol, total cholesterol, triglycerides, TSH, and UACR. These delta (change-over-time) features were calculated by taking the difference between the two measurements for each variable, dividing by the number of days between the measurements, and then multiplying by 365 to express the variation on a per-year basis.

Finally, with the aim of supporting the subsequent learning approaches, the preprocessed data was used to produce three dataset versions. The *static* version excluded time-dependent features and retained only the primary measurement for variables with multiple entries. The *slope-based* version was similar to the static version but extended with the delta features. The *longitudinal* version included the complete dataset with all available features.

Class imbalance was addressed by evaluating different sampling strategies for each outcome-model--dataset combination, including baseline (no sampling), random undersampling, random oversampling, and SMOTE. The strategy that achieved the best performance was selected for each case. To prevent excessive data loss in highly imbalanced outcomes—particularly in certain time intervals—undersampling was limited so that at least 10% of the original dataset was retained. The stroke and arterial embolism outcomes were especially unbalanced and were combined across all time intervals to create a single composite endpoint, improving the stability and reliability of model training.

Finally, at the modeling stage, the training and test datasets were scaled using StandardScaler

parameterized on the training data. This approach ensures that information from the test set does not influence the training data. Standardization was only applied to algorithms sensitive to feature scales, such as Naïve Bayes, Logistic Regression, and Multi-Layer Perceptron.

5.2 Classical Risk Calculators

To evaluate the performance of classical risk calculators within the AF cohort, three representative models were selected: CHA_2DS_2-VASc, which follows a point-based system, and CHARGE-AF and GARFIELD-AF, which are based on Cox regression. CHARGE-AF was chosen to predict AF due to its demonstrated success in previous studies and its compatibility with the clinical features available in our dataset, enabling a meaningful comparison. CHA_2DS_2-VASc and GARFIELD-AF were selected for their ability to predict stroke, with GARFIELD-AF also capable of predicting mortality in AF cohorts. These scores serve as a baseline for comparison with subsequently developed machine learning models, as they are specifically tailored for these comorbidities in AF cohorts. To ensure that the calculators provide true predictions of future AF risk, they were applied using only data available prior to AF diagnosis, without relying on any information obtained after the event.

5.2.1 CHARGE-AF

To assess whether the cohort exhibits strong indicators of AF internally, the CHARGE-AF score was computed in the cohort, predicting the risk of developing AF in 5 years. The calculus of the CHARGE-AF score requires information on the patient's age, ethnicity, height, weight, systolic and diastolic blood pressure, smoking status (current), use of antihypertensive medication, presence of diabetes, history of heart failure, and myocardial infarction. All variables were obtainable from the dataset, except ethnicity. Given that the cohort is Portuguese, participants were categorized as Caucasian by default. Directly available features include height, weight, systolic and diastolic blood pressure, current smoking status, heart failure, and myocardial infarction. Antihypertensive medication and diabetes were also available as they were derived from other features. Age was the only feature that required additional processing, as it was provided in ranges due to HIPAA compliance. To calculate it, each range was replaced with its midpoint (e.g., the range 70–79 was represented as 75). The score was then computed using the original regression coefficients [1], summarized in Table 5.1, together with the Cox proportional hazards formula (Equation 5.1).

Table 5.1: Original coefficients for final multivariable model for 5-year risk of AF CHARGE-AF [1]

Variable	Estimated β
Age (5 years)	0.508
Race (white)	0.465
Height (10 cm)	0.248
Weight (15 kg)	0.115
Systolic BP (20 mm Hg)	0.197
Diastolic BP (10 mm Hg)	-0.101
Smoking (current)	0.359
Antihypertensive medication use (Yes)	0.349
Diabetes (Yes)	0.237
Heart failure (Yes)	0.701
Myocardial infarction (Yes)	0.496

$$\text{CHARGE-AF} = 1 - \left(0.9718812736^{\exp(\sum \beta X - 12.58156)}\right) \tag{5.1}$$

1. β is the regression coefficient associated with each risk factor.

2. X is the level (value) for each corresponding risk factor.

5.2.2 CHA_2DS_2-VASc

To access stroke risk in the cohort CHA_2DS_2-VASc score is further calculated, estimating stroke risk over a 1-year period. CHA_2DS_2-VASc is a risk-stratification score ranging from 0 to 9, depending on the number and weight of the score's risk components. Table 5.2 shows each risk factor and the corresponding score.

Table 5.2: CHA_2DS_2-VASc risk factors and correspondent score [2]

Risk Factor	Score
Congestive heart failure (C)	1
Hypertension (H)	1
Age 75+ years (A)	2
Diabetes mellitus (D)	1
Stroke (S)	2
Vascular disease (V)	1
Age 65-74 years (A)	1
Sex category - female (Sc)	1

* If a woman is assigned only one point due to sex category, her score should be adjusted to zero.

The dataset contains all necessary information to compute CHA_2DS_2-VASc score. Data on congestive heart failure (equivalent to heart failure), hypertension, diabetes mellitus, stroke, and sex are directly available. Vascular disease was derived in the data preprocessing. Age was calculated using the middle value of the category interval: patients aged 60–69 were assigned a score of 1 (using 65 as the representative age), while those aged 70–79 or older were assigned a score of 2.

5.2.3 GARFIELD-AF

The GARFIELD-AF score was calculated for all patients in the cohort to evaluate their risk of stroke and all-cause mortality. Calculation of the GARFIELD-AF score requires the following variables: age, sex, ethnicity, diastolic blood pressure, pulse, history of heart failure, vascular disease, prior stroke, history of bleeding, diabetes, moderate-to-severe chronic kidney disease, dementia, current smoking, and use of vitamin K antagonist (VKA) or non–vitamin K antagonist oral anticoagulants (NOACs). Among these, sex, diastolic blood pressure, history of heart failure, prior stroke, current smoking, and chronic kidney disease were directly available from the dataset. Vascular disease and diabetes were derived during data preprocessing. Age was estimated as the midpoint of each age category, similarly to the other classical calculators. Variables not captured in the dataset were handled with assumptions: all patients were considered Caucasian, pulse was set at 80 bpm, and patients were assumed to have no history of bleeding or dementia. Regarding anticoagulant use, VKA treatment was assumed absent, and NOAC treatment was considered present for patients recorded as taking any anticoagulant. The score for each outcome was then computed using the original regression coefficients, summarized in Table 5.3, together with the Cox proportional hazards formulas (Equation 5.2),

$$\text{GARFIELD-AF all-cause mortality} = 1 - 0.987921904^{\exp(\sum \beta X)}, \tag{5.2}$$

for all-cause mortality, and Equation 5.3,

$$\text{GARFIELD-AF ischemic stroke/SE} = 1 - 0.9925445321^{\exp(\sum \beta X)}, \tag{5.3}$$

1. β is the regression coefficient associated with each risk factor.

2. X is the value or presence of each corresponding risk factor, with the respective adjustments.

3. All-cause mortality equation is for estimating risk at 6 months, and the ischemic stroke/SE formula is for estimating risk at 1 year. 4. The original GARFIELD-AF equations are expressed as percentages and are therefore multiplied by 100.

for ischemic stroke or systemic embolism. Other GARFIELD-AF formulas can be found at https://af.garfieldregistry.org/garfield-af-risk-calculator.

Table 5.3: GARFIELD-AF model coefficients for 6-month all-cause mortality and 1-year ischemic stroke/systemic embolism

Variable	All-cause mortality β	Ischemic stroke/SE β
Female sex	-0.306202287	-
Heart failure	0.693789082	0.233182644
Vascular disease	0.306120964	0.197919709
Prior stroke	0.265852980	0.800863063
History of bleeding	0.385407386	0.298839670
Diabetes	0.280133213	0.211995445
Moderate-to-severe CKD	0.377903886	0.349516938
Dementia	0.489453313	0.513221391
Current smoking	0.345481149	0.478831506
OAC treatment: NOAC	-0.414591263	-0.572199357
OAC treatment: VKA	-0.185935610	-0.352373263
Ethnicity: Hispanic/Latino	0.157023564	-
Ethnicity: Asian	-0.609609055	-
Ethnicity: Black/Mixed/Other	0.375675102	-
(Age-65) (if ≤ 65)	0.031050027	0.039138147
(Age-65) (if > 65)	0.064594824	-
(Weight-75) (if ≤ 75)	-0.021535182	-
(Pulse-120) (if ≤ 120)	0.007678035	-
(Diastolic BP-80) (if ≤ 80)	-0.019304333	-
(Diastolic BP-80) (if > 80)	-	0.015900160

5.3 Machine Learning Predictors

To predict the outcomes, a set of well-established machine learning models was implemented. These included Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, XGBoost, and a Multi-Layer Perceptron (MLP). This selection encompasses both classical statistical methods and modern ensemble and neural network approaches, allowing for a comprehensive evaluation of predictive performance across different modeling strategies.

Each model was trained for every outcome variable at the 6-month time horizon, except for stroke combined with arterial embolism, which was trained using data from the full 10-year cohort due to class imbalance. Models were trained separately for each dataset type—static, slope-based, and longitudinal—and both independently for each outcome to avoid multi-target bias, as well as jointly in a multi-target framework.

Robustness was ensured through 5-fold cross-validation, with model hyperparameters optimized via Bayesian optimization using the F_2 score (Equation 5.4),

$$F_2 = (1 + 2^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(2^2 \cdot \text{Precision}) + \text{Recall}}, \tag{5.4}$$

as the objective, and final selection was also guided by the F_2 score during sampling. Using the F-

measure with $\beta = 2$ gives greater weight to recall, ensuring higher sensitivity to false negatives compared to false positives. For each outcome, the best-performing predictor was identified through rank fusion of F_1, F_2, and AUC (Equation 5.5),

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}). \tag{5.5}$$

1. TPR: True positive rate

2. FPR: False positive rate

The best-performing model was subsequently used to assess performance across all time horizons and subjected to explainability analysis using SHAP-based feature importance.

6

Results and Evaluation

Contents

This chapter presents the results of the proposed predictive pipelines targeting atrial fibrillation and related outcomes. We first evaluate a classical AF risk score within the cohort, and then examine both traditional risk scores and machine learning models across multiple outcomes to identify the best-performing predictors. Additionally, models are trained on the previously introduced dataset versions to assess the influence of temporal information on predictive performance.

6.1 Classical AF Risk Calculator: CHARGE-AF

In the cohort, the CHARGE-AF score had a mean of 0.2 risk, with a standard deviation of 0.19. Risk categories are defined as follows: low risk < 0.025, medium risk 0.025–0.05, and high risk > 0.05 for developing AF within five years [1]. The distribution of patients across these categories is presented in Table 6.1.

Table 6.1: Distribution of CHARGE-AF risk categories in the study cohort

Risk category	Count	Percentage
Low risk (<2.5%)	658	10.5%
Moderate risk (2.5% to 5%)	676	10.8%
High risk (>5%)	4889	78.7%

CHARGE-AF classified only 10.5% of AF patients as low risk, whereas 78.7% were categorized as high risk. This indicates that our cohort has a high prevalence of AF risk factors, with the majority of patients identified as high risk. Only a small proportion fell into the low or medium-risk categories, suggesting that the score effectively highlights individuals with an elevated likelihood of developing AF within five years.

Figure 6.1 shows the distribution of CHARGE-AF scores in the target AF cohort. The left panel presents the full histogram, while the right panel focuses on the clinically relevant range below 0.25. For reference, Figure 6.2 shows the histogram of scores from the CHARGE-AF derivation cohort. Comparing the two distributions highlights differences in score ranges and density patterns between the analyzed AF cohort and the community-based cohort that is not restricted to AF.

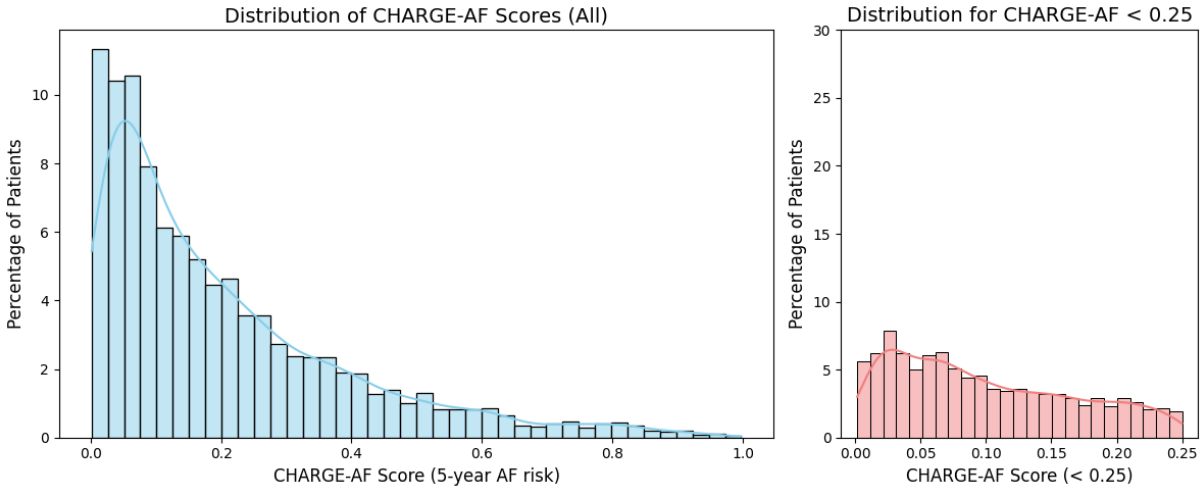

Figure 6.1: Distribution of CHARGE-AF scores in the cohort

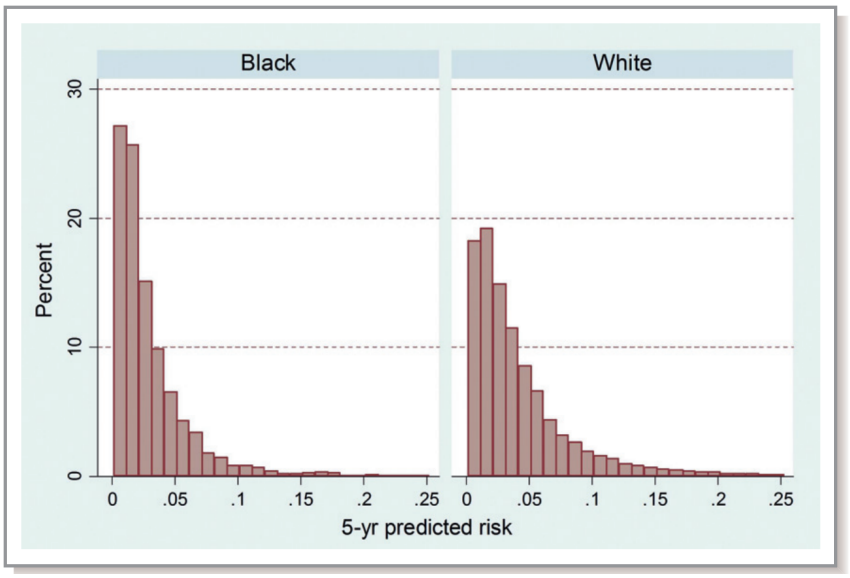

Figure 6.2: Distribution of CHARGE-AF scores in the CHARGE-AF derivation cohorts (ARIC, CHS and FHS), separated by ethnicity. Source: [1]

6.2 Prediction of AF-based clinical outcomes

This section assesses the capacity of several machine learning models for predicting six clinical endpoints under three major data settings: (i) static, (ii) slope-based, and (iii) longitudinal; described in depth in Chapter 5.1. Model performance is compared against classical methods for stroke and all-cause mortality outcomes. We also experimented with multi-label classification, but the results were substantially poorer and are therefore not reported. Each model was tuned using Bayesian optimization with the F_2 score as the objective, and also selected based on the F_2 score during sampling. The best-performing models were then selected using a rank fusion of F_1, F_2, and AUC, and were carried forward for further analyses, including feature importance assessment and outcome prediction across additional time horizons.

6.2.1 Stroke and Arterial Embolism Outcomes

The CHA_2DS_2-VASc score was developed to estimate 1-year stroke or systemic embolism risk in patients with atrial fibrillation, with traditional risk categories defined as 0 for low risk, 1 for intermediate risk, and $\geq$2 for high risk [2]. According to the 2020 ESC guidelines for the diagnosis and management of AF, risk stratification should be interpreted in a sex-specific manner: low risk (score = 0 in men, or 1 in women), for whom antithrombotic therapy is generally not recommended; intermediate risk (score = 1 in men, or 2 in women), for whom oral anticoagulation (OAC) may be considered; and high risk (score $\geq$2 in men, or $\geq$3 in women), for whom OAC is recommended [135].

In the target cohort, the distribution of CHA_2DS_2-VASc scores is shown in Figure 6.3 (left), with a mean of 4.09 and a standard deviation of 1.44. Figure 6.3 (right) shows the corresponding ROC curve, with an area under the curve (AUC) of 0.588. Most patients had elevated scores, with approximately 94% classified as high risk, 5% as intermediate risk, and 1% as low risk according to the ESC 2020 thresholds. These results are consistent with the presence of a high-risk, frail population and indicate that our dataset contains strong predictors of stroke and systemic embolism.

The ROC curve results were also favorable, with the CHA_2DS_2-VASc score achieving an AUC comparable to that of its original derivation cohort (0.588 vs. 0.606). However, it is important to note that the original derivation cohort [2] primarily included patients not receiving anticoagulants, although antiplatelet therapy was allowed. In contrast, a substantial proportion of our cohort (44%) was on anticoagulants, in addition to other cardiovascular medications, which can significantly alter individual stroke risk profiles.

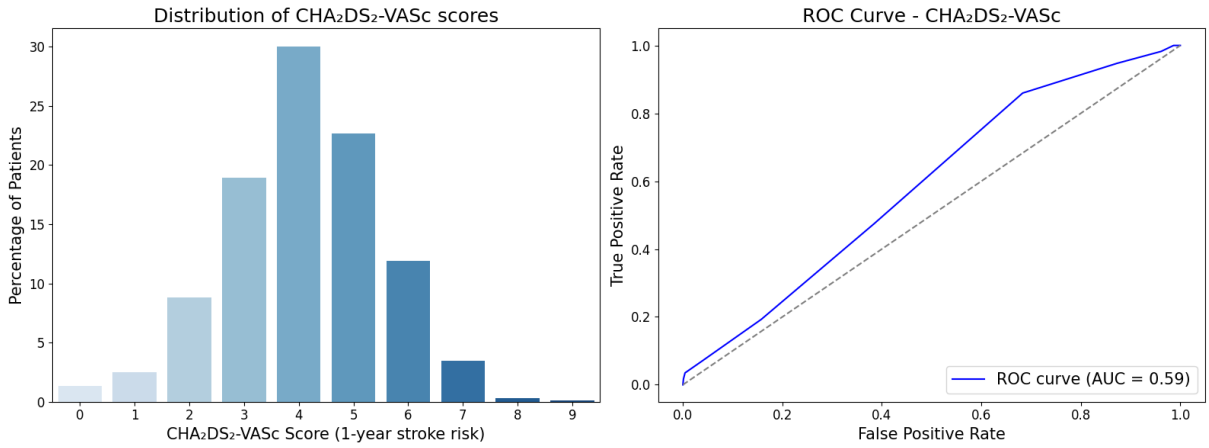

Figure 6.3: Distribution of CHA_2DS_2-VASc scores in the conducted cohort, and AUC for predicting stroke or systemic embolism at 1 year

In Figure 6.4, the left panel shows the distribution of GARFIELD-AF risk estimates for 1-year ischemic stroke or systemic embolism across the cohort, and the right panel shows the corresponding ROC curve. Although GARFIELD-AF has not defined formal risk categories, higher deciles of predicted risk correspond to increased observed stroke incidence [128]. GARFIELD-AF also demonstrates superior predictive performance compared with CHA_2DS_2-VASc, achieving a higher AUC both in the derivation cohort (0.65 vs. 0.59) and in our treated cohort (0.633 vs. 0.588). In this task, an AUC of 0.63 in a validation cohort represents a strong performance, demonstrating the efficacy of GARFIELD-AF in predicting stroke and systemic embolism. It is also noteworthy that several variables required to calculate GARFIELD-AF—such as pulse, ethnicity, bleeding history, dementia, and type of anticoagulant (NOAC vs. VKA)—had to be approximated or inferred from our available dataset, which may have affected

the predictive accuracy of the score. Nevertheless, the good AUCs observed for both CHA_2DS_2-VASc and GARFIELD-AF demonstrate the strength of these predictors, and indicate that the dataset provides reliable indicators of stroke and arterial embolism risk.

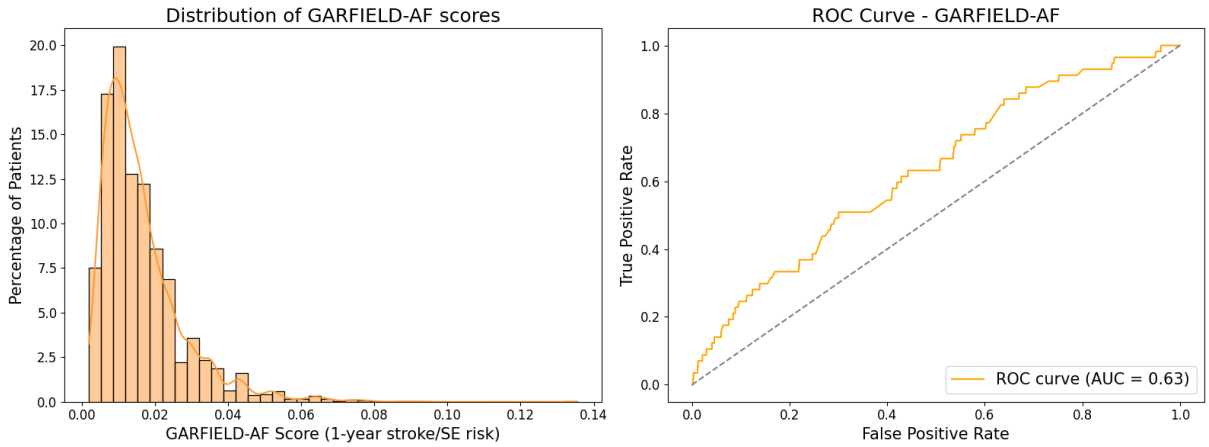

Figure 6.4: GARFIELD-AF Area Under the Operating Curve for predicting stroke and artery embolism at 1-year

Table 6.2 summarizes the performance of the hyperparameterized learning models in predicting stroke or systemic embolism over a 10-year period. The best-performing model was Logistic Regression (LR), with the Multi-Layer Perceptron (MLP) and XGBoost trailing behind. Among the models, the static XGBoost achieved the highest precision, while the slope-based LR attained the highest F_1 and F_2 scores, reflecting a strong balance between precision and sensitivity. Additionally, LR achieved the highest AUC (0.634) and the lowest NNS (29.24), highlighting its superior overall performance.

Table 6.2: Performance of machine learning models in predicting stroke and systemic embolism, reported as mean ± standard deviation for accuracy, precision, sensitivity, F_1 score, F_2 score, AUC, and NNS.

Model	Accuracy	Precision	Sensitivity	F_1 score	F_2 score	AUC	NNS	Rank
Static NB	0.369 ± 0.264	0.024 ± 0.002	0.667 ± 0.280	0.045 ± 0.002	0.100 ± 0.008	0.570 ± 0.053	41.97 ± 3.17	11
Static LR	0.522 ± 0.017	0.030 ± 0.001	0.640 ± 0.028	0.056 ± 0.003	0.125 ± 0.006	0.605 ± 0.022	33.93 ± 1.76	3
Static DT	0.763 ± 0.033	0.021 ± 0.008	0.206 ± 0.075	0.038 ± 0.014	0.074 ± 0.027	0.515 ± 0.034	61.25 ± 37.04	17
Static RF	0.761 ± 0.016	0.016 ± 0.005	0.161 ± 0.057	0.029 ± 0.010	0.057 ± 0.019	0.478 ± 0.018	70.72 ± 24.83	18
Static XGB	0.908 ± 0.019	**0.038 ± 0.018**	0.117 ± 0.051	0.056 ± 0.024	0.080 ± 0.033	0.576 ± 0.029	33.63 ± 15.25	6
Static MLP	0.859 ± 0.018	0.036 ± 0.012	0.205 ± 0.076	0.061 ± 0.021	0.106 ± 0.037	0.574 ± 0.043	32.59 ± 15.23	4
Slope-based NB	0.131 ± 0.036	0.023 ± 0.001	**0.904 ± 0.076**	0.044 ± 0.002	0.103 ± 0.005	0.529 ± 0.039	44.11 ± 2.09	12
Slope-based LR	0.612 ± 0.031	0.034 ± 0.002	0.603 ± 0.059	**0.065 ± 0.004**	**0.140 ± 0.008**	**0.634 ± 0.022**	**29.24 ± 1.70**	**1**
Slope-based DT	0.784 ± 0.028	0.023 ± 0.006	0.220 ± 0.080	0.042 ± 0.011	0.082 ± 0.024	0.503 ± 0.034	45.48 ± 11.42	16
Slope-based RF	0.788 ± 0.026	0.026 ± 0.007	0.243 ± 0.089	0.047 ± 0.013	0.091 ± 0.027	0.538 ± 0.035	41.21 ± 12.02	13
Slope-based XGB	0.898 ± 0.011	0.033 ± 0.010	0.125 ± 0.038	0.052 ± 0.016	0.080 ± 0.024	0.574 ± 0.026	33.96 ± 13.31	9
Slope-based MLP	0.856 ± 0.014	0.035 ± 0.008	0.205 ± 0.047	0.060 ± 0.013	0.104 ± 0.023	0.575 ± 0.033	29.91 ± 6.45	5
Longitudinal NB	0.266 ± 0.096	0.025 ± 0.002	0.824 ± 0.078	0.048 ± 0.004	0.110 ± 0.009	0.562 ± 0.051	40.71 ± 3.31	8
Longitudinal LR	0.774 ± 0.008	0.034 ± 0.008	0.338 ± 0.071	0.063 ± 0.014	0.122 ± 0.027	0.611 ± 0.032	30.65 ± 7.45	2
Longitudinal DT	0.801 ± 0.020	0.025 ± 0.014	0.214 ± 0.130	0.044 ± 0.025	0.084 ± 0.048	0.561 ± 0.043	74.32 ± 73.85	15
Longitudinal RF	0.779 ± 0.022	0.029 ± 0.008	0.273 ± 0.081	0.052 ± 0.015	0.101 ± 0.029	0.534 ± 0.029	37.37 ± 9.72	10
Longitudinal XGB	**0.915 ± 0.021**	0.031 ± 0.005	0.096 ± 0.045	0.045 ± 0.011	0.065 ± 0.023	0.567 ± 0.066	33.60 ± 5.50	14
Longitudinal MLP	0.857 ± 0.006	0.029 ± 0.008	0.169 ± 0.049	0.050 ± 0.014	0.086 ± 0.025	0.591 ± 0.032	37.45 ± 11.74	7

Note: Undersampling was applied to Decision Tree, Random Forest, XGBoost, and Multi-Layer Perceptron models. Naive Bayes was used as a baseline, while Logistic Regression employed oversampling for class balancing.

The Naive Bayes demonstrated high sensitivity, while the remaining models showed comparatively

limited performance, generally characterized by low precision and sensitivity. Incorporating slope and other longitudinal features enhanced model performance, highlighting the added value of temporal variables in predicting stroke or systemic embolism.

Figure 6.5 presents the SHAP feature importance for the slope-based Logistic Regression model, which was the best-performing model. The analysis suggests that shorter height is associated with an increased risk of stroke or systemic embolism. Although height is not traditionally recognized as a stroke/SE risk factor, prior studies have linked shorter stature to cardiovascular disease or stroke, proposing that it may serve as a marker of early-life social and physical conditions [136].

Low body weight and BMI also emerged as predictors of increased risk, whereas higher values seemed protective—a pattern consistent with the so-called “obesity paradox”. This paradox is thought to arise from factors such as malnutrition, frailty, and the limitations of BMI and weight in distinguishing fat from lean mass.

Blood pressure was another strong predictor, with both systolic and diastolic elevations associated with higher risk. Elevated systolic pressure is well established as a stroke risk factor, while diastolic pressure is generally less predictive; however, both contributed meaningfully in this model.

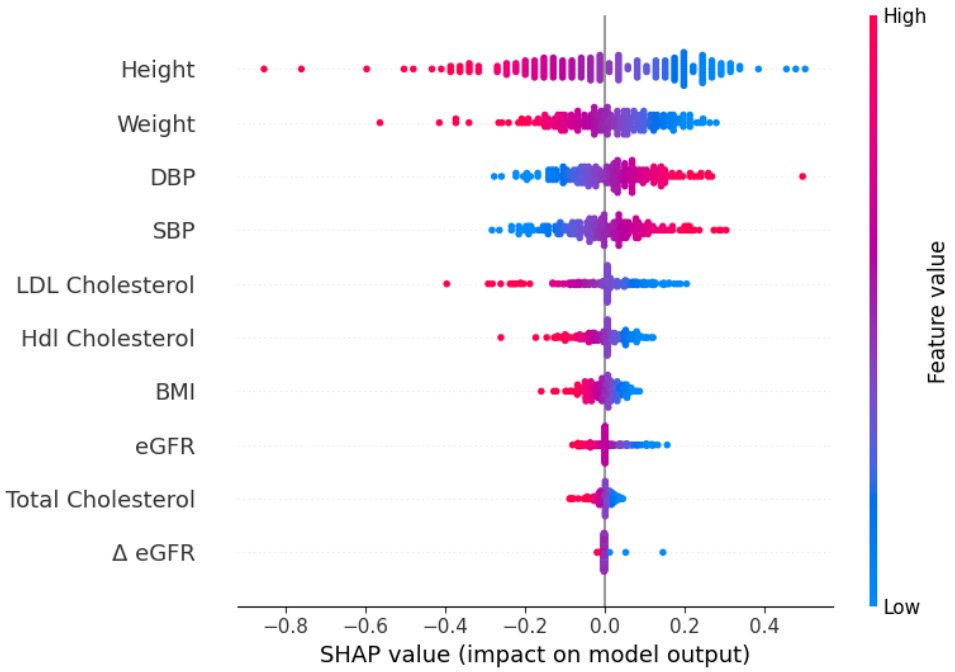

Figure 6.5: SHAP value of slope-based Logistic Regression model on predicting stroke and arterial embolism

Interestingly, lower LDL, HDL, and total cholesterol levels were also linked to a higher risk. While elevated LDL is classically considered harmful, the inverse association observed here may reflect the widespread use of statins in this population, which lowers LDL and therefore total cholesterol. Similar

"cholesterol paradox" findings have been reported in older, frail, and atrial fibrillation cohorts [137].

Additionally, lower eGFR values were associated with a higher risk, consistent with existing evidence that impaired kidney function contributes to stroke and embolism. Changes in eGFR over time also influenced model performance, with declining eGFR linked to additional risk, in line with baseline eGFR findings. Although most slope features did not rank among the most important predictors in the model, they nonetheless made a meaningful contribution to risk prediction.

Age, despite being a well-established predictor, did not rank among the strongest features in this model. This may reflect overlap with other clinical variables such as blood pressure, renal function, and other measures, which may have captured much of the age-related risk.

The model was subsequently trained across additional time horizons, including 1 year, to assess its performance and compare it with CHA_2DS_2-VASc and GARFIELD-AF. The results for these time ranges are presented in Table 6.3. Model performance increases from 6 months to 1 year and is similar at 1 year and 2 years; however, the highest AUC is observed at 1 year. The 1-month and 3-month horizons were excluded, as the degree of class imbalance prevented the generation of reliable predictions.

Table 6.3: Performance of slope-based Logistic Regression model in predicting stroke and systemic embolism at different time intervals, reported as mean ± standard deviation for accuracy, precision, sensitivity, F_1 score, F_2 score, AUC, and NNS.

Interval	Accuracy	Precision	Sensitivity	F_1 score	F_2 score	AUC	NNS
6 months	0.786 ± 0.033	0.010 ± 0.004	0.339 ± 0.169	0.020 ± 0.007	0.046 ± 0.017	0.582 ± 0.078	113.80 ± 53.30
1 year	0.758 ± 0.030	0.018 ± 0.007	0.456 ± 0.209	0.034 ± 0.014	0.076 ± 0.031	0.651 ± 0.098	95.705 ± 96.187
2 years	0.686 ± 0.044	0.018 ± 0.002	0.425 ± 0.092	0.034 ± 0.004	0.076 ± 0.010	0.562 ± 0.055	57.130 ± 6.280

Note: SMOTE was applied to 6 months interval, undersampling was applied to 1 year interval, and oversampling was applied to 2 year interval.

The static Logistic Regression model achieved only modest predictive performance, with very low precision and sensitivity. Nevertheless, its AUC for 1-year stroke risk (0.651) exceeded that of CHA_2DS_2-VASc (0.588) and GARFIELD-AF (0.633), outperforming the classical risk calculators. However, the weaker results at the 6-month and 2-year horizons indicate that the predictions are not yet reliable, and that addressing the severe data imbalance will likely require either a specifically tailored algorithm or a larger dataset to achieve both robust performance and greater reliability.

6.2.2 All-Cause Death Outcome

Figure 6.6 shows the distribution of GARFIELD-AF predictions at 6 months in the cohort (left panel) and the corresponding ROC curve (right panel). The distribution indicates that several patients are at high risk ($\geq$ 10%) of mortality within the next six months, with some exhibiting even higher predicted probabilities. An AUC of 0.72 demonstrates good discriminatory ability, indicating that the model can reliably differentiate between patients who did and did not experience all-cause mortality within six months.

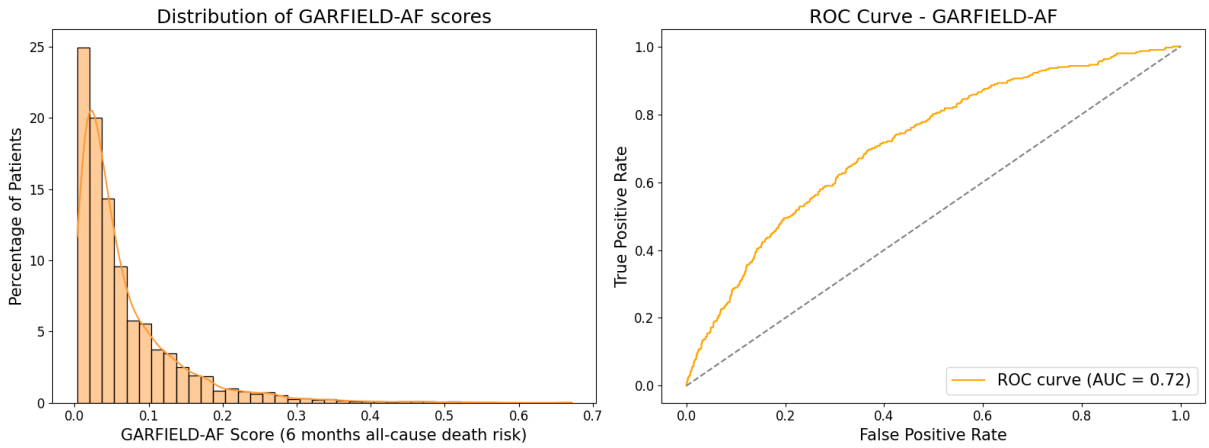

Figure 6.6: GARFIELD-AF Area Under the Operating Curve for predicting all-cause death at 6 months

Table 6.4 presents the performance of the ML models in predicting all-cause death at 6 months. XGBoost, Random Forest, and Logistic Regression achieved the best overall performance, combining high sensitivity with strong precision. Other models performed less well: Naive Bayes showed higher sensitivity but at the expense of precision, while the MLP and Decision Tree exhibited generally poor performance. Incorporating slope features did not improve predictive ability, whereas including the full set of longitudinal features enhanced model performance, indicating that these longitudinal features contain valuable predictive information.

Table 6.4: Performance of machine learning models in predicting all-cause death at 6 months, reported as mean ± standard deviation for accuracy, precision, sensitivity, F_1 score, F_2 score, AUC, and NNS.

Model	Accuracy	Precision	Sensitivity	F_1 score	F_2 score	AUC	NNS	Rank
Static NB	0.662 ± 0.013	0.092 ± 0.008	0.663 ± 0.051	0.161 ± 0.014	0.295 ± 0.024	0.705 ± 0.024	10.99 ± 0.90	15
Static LR	0.615 ± 0.032	0.096 ± 0.004	0.815 ± 0.063	0.171 ± 0.006	0.325 ± 0.011	0.774 ± 0.020	10.45 ± 0.42	4
Static DT	0.637 ± 0.037	0.092 ± 0.011	0.717 ± 0.069	0.162 ± 0.018	0.303 ± 0.030	0.691 ± 0.038	11.04 ± 1.19	14
Static RF	0.654 ± 0.043	0.102 ± 0.009	0.771 ± 0.051	0.180 ± 0.015	0.332 ± 0.023	0.769 ± 0.024	9.91 ± 0.98	5
Static XGB	0.662 ± 0.043	0.100 ± 0.012	0.730 ± 0.035	0.176 ± 0.018	0.322 ± 0.024	0.762 ± 0.014	10.12 ± 1.17	9
Static MLP	0.671 ± 0.012	0.096 ± 0.006	0.680 ± 0.046	0.168 ± 0.010	0.306 ± 0.019	0.727 ± 0.020	10.48 ± 0.63	10
Slope-based NB	0.276 ± 0.023	0.059 ± 0.003	**0.922 ± 0.039**	0.111 ± 0.006	0.234 ± 0.012	0.694 ± 0.027	17.05 ± 0.99	18
Slope-based LR	0.610 ± 0.018	0.095 ± 0.003	0.822 ± 0.056	0.170 ± 0.006	0.325 ± 0.013	0.772 ± 0.021	10.54 ± 0.36	6
Slope-based DT	0.597 ± 0.048	0.084 ± 0.004	0.737 ± 0.097	0.151 ± 0.007	0.288 ± 0.017	0.686 ± 0.031	11.87 ± 0.53	17
Slope-based RF	0.647 ± 0.041	0.099 ± 0.012	0.761 ± 0.077	0.175 ± 0.020	0.324 ± 0.034	0.767 ± 0.034	10.28 ± 1.25	8
Slope-based XGB	0.662 ± 0.024	0.101 ± 0.008	0.751 ± 0.053	0.179 ± 0.014	0.329 ± 0.023	0.755 ± 0.037	9.92 ± 0.83	7
Slope-based MLP	0.673 ± 0.014	0.094 ± 0.004	0.656 ± 0.029	0.164 ± 0.006	0.298 ± 0.010	0.716 ± 0.018	10.71 ± 0.42	6
Longitudinal NB	0.452 ± 0.030	0.073 ± 0.006	0.865 ± 0.024	0.134 ± 0.010	0.272 ± 0.017	0.708 ± 0.024	13.86 ± 1.09	16
Longitudinal LR	0.655 ± 0.020	0.102 ± 0.005	0.778 ± 0.035	0.180 ± 0.007	0.334 ± 0.011	0.774 ± 0.011	9.82 ± 0.43	2
Longitudinal DT	0.633 ± 0.049	0.094 ± 0.011	0.744 ± 0.070	0.166 ± 0.017	0.310 ± 0.027	0.690 ± 0.039	10.80 ± 1.16	11
Longitudinal RF	0.659 ± 0.040	0.105 ± 0.017	0.778 ± 0.068	0.184 ± 0.029	0.339 ± 0.047	0.753 ± 0.068	9.85 ± 1.72	3
Longitudinal XGB	0.695 ± 0.021	**0.111 ± 0.012**	0.741 ± 0.051	**0.192 ± 0.020**	**0.346 ± 0.033**	**0.779 ± 0.031**	**9.16 ± 1.04**	**1**
Longitudinal MLP	**0.680 ± 0.011**	0.091 ± 0.008	0.623 ± 0.052	0.159 ± 0.014	0.288 ± 0.025	0.723 ± 0.019	11.02 ± 0.97	13

Note: Undersampling was applied to Decision Tree, Random Forest, XGBoost, and Multi-Layer Perceptron models. Naive Bayes was used as a baseline, while Logistic Regression employed SMOTE for class balancing.

The best-performing model was the longitudinal XGBoost, and its AUC for 6-month all-cause death risk (0.779) exceeded that of GARFIELD-AF (0.715), showing once again the efficacy of machine learning.

The longitudinal XGBoost has its feature importance presented in Figure 6.7. Age emerged as the strongest predictor, consistent with its role as a major risk factor, with patients aged 80–89—and especially those over 90—strongly related with high risk of all-cause mortality. A clinical history of cancer, heart failure, and COPD also contributed to higher risk, in line with established medical knowledge.

The obesity paradox was observed once again, with lower BMI values associated with a higher risk. Similarly, lower HDL cholesterol levels were also linked to a higher risk, aligning with established medical knowledge. Medication use also contributed to model performance; for instance, insulin use and elevated HbA1c levels, indicating prolonged hyperglycemia over the preceding 2–3 months and consistent with the diagnostic criteria for diabetes, were both related to higher risk. These findings align well with existing clinical understanding.

Among the slope-based features, decreases in creatinine, glycemia, and HDL cholesterol appeared to increase risk. Although elevated creatinine is typically indicative of impaired renal function, a decrease in creatinine may instead reflect reduced muscle mass or underlying liver disease—both recognized markers of frailty. A decline in HDL cholesterol increasing risk is consistent with established knowledge, whereas a decrease in glycemia increasing risk is contrary to expectations. This counterintuitive finding may occur because substantial declines in glycemia often follow initially elevated baseline levels, suggesting an underlying diagnosis of diabetes.

The longitudinal (temporal) features also showed high relevance within the model. The time-related variables that are associated with binary indicators, enable the model to capture both the occurrence and timing of clinical events. Typically, lower time values indicate a positive event, while higher values correspond to its absence.

Creatinine time had a notable influence, with more recent measurements associated with a higher risk—likely reflecting its connection to current renal function status. Primary and secondary care visit times were also informative: individuals without recent primary care visits tended to have lower risk, possibly reflecting overall good health and reduced healthcare utilization, while recent visits were associated with increased mortality risk. Conversely, shorter intervals since a secondary care visit were associated with a higher risk, suggesting that more frequent specialist monitoring likely reflects greater patient frailty.

Medication- and test-related time variables also provided insights. Recent digoxin use was associated with increased risk, as expected given its prescription in patients with cardiac dysfunction. Similarly, recent BMI and glycemia measurements correlated with a higher risk, potentially reflecting closer monitoring of patients with existing health concerns. However, the relevance of Loop Diuretics time was less clear.

Overall, the interpretation of these temporal features is complex and not always directly aligned with medical knowledge. Many of these associations likely reflect patterns of healthcare utilization, where

sicker or frailer individuals undergo more frequent testing and medication changes. It is not clear if the models are creating relationships between the variable and its timing.

The dispensation of some medications appears to be associated with an elevated risk. However, these associations are unlikely to represent direct causal effects of the drugs themselves. Rather, they likely capture the higher baseline risk of patients with complex comorbidities for which these medications are prescribed. Accordingly, their predictive value reflects underlying disease burden rather than harmful effects of the treatments.

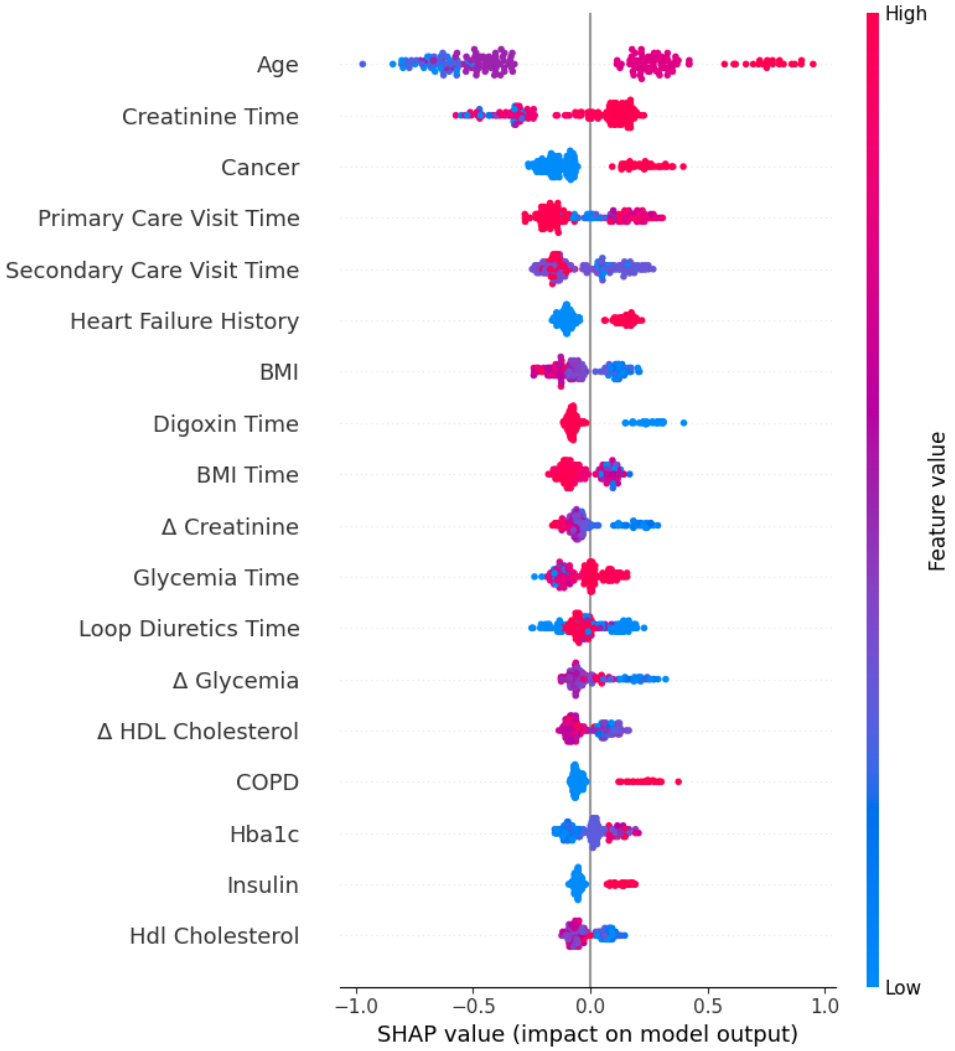

Figure 6.7: SHAP value of longitudinal XGBoost model on predicting all-cause death at 6 months

To identify the most influential predictors in a static model, static Logistic Regression was selected, and its feature importance is illustrated in Figure 6.8. The results closely mirror those of the slope-based Logistic Regression model for stroke/SE prediction. Similar features, including weight, HDL choles-

terol, systolic blood pressure (SBP), estimated glomerular filtration rate (eGFR), and total cholesterol, consistently emerged as important predictors. Weight reflected the obesity paradox, with higher values associated with lower risk. Low HDL cholesterol increased risk, as expected, while reduced eGFR indicated kidney dysfunction. Interestingly, lower total cholesterol was associated with higher risk, contrary to expectations, but potentially explained by statin use. Elevated SBP was linked to greater risk of death, consistent with established medical knowledge, whereas low diastolic blood pressure appeared to increase risk—a surprising finding, as it is not typically considered a strong risk factor in these populations. Age also contributed modestly, with higher values associated with increased risk, as expected.

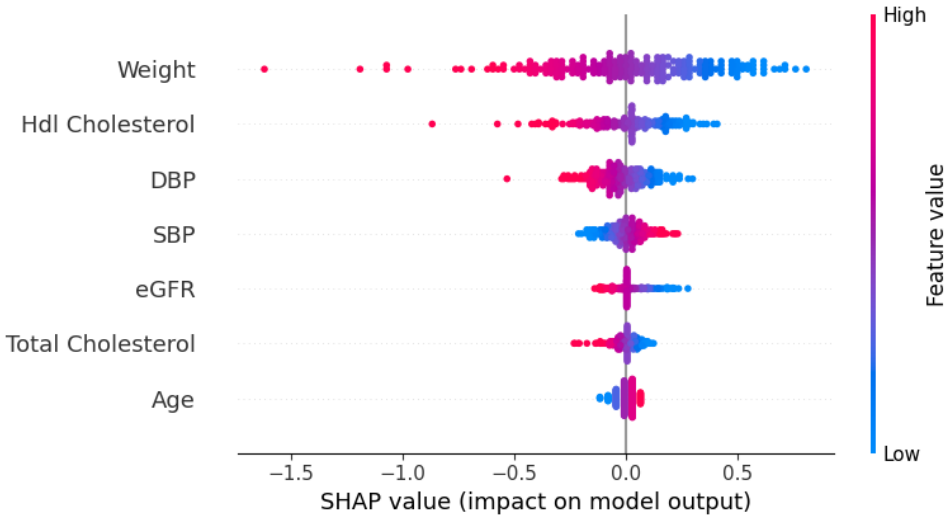

Figure 6.8: SHAP value of static Logistic Regression model on predicting all-cause death at 6 months

The longitudinal XGBoost model was then further evaluated across different time intervals. Table 6.5 reports the results: model performance generally improves as the time horizon increases, reflecting the larger number of positive cases; however, the highest AUC is observed at 3 months.

Table 6.5: Performance of longitudinal XGBoost model in predicting all-cause death at different time intervals, reported as mean ± standard deviation for accuracy, precision, sensitivity, F_1 score, F_2 score, AUC, and NNS.

Interval	Accuracy	Precision	Sensitivity	F_1 score	F_2 score	AUC	NNS
1 month	0.936 ± 0.010	0.039 ± 0.011	0.153 ± 0.060	0.061 ± 0.019	0.095 ± 0.031	0.722 ± 0.049	30.13 ± 15.00
3 months	0.828 ± 0.018	0.098 ± 0.014	0.523 ± 0.110	0.165 ± 0.025	0.279 ± 0.047	0.803 ± 0.030	10.43 ± 1.64
6 months	0.695 ± 0.021	0.111 ± 0.012	0.741 ± 0.051	0.192 ± 0.020	0.346 ± 0.033	0.779 ± 0.031	9.16 ± 1.04
1 year	0.659 ± 0.031	0.144 ± 0.014	0.770 ± 0.053	0.243 ± 0.022	0.412 ± 0.033	0.768 ± 0.038	7.00 ± 0.73
2 years	0.659 ± 0.013	0.197 ± 0.009	0.741 ± 0.038	0.312 ± 0.014	0.478 ± 0.022	0.753 ± 0.013	5.08 ± 0.24

6.2.3 Cardiovascular Death Outcome

Although GARFIELD-AF was not specifically designed to predict cardiovascular death, the all-cause death model achieved an AUC of 0.723 for cardiovascular death at 6 months in this cohort. This suggests

that GARFIELD-AF retains predictive ability for cardiovascular mortality in addition to its original focus on all-cause death.

The machine learning models predicting six-month cardiovascular mortality are summarized in Table 6.6. Overall, Logistic Regression and XGBoost demonstrated strong performance; however, the longitudinal Random Forest outperformed them in terms of precision, achieving the highest F_1, F_2, and NNS scores, while maintaining a competitive AUC. The Multi-Layer Perceptron also showed respectable performance, whereas Naive Bayes and Decision Trees performed comparatively worse. Incorporating slope-based features did not appear to enhance predictive ability, but using the full longitudinal data generally improved model performance across most algorithms.

Table 6.6: Performance of machine learning models in predicting cardiovascular death at 6 months, reported as mean ± standard deviation for accuracy, precision, sensitivity, F_1 score, F_2 score, AUC, and NNS.

Model	Accuracy	Precision	Sensitivity	F_1 score	F_2 score	AUC	NNS	Rank
Static NB	0.717 ± 0.040	0.070 ± 0.007	0.538 ± 0.078	0.123 ± 0.012	0.228 ± 0.021	0.715 ± 0.027	14.53 ± 1.50	14
Static LR	0.656 ± 0.024	0.080 ± 0.004	0.798 ± 0.071	0.145 ± 0.007	**0.285 ± 0.014**	0.774 ± 0.018	12.54 ± 0.53	2
Static DT	0.715 ± 0.037	0.068 ± 0.005	0.534 ± 0.073	0.121 ± 0.008	0.225 ± 0.017	0.651 ± 0.015	14.70 ± 1.03	15
Static RF	0.740 ± 0.095	0.078 ± 0.019	0.520 ± 0.138	0.133 ± 0.029	0.235 ± 0.039	0.707 ± 0.042	13.54 ± 2.98	10
Static XGB	0.798 ± 0.022	0.090 ± 0.011	0.498 ± 0.077	0.153 ± 0.018	0.261 ± 0.032	0.757 ± 0.025	11.19 ± 1.19	5
Static MLP	0.756 ± 0.006	0.079 ± 0.008	0.533 ± 0.060	0.138 ± 0.014	0.249 ± 0.026	0.712 ± 0.030	12.73 ± 1.20	9
Slope-based NB	0.387 ± 0.044	0.050 ± 0.004	**0.874 ± 0.030**	0.095 ± 0.008	0.204 ± 0.015	0.706 ± 0.032	20.02 ± 1.59	16
Slope-based LR	0.639 ± 0.020	0.078 ± 0.002	0.825 ± 0.057	0.143 ± 0.003	0.284 ± 0.008	**0.776 ± 0.016**	12.75 ± 0.27	3
Slope-based DT	0.717 ± 0.016	0.066 ± 0.010	0.507 ± 0.063	0.117 ± 0.018	0.217 ± 0.032	0.650 ± 0.046	15.51 ± 2.19	18
Slope-based RF	0.746 ± 0.090	0.078 ± 0.009	0.521 ± 0.145	0.133 ± 0.009	0.235 ± 0.012	0.684 ± 0.051	13.04 ± 1.54	13
Slope-based XGB	0.798 ± 0.015	0.089 ± 0.007	0.489 ± 0.076	0.150 ± 0.014	0.256 ± 0.028	0.753 ± 0.038	11.37 ± 0.92	7
Slope-based MLP	0.749 ± 0.018	0.076 ± 0.010	0.520 ± 0.058	0.132 ± 0.017	0.239 ± 0.030	0.701 ± 0.031	13.47 ± 1.94	12
Longitudinal NB	0.538 ± 0.030	0.062 ± 0.006	0.816 ± 0.063	0.115 ± 0.011	0.237 ± 0.022	0.722 ± 0.031	16.32 ± 1.43	11
Longitudinal LR	0.761 ± 0.034	0.087 ± 0.014	0.583 ± 0.130	0.151 ± 0.024	0.271 ± 0.046	0.727 ± 0.041	11.74 ± 1.69	6
Longitudinal DT	0.731 ± 0.062	0.069 ± 0.019	0.493 ± 0.142	0.120 ± 0.031	0.218 ± 0.054	0.641 ± 0.049	15.73 ± 4.32	17
Longitudinal RF	**0.825 ± 0.013**	**0.105 ± 0.010**	0.502 ± 0.059	**0.173 ± 0.016**	**0.285 ± 0.028**	0.764 ± 0.039	**9.61 ± 0.82**	1
Longitudinal XGB	0.796 ± 0.017	0.093 ± 0.009	0.516 ± 0.052	0.157 ± 0.015	0.269 ± 0.024	0.750 ± 0.024	10.93 ± 1.28	4
Longitudinal MLP	0.772 ± 0.014	0.087 ± 0.009	0.547 ± 0.063	0.149 ± 0.016	0.265 ± 0.028	0.734 ± 0.033	11.67 ± 1.13	8

Note: Undersampling was applied to Decision Tree, Random Forest, XGBoost, and Multi-Layer Perceptron models. Naive Bayes was used as a baseline, while Logistic Regression employed SMOTE for class balancing.

The best-performing predictor was the longitudinal Random Forest model, with feature importance depicted using SHAP values in Figure 6.9. Age again emerged as the most influential risk factor, consistent with established medical knowledge. Insulin use and insulin-related time variables suggested that taking insulin increases cardiovascular death risk, likely reflecting the underlying health status and comorbidities of patients requiring insulin therapy.

Most of the other important features in the model were temporal in nature. For example, recent measurements of eGFR or creatinine were associated with a higher risk. The implications of other time-related features, as well as features not displayed in the plot, are less clear. This underscores a key challenge: although incorporating temporal information improves model performance, interpreting these features in a clinically meaningful way remains difficult.

To examine key predictors more clearly, the static Logistic Regression model was selected, with feature importance illustrated in Figure 6.10. Once again, weight emerged as a highly influential feature,

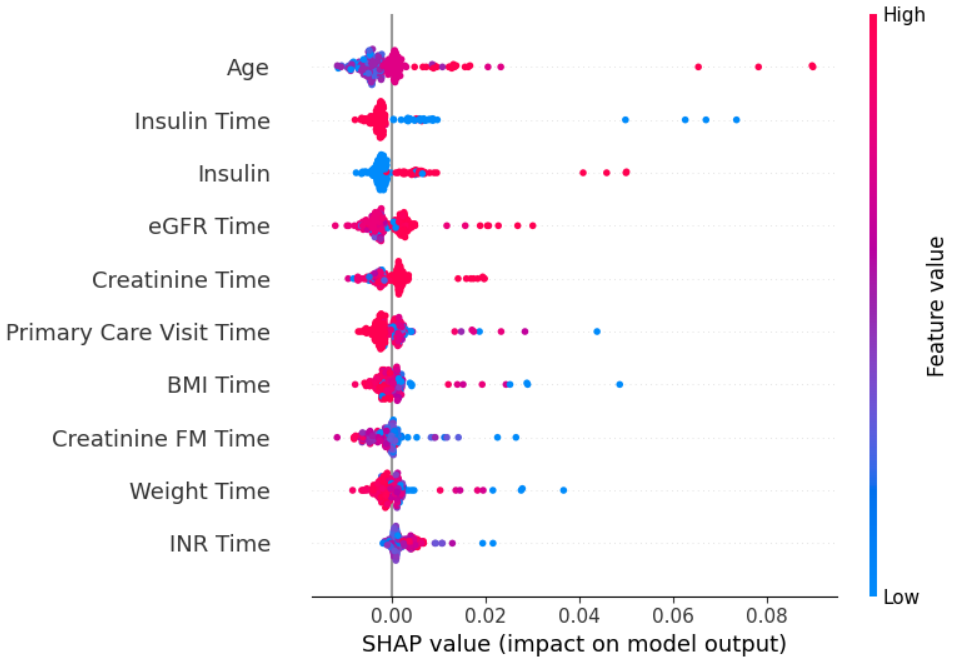

Figure 6.9: SHAP value of longitudinal Random Forest model on predicting cardiovascular death at 6-months

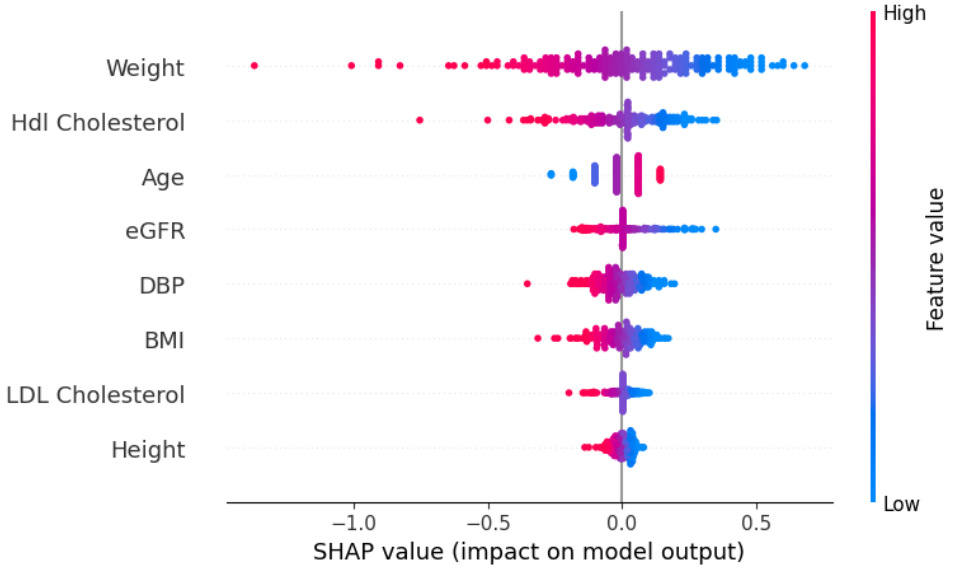

Figure 6.10: SHAP value of static Logistic Regression model on predicting cardiovascular death at 6-months

reflecting the obesity paradox, where lower weight and BMI were associated with increased cardiovascular death risk. Low HDL cholesterol consistently appears as a risk factor, aligning with established medical knowledge. Age, particularly values above 80 years, was associated with a higher risk, as expected. Reduced eGFR also contributed to risk, indicating kidney dysfunction. Low diastolic blood pressure (DBP) appeared as a risk factor once more, consistent with previous findings. Low LDL cholesterol

emerged as increasing risk, contrary to typical expectations, but potentially explained by statin usage. Finally, height appeared once again, inversely related to cardiovascular death risk.

The longitudinal Random Forest model was further evaluated across different time intervals. Table 6.7 reports the results: model performance generally improves as the time horizon increases, reflecting the larger number of positive cases; however, the highest AUC is observed at 6 months.

Table 6.7: Performance of longitudinal Random Forest model in predicting cardiovascular death at different time intervals, reported as mean ± standard deviation for accuracy, precision, sensitivity, F_1 score, F_2 score, AUC, and NNS.

Interval	Accuracy	Precision	Sensitivity	F_1 score	F_2 score	AUC	NNS
1 month	0.913 ± 0.013	0.027 ± 0.013	0.214 ± 0.101	0.048 ± 0.023	0.089 ± 0.043	0.563 ± 0.111	48.75 ± 25.00
3 months	0.823 ± 0.046	0.066 ± 0.025	0.417 ± 0.045	0.113 ± 0.037	0.197 ± 0.048	0.667 ± 0.053	17.14 ± 5.64
6 months	0.825 ± 0.013	0.105 ± 0.010	0.502 ± 0.059	0.173 ± 0.016	0.285 ± 0.028	0.764 ± 0.039	9.61 ± 0.82
1 year	0.610 ± 0.045	0.103 ± 0.010	0.754 ± 0.070	0.180 ± 0.016	0.331 ± 0.026	0.738 ± 0.034	9.84 ± 0.93
2 years	0.639 ± 0.014	0.153 ± 0.006	0.730 ± 0.061	0.253 ± 0.012	0.416 ± 0.024	0.743 ± 0.020	6.54 ± 0.27

6.2.4 Heart Failure Hospitalization Outcome

Table 6.8 summarizes the performance of the hyperparameterized learning models in predicting heart failure hospitalization. Models incorporating longitudinal predictors demonstrated the best overall performance, with improvements observed across all algorithms. Among them, longitudinal XGBoost achieved the strongest results, consistently exhibiting high precision and sensitivity, which translated into the highest F_1, F_2, and AUC scores. While the other models also performed well, longitudinal XGBoost consistently outperformed them across all metrics. Incorporating slope-based features did not appear to enhance model performance over the static predictors.

Table 6.8: Performance of machine learning models in predicting heart failure hospitalization at 6 months, reported as mean ± standard deviation for accuracy, precision, sensitivity, F_1 score, F_2 score, AUC, and NNS.

Model	Accuracy	Precision	Sensitivity	F_1 score	F_2 score	AUC	NNS	Rank
Static NB	0.712 ± 0.076	0.238 ± 0.028	0.564 ± 0.097	0.330 ± 0.020	0.434 ± 0.024	0.721 ± 0.023	4.27 ± 0.58	15
Static LR	0.682 ± 0.024	0.246 ± 0.018	0.756 ± 0.048	0.371 ± 0.024	0.534 ± 0.031	0.775 ± 0.018	4.08 ± 0.30	5
Static DT	0.676 ± 0.052	0.231 ± 0.026	0.675 ± 0.080	0.342 ± 0.028	0.483 ± 0.036	0.718 ± 0.022	4.38 ± 0.45	12
Static RF	0.647 ± 0.066	0.234 ± 0.031	0.780 ± 0.045	0.358 ± 0.033	0.528 ± 0.024	0.747 ± 0.041	4.35 ± 0.58	10
Static XGB	0.683 ± 0.032	0.248 ± 0.021	0.756 ± 0.018	0.372 ± 0.022	0.535 ± 0.018	0.766 ± 0.017	4.07 ± 0.33	4
Static MLP	0.666 ± 0.016	0.219 ± 0.010	0.658 ± 0.016	0.328 ± 0.011	0.469 ± 0.012	0.712 ± 0.015	4.58 ± 0.21	18
Slope-based NB	0.571 ± 0.029	0.193 ± 0.008	0.776 ± 0.039	0.309 ± 0.009	0.484 ± 0.013	0.714 ± 0.020	5.18 ± 0.21	16
Slope-based LR	0.683 ± 0.022	0.246 ± 0.017	0.749 ± 0.053	0.370 ± 0.023	0.531 ± 0.031	0.774 ± 0.020	4.09 ± 0.28	7
Slope-based DT	0.682 ± 0.039	0.234 ± 0.022	0.676 ± 0.071	0.346 ± 0.027	0.488 ± 0.036	0.713 ± 0.031	4.32 ± 0.41	13
Slope-based RF	0.659 ± 0.048	0.235 ± 0.021	0.763 ± 0.051	0.358 ± 0.023	0.524 ± 0.020	0.764 ± 0.015	4.29 ± 0.39	9
Slope-based XGB	0.668 ± 0.052	0.242 ± 0.023	0.765 ± 0.044	0.366 ± 0.024	0.531 ± 0.016	0.751 ± 0.032	4.18 ± 0.42	8
Slope-based MLP	0.648 ± 0.013	0.212 ± 0.010	0.676 ± 0.032	0.322 ± 0.014	0.470 ± 0.021	0.714 ± 0.010	4.74 ± 0.23	17
Longitudinal NB	0.606 ± 0.025	0.207 ± 0.010	0.769 ± 0.026	0.326 ± 0.012	0.498 ± 0.013	0.716 ± 0.017	4.84 ± 0.23	14
Longitudinal LR	0.654 ± 0.027	0.238 ± 0.015	**0.810 ± 0.038**	0.368 ± 0.019	0.547 ± 0.023	0.782 ± 0.014	4.22 ± 0.26	2
Longitudinal DT	0.699 ± 0.068	0.263 ± 0.033	0.745 ± 0.092	0.385 ± 0.025	0.538 ± 0.014	0.762 ± 0.018	3.86 ± 0.49	3
Longitudinal RF	0.654 ± 0.043	0.240 ± 0.029	0.808 ± 0.019	0.369 ± 0.034	0.546 ± 0.030	0.758 ± 0.035	4.22 ± 0.44	6
Longitudinal XGB	**0.713 ± 0.041**	**0.274 ± 0.025**	0.778 ± 0.026	**0.404 ± 0.027**	**0.567 ± 0.021**	**0.801 ± 0.020**	**3.69 ± 0.38**	**1**
Longitudinal MLP	0.670 ± 0.010	0.230 ± 0.011	0.710 ± 0.039	0.348 ± 0.017	0.501 ± 0.025	0.739 ± 0.021	4.35 ± 0.20	11

Note: Undersampling was applied to Decision Tree, Random Forest, XGBoost, and Multi-Layer Perceptron models. Naive Bayes was trained as baseline, while Logistic Regression used oversampling for class balancing.

Figure 6.12 illustrates the feature importance of the longitudinal XGBoost model, which achieved the best performance in predicting heart failure hospitalizations. Among the non-temporal variables, a

history of heart failure or valvular heart disease increased risk, consistent with medical knowledge. Loop diuretic use was associated with a higher risk, likely reflecting the underlying comorbidities of patients who require them. Low systolic blood pressure (SBP) was linked to increased risk, potentially indicating poor cardiac output, while elevated creatinine suggested kidney dysfunction. Older age also increased risk, as expected.

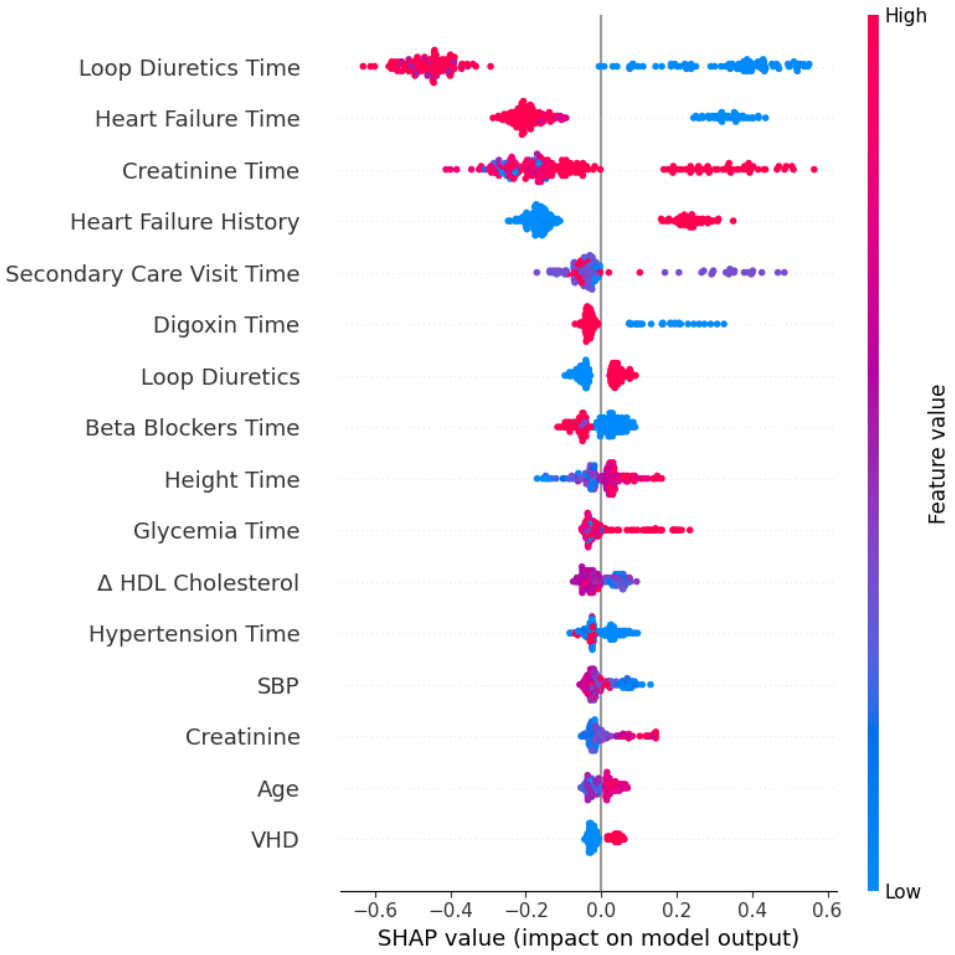

Figure 6.11: SHAP value of longitudinal XGBoost model on predicting heart failure hospitalization at 6 months

Among the slope-based features, a decrease in HDL cholesterol appeared to increase risk, in line with expectations.

Many of the other influential variables were temporal in nature. Time-related features such as loop diuretics, digoxin, and beta-blocker reflect their usage, which likely reflect underlying comorbidities. Secondary care visit time was also relevant, with patients having intermediate values showing a higher risk. Similarly, the timing of creatinine exams, height records, and glycemia measurements indicated that more recent measurements were associated with increased risk. An older hypertension diagnosis also

contributed to risk.

Once again, despite increasing model performance, these temporal variables are challenging to interpret individually, as their influence is likely intertwined with multiple other features, highlighting the complexity of translating temporal model insights into direct clinical understanding.

To analyze one static model, Figure 6.12 shows the feature importance of the static XGBoost, which achieved the best static performance in predicting heart failure hospitalizations. Several features were consistent with those identified by the longitudinal XGBoost model, reflecting agreement with the identified risk factors. These included variables related to heart failure history, age, and vascular heart disease, as well as medication use (loop diuretics, digoxin, or beta blockers), and clinical measures such as creatinine, blood glucose, HDL cholesterol, and systolic blood pressure.

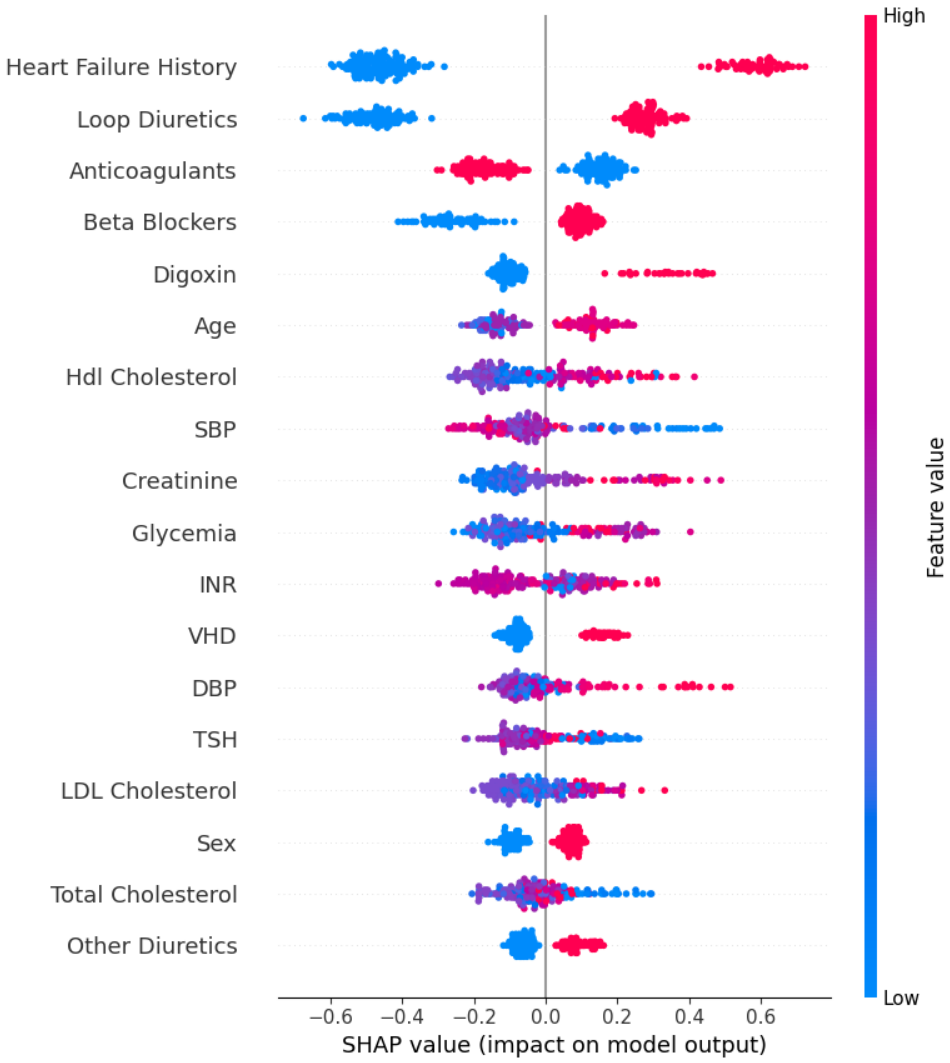

Figure 6.12: SHAP value of static XGBoost model on predicting heart failure hospitalization at 6 months

A prior history of heart failure was the strongest predictor, substantially increasing risk, while valvular

heart disease also contributed meaningfully—both well-established risk factors. Advanced age, particularly $\geq$80 years, and female sex were additional important predictors, consistent with clinical knowledge.

Medication use revealed nuanced associations. Treatment with beta-blockers, digoxin, or diuretics (loop or other) was associated with increased risk, likely reflecting the higher disease burden of patients prescribed these therapies. In contrast, anticoagulant therapy stood out: unlike other medications, the absence of anticoagulant use was linked to substantially higher hospitalization risk. This finding highlights anticoagulants as highly protective and clinically important, as patients receiving these therapies appeared better shielded from hospitalization.

Blood pressure also played a key role. Low systolic blood pressure was associated with increased risk, likely reflecting poor cardiac output. Although this may seem counterintuitive—given that hypertension is a major risk factor for developing heart failure—it can also be explained by the widespread use of antihypertensive medications within the population. High diastolic blood pressure also contributed to risk, though its cardiovascular implications are generally less pronounced in older populations. Other parameters, including TSH, HDL, and LDL cholesterol, influenced outcomes in a complex, multivariate manner, with both high and low values contributing to risk depending on context. These findings emphasize the value of multivariate modeling over simple univariate analyses when evaluating predictors of heart failure hospitalization.

Several other predictors appeared to influence outcomes in a complex, multivariate manner, with both high and low values contributing to risk depending on context. This underscores the importance of a multivariate modeling approach over simple univariate associations when evaluating predictors of heart failure hospitalization.

The model was further evaluated across different time intervals. Table 6.9 reports the results: as the time horizon increases, F_1, F_2, and precision increase, as sensitivity decreases; interestingly, the highest AUC is observed at 1 month.

Table 6.9: Performance of longitudinal XGBoost model in predicting heart failure hospitalizations at different time intervals, reported as mean ± standard deviation for accuracy, precision, sensitivity, F_1 score, F_2 score, AUC, and NNS.

Interval	Accuracy	Precision	Sensitivity	F_1 score	F_2 score	AUC	NNS
1 month	0.712 ± 0.010	0.135 ± 0.005	0.839 ± 0.032	0.232 ± 0.009	0.410 ± 0.014	0.839 ± 0.010	7.44 ± 0.28
3 months	0.700 ± 0.042	0.207 ± 0.021	0.759 ± 0.021	0.325 ± 0.025	0.494 ± 0.022	0.800 ± 0.012	4.88 ± 0.54
6 months	0.713 ± 0.041	0.274 ± 0.025	0.778 ± 0.026	0.404 ± 0.027	0.567 ± 0.021	0.801 ± 0.020	3.69 ± 0.38
1 year	0.702 ± 0.037	0.319 ± 0.024	0.735 ± 0.036	0.444 ± 0.021	0.582 ± 0.015	0.781 ± 0.021	3.15 ± 0.26
2 years	0.701 ± 0.028	0.388 ± 0.025	0.745 ± 0.030	0.510 ± 0.020	0.628 ± 0.017	0.774 ± 0.015	2.59 ± 0.17

6.2.5 Inpatient Visit Outcome

Table 6.10 summarizes the performance of the hyperparameterized learning models in predicting inpatient visit at 6 months. Logistic Regression, Random Forest, and XGBoost were the top-performing models, demonstrating both high precision and high sensitivity. Naive Bayes and Decision Tree mod-

els also performed well, with Naive Bayes showing higher sensitivity but lower precision, and Decision Trees showing higher precision but lower sensitivity. The MLP models underperformed relative to the other approaches. Overall, both static and slope-based predictors outperformed longitudinal predictors.

Table 6.10: Performance of machine learning models in predicting inpatient visit at 6 months, reported as mean ± standard deviation for accuracy, precision, sensitivity, F_1 score, F_2 score, AUC, and NNS.

Model	Accuracy	Precision	Sensitivity	F_1 score	F_2 score	AUC	NNS	Rank
Static NB	0.597 ± 0.120	0.300 ± 0.032	0.597 ± 0.165	0.389 ± 0.011	0.483 ± 0.055	0.652 ± 0.018	3.37 ± 0.41	13
Static LR	0.587 ± 0.021	0.294 ± 0.008	0.664 ± 0.052	0.407 ± 0.013	0.530 ± 0.027	0.663 ± 0.015	3.40 ± 0.10	11
Static DT	0.617 ± 0.035	0.291 ± 0.029	0.545 ± 0.038	0.379 ± 0.032	0.463 ± 0.034	0.611 ± 0.031	3.47 ± 0.35	16
Static RF	0.600 ± 0.029	0.301 ± 0.022	0.654 ± 0.023	0.412 ± 0.023	0.530 ± 0.022	0.666 ± 0.026	3.34 ± 0.23	10
Static XGB	0.620 ± 0.008	**0.311 ± 0.007**	0.641 ± 0.032	0.419 ± 0.012	0.528 ± 0.020	0.677 ± 0.015	3.22 ± 0.08	6
Static MLP	0.582 ± 0.006	0.276 ± 0.008	0.587 ± 0.029	0.375 ± 0.013	0.479 ± 0.020	0.616 ± 0.011	3.63 ± 0.11	15
Slope-based NB	0.320 ± 0.011	0.229 ± 0.003	**0.921 ± 0.021**	0.367 ± 0.005	0.574 ± 0.009	0.632 ± 0.021	4.37 ± 0.06	8
Slope-based LR	0.600 ± 0.017	0.303 ± 0.012	0.670 ± 0.036	0.417 ± 0.015	0.539 ± 0.022	0.668 ± 0.017	3.30 ± 0.13	7
Slope-based DT	0.619 ± 0.046	0.291 ± 0.031	0.532 ± 0.053	0.375 ± 0.033	0.455 ± 0.039	0.609 ± 0.029	3.48 ± 0.39	18
Slope-based RF	0.603 ± 0.015	0.308 ± 0.013	0.689 ± 0.024	**0.426 ± 0.017**	0.552 ± 0.020	**0.680 ± 0.017**	3.25 ± 0.13	**1**
Slope-based XGB	0.621 ± 0.017	0.314 ± 0.014	0.647 ± 0.022	0.423 ± 0.016	0.534 ± 0.018	0.674 ± 0.016	**3.19 ± 0.14**	5
Slope-based MLP	0.579 ± 0.010	0.272 ± 0.008	0.578 ± 0.019	0.370 ± 0.011	0.472 ± 0.014	0.614 ± 0.014	3.68 ± 0.11	17
Longitudinal NB	0.383 ± 0.053	0.240 ± 0.010	0.869 ± 0.052	0.376 ± 0.011	0.570 ± 0.015	0.638 ± 0.026	4.17 ± 0.18	9
Longitudinal LR	0.600 ± 0.010	0.307 ± 0.010	0.692 ± 0.040	0.425 ± 0.016	**0.553 ± 0.026**	0.678 ± 0.019	3.26 ± 0.11	3
Longitudinal DT	**0.631 ± 0.038**	0.314 ± 0.024	0.599 ± 0.064	0.410 ± 0.022	0.505 ± 0.035	0.649 ± 0.030	3.20 ± 0.24	12
Longitudinal RF	0.484 ± 0.060	0.272 ± 0.017	0.831 ± 0.045	0.409 ± 0.017	0.587 ± 0.012	0.619 ± 0.033	3.69 ± 0.25	4
Longitudinal XGB	0.539 ± 0.036	0.291 ± 0.014	0.793 ± 0.041	0.425 ± 0.011	0.588 ± 0.012	0.670 ± 0.031	3.45 ± 0.16	2
Longitudinal MLP	0.596 ± 0.010	0.284 ± 0.009	0.587 ± 0.014	0.383 ± 0.011	0.484 ± 0.013	0.630 ± 0.010	3.52 ± 0.11	14

Note: Undersampling was applied to Decision Tree, Random Forest, XGBoost, and Multi-Layer Perceptron models. Naive Bayes was trained with SMOTE, while Logistic Regression used oversampling for class balancing.

The best-performing model was the slope-based Random Forest, with SHAP feature importance shown in Figure 6.13. Both history of heart failure and valvular heart disease again emerged as risk factors, consistent with medical knowledge.

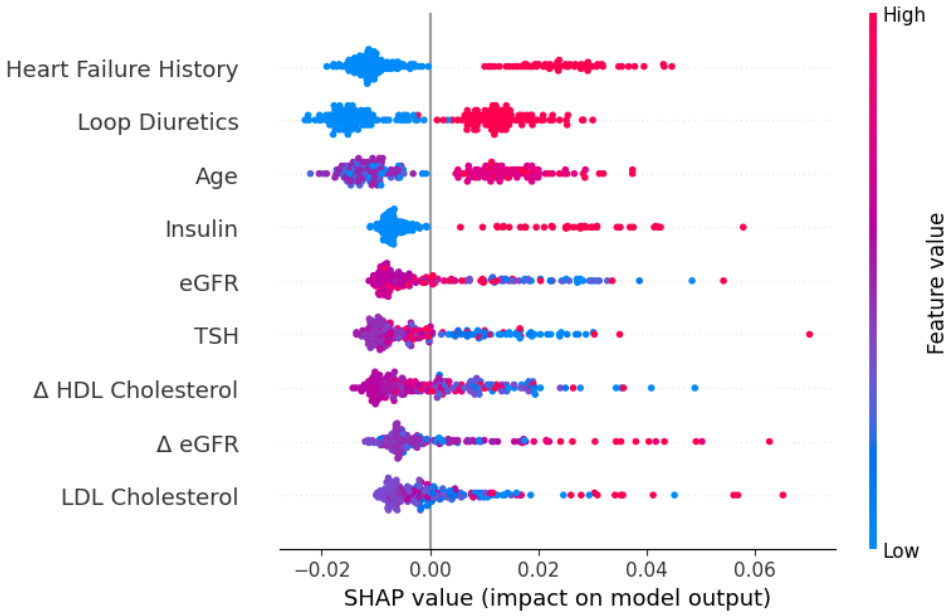

Figure 6.13: SHAP value of slope-based Random Forest model on inpatient visit at 6 months

Age was also among the top predictors, with patients aged 80 or older showing a higher inpatient visit risk. Use of loop diuretics or insulin was associated with increased risk, likely reflecting the underlying comorbidities of patients requiring these medications. Additionally, low eGFR was linked to higher risk, consistent with kidney dysfunction, and low TSH levels were associated with increased risk, potentially indicating hyperthyroidism. LDL cholesterol levels and the variability over time of HDL cholesterol and eGFR seem to have mixed relations with inpatient visit risk.

Other features not shown in the plot but still contributing to the model included UACR, creatinine, glycemia, and diastolic blood pressure, although their associations were more mixed.

The model was further evaluated across different time intervals. Table 6.11 reports the results: apart from sensitivity, metrics improve as the time horizon increases, reflecting the larger number of positive cases; however, the highest AUC is observed at 3 months.

Table 6.11: Performance of slope-based Random Forest model in predicting inpatient visit at different time intervals, reported as mean ± standard deviation for accuracy, precision, sensitivity, F_1 score, F_2 score, AUC, and NNS.

Interval	Accuracy	Precision	Sensitivity	F_1 score	F_2 score	AUC	NNS
1 month	0.596 ± 0.007	0.137 ± 0.007	0.664 ± 0.050	0.227 ± 0.013	0.375 ± 0.024	0.673 ± 0.021	7.32 ± 0.41
3 months	0.602 ± 0.011	0.239 ± 0.010	0.683 ± 0.035	0.354 ± 0.015	0.498 ± 0.022	0.683 ± 0.020	4.18 ± 0.17
6 months	0.603 ± 0.015	0.308 ± 0.013	0.689 ± 0.024	0.426 ± 0.017	0.552 ± 0.020	0.680 ± 0.017	3.25 ± 0.13
1 year	0.596 ± 0.024	0.392 ± 0.023	0.687 ± 0.043	0.499 ± 0.030	0.597 ± 0.037	0.666 ± 0.038	2.56 ± 0.16
2 years	0.604 ± 0.009	0.493 ± 0.009	0.678 ± 0.032	0.571 ± 0.017	0.631 ± 0.025	0.665 ± 0.013	2.03 ± 0.04

6.2.6 Acute Coronary Syndrome Outcome

Table 6.12 presents the performance of the hyperparameterized learning models in predicting acute coronary syndrome (ACS). Several models achieved reasonable performance, with longitudinal XGBoost showing particularly strong results.

Table 6.12: Performance of machine learning models in predicting acute coronary syndrome at 6 months, reported as mean ± standard deviation for accuracy, precision, sensitivity, F_1 score, F_2 score, AUC, and NNS.

Model	Accuracy	Precision	Sensitivity	F_1 score	F_2 score	AUC	NNS	Rank
Static NB	0.845 ± 0.016	0.036 ± 0.008	0.473 ± 0.143	0.066 ± 0.015	0.136 ± 0.032	**0.737 ± 0.038**	29.60 ± 6.93	3
Static LR	0.812 ± 0.027	0.032 ± 0.008	0.503 ± 0.104	0.060 ± 0.014	0.127 ± 0.029	0.734 ± 0.055	33.06 ± 8.08	6
Static DT	0.913 ± 0.030	0.030 ± 0.009	0.180 ± 0.052	0.050 ± 0.013	0.085 ± 0.016	0.609 ± 0.077	37.25 ± 11.51	14
Static RF	0.916 ± 0.020	0.028 ± 0.021	0.178 ± 0.119	0.048 ± 0.035	0.084 ± 0.059	0.604 ± 0.051	—	15
Static XGB	0.957 ± 0.014	0.058 ± 0.037	0.139 ± 0.090	0.077 ± 0.043	0.102 ± 0.055	0.703 ± 0.077	—	9
Static MLP	0.945 ± 0.007	0.057 ± 0.020	0.236 ± 0.106	0.091 ± 0.033	**0.143 ± 0.054**	0.678 ± 0.033	20.33 ± 7.69	2
Slope Based NB	0.717 ± 0.046	0.023 ± 0.004	0.572 ± 0.170	0.045 ± 0.008	0.100 ± 0.020	0.712 ± 0.067	44.49 ± 9.02	12
Slope Based LR	0.809 ± 0.024	0.032 ± 0.008	0.517 ± 0.125	0.060 ± 0.014	0.128 ± 0.030	0.729 ± 0.060	33.11 ± 8.23	8
Slope Based DT	0.888 ± 0.015	0.023 ± 0.009	0.208 ± 0.085	0.042 ± 0.016	0.080 ± 0.031	0.544 ± 0.042	48.81 ± 17.24	18
Slope Based RF	0.907 ± 0.029	0.024 ± 0.010	0.166 ± 0.068	0.042 ± 0.017	0.075 ± 0.030	0.597 ± 0.047	48.32 ± 17.29	17
Slope Based XGB	0.961 ± 0.005	0.063 ± 0.023	0.167 ± 0.072	0.091 ± 0.033	0.124 ± 0.048	0.691 ± 0.066	20.02 ± 12.19	7
Slope Based MLP	0.943 ± 0.003	0.045 ± 0.025	0.193 ± 0.111	0.073 ± 0.040	0.117 ± 0.064	0.661 ± 0.047	33.66 ± 21.99	11
Longitudinal NB	0.764 ± 0.023	0.028 ± 0.006	**0.587 ± 0.184**	0.054 ± 0.012	0.119 ± 0.028	0.720 ± 0.078	36.97 ± 8.13	10
Longitudinal LR	0.759 ± 0.025	0.029 ± 0.009	0.586 ± 0.140	0.055 ± 0.017	0.121 ± 0.035	**0.737 ± 0.059**	37.53 ± 10.20	4
Longitudinal DT	0.903 ± 0.017	0.030 ± 0.012	0.225 ± 0.108	0.053 ± 0.022	0.097 ± 0.041	0.604 ± 0.099	40.40 ± 17.41	13
Longitudinal RF	0.904 ± 0.013	0.025 ± 0.011	0.196 ± 0.107	0.044 ± 0.020	0.082 ± 0.040	0.580 ± 0.088	48.07 ± 18.99	16
Longitudinal XGB	**0.964 ± 0.006**	**0.073 ± 0.015**	0.181 ± 0.073	**0.103 ± 0.028**	0.138 ± 0.047	0.715 ± 0.066	**14.20 ± 2.89**	**1**
Longitudinal MLP	0.951 ± 0.008	0.058 ± 0.021	0.223 ± 0.115	0.092 ± 0.035	0.141 ± 0.061	0.698 ± 0.060	19.73 ± 7.95	5

Note: Undersampling was applied to Logistic Regression, Decision Tree, Random Forest, XGBoost, and Multi-Layer Perceptron models. Naive Bayes was used as a baseline.

It attained the highest precision, leading to the highest F_1 score and the lowest NNS, while maintaining a competitive AUC; however, its sensitivity was lower compared to other models. In contrast, Naive Bayes, Logistic Regression, and MLP achieved higher sensitivities but at the cost of lower precision. Random Forest and Decision Tree models exhibited relatively poor performance. Incorporating slope-based features did not improve predictive ability, whereas using the full set of longitudinal features generally enhanced model performance.

The best-performing model was the longitudinal XGBoost, with SHAP feature importance illustrated in Figure 6.14. Among the static features, high triglyceride levels were associated with an increased risk of acute coronary syndrome (ACS). Once again, lower body weight was linked to higher risk; however, contrary to both other models and established medical knowledge, greater height also appeared to increase risk. Low HDL cholesterol was associated with increased risk, consistent with known clinical evidence. Interestingly, low UACR values were also linked to higher risk—a contradictory finding, as lower UACR typically indicates normal kidney function. The presence of myocardial infarction or unstable angina strongly increased the predicted risk of ACS, as expected.

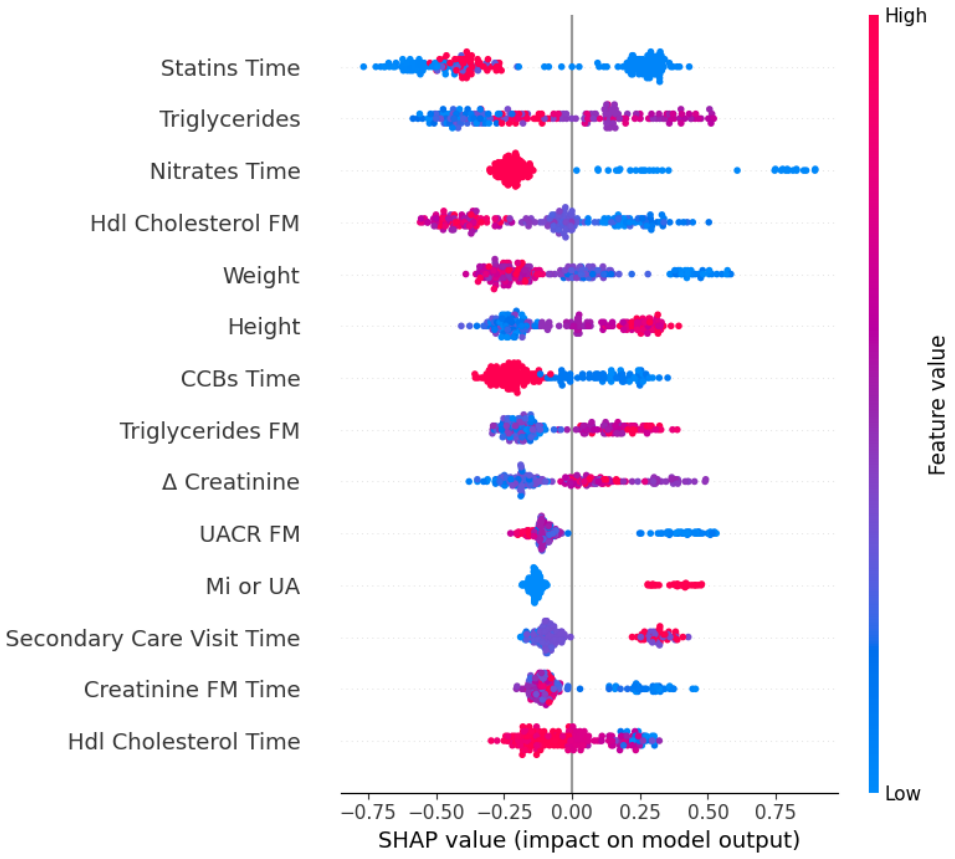

Figure 6.14: SHAP value of longitudinal XGBoost model on acute coronary syndrome at 6 months

Among the slope-based features, rising creatinine levels were associated with a higher risk, aligning with the understanding that elevated creatinine reflects impaired kidney function.

Regarding the time-dependent variables, statin, nitrate, and calcium channel blocker use were associated with increased predicted risk, likely reflecting the higher baseline risk of patients prescribed these medications. Additionally, a longer interval since the last secondary care visit was associated with higher risk, possibly indicating greater frailty or reduced medical follow-up. Shorter intervals since creatinine or HDL measurements were also linked to higher predicted risk, suggesting that patients undergoing recent testing may represent those under closer clinical scrutiny.

Overall, the relationships captured by the time-dependent variables remain difficult to interpret clinically, as their associations may reflect complex patterns within the dataset rather than direct causal mechanisms.

To further analyze a static model, Figure 6.15 presents the SHAP feature importance for the static Logistic Regression model. This model was selected due to the mixed interpretability of the MLP SHAP plot and because it shared the same key features as the Naive Bayes model—the third-best overall performer—while offering clearer visualization.

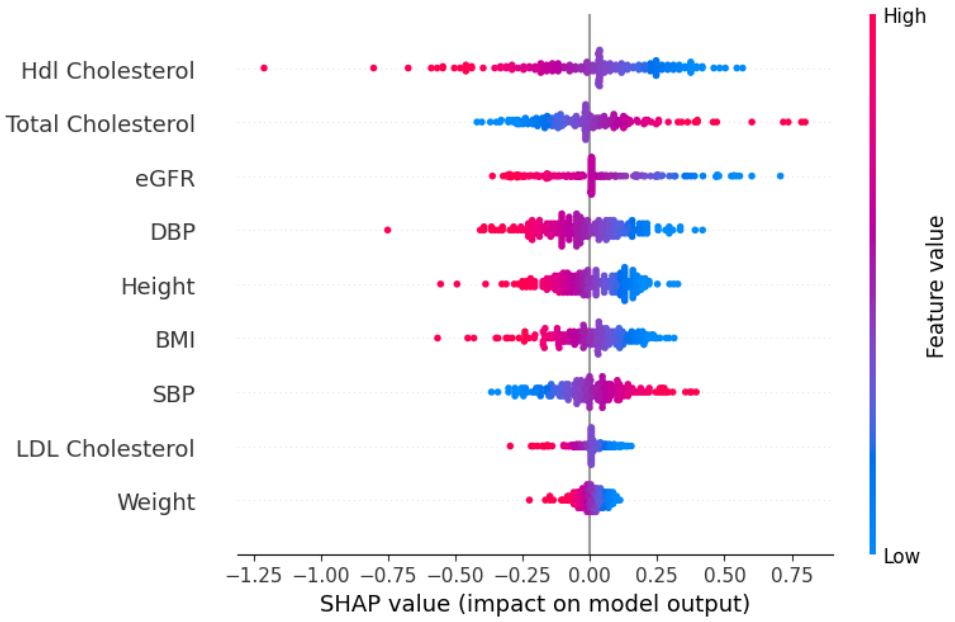

Figure 6.15: SHAP value of static Logistic Regression model on acute coronary syndrome at 6 months

The Logistic Regression model identified similar risk factors to those observed previously. Low HDL cholesterol and high total cholesterol were associated with increased risk, consistent with established medical knowledge. Low LDL cholesterol appeared to increase risk, likely reflecting the confounding effect of statin use. Reduced eGFR was linked to a higher risk, aligning with known associations between

kidney dysfunction and cardiovascular events. High systolic blood pressure increased risk, as expected, while low systolic blood pressure also appeared as a risk factor, consistent with patterns observed in other outcomes.

Interestingly, low height re-emerged as a risk factor—opposite to what was observed in the longitudinal XGBoost model—potentially indicating an interaction between height and other covariates. Finally, low BMI and body weight were again associated with higher risk, supporting the "obesity paradox" described in prior literature.

The model longitudinal XGBoost was further evaluated across different time intervals. Table 6.5 reports the results: model performance generally improves until 1 1-year time interval, which also has the highest AUC.

Table 6.13: Performance of static Logistic Regression model in predicting acute coronary syndrome at different time intervals, reported as mean ± standard deviation for accuracy, precision, sensitivity, F_1 score, F_2 score, AUC, and NNS.

Interval	Accuracy	Precision	Sensitivity	F_1 score	F_2 score	AUC	NNS
1 month	0.993 ± 0.004	0.067 ± 0.133	0.033 ± 0.067	0.044 ± 0.089	0.037 ± 0.074	0.592 ± 0.089	—
3 months	0.986 ± 0.004	0.013 ± 0.027	0.020 ± 0.040	0.016 ± 0.032	0.018 ± 0.036	0.648 ± 0.073	—
6 months	0.964 ± 0.006	0.073 ± 0.015	0.181 ± 0.073	0.103 ± 0.028	0.138 ± 0.047	0.715 ± 0.066	14.20
1 year	0.942 ± 0.014	0.079 ± 0.040	0.237 ± 0.119	0.117 ± 0.057	0.166 ± 0.080	0.718 ± 0.017	17.89
2 years	0.886 ± 0.008	0.070 ± 0.011	0.292 ± 0.069	0.113 ± 0.020	0.179 ± 0.035	0.682 ± 0.031	14.66

7

Clinical Decision Support Tool

Contents

The implemented clinical decision support tool is structured around two main components: a backend API and a user-friendly graphical user interface (GUI). This section provides an overview of the design and functionality of each developed component. The source code is available at https://github.com/rique-git/Medical-Tool-Interface-AF.

7.1 Application Programming Interface (API)

The API is developed to facilitate integration into existing clinical systems and workflows. It provides a standardized way to interact with the predictive models and the web interface, ensuring that functionalities can be extended and adapted as needed.

The API is implemented using `FastAPI`, a modern and efficient Python web framework. It exposes endpoints such as `/predict`, which accept patient data in JSON format, process it through the trained machine learning model, and return a prediction along with the associated probability.

This design separates the user interface from the computational backend, allowing different client applications (e.g., the `Dash` web application, hospital information systems, or third-party services) to access the predictive functionality in a consistent and secure manner. Responses are returned in JSON, ensuring interoperability with a wide range of technologies.

Running on an ASGI server such as `uvicorn`, the API can be deployed as a standalone service or containerized for scalability. This modular architecture improves maintainability and makes it possible to update or replace models without altering the user interface or external integrations.

7.2 Graphical User Interface (GUI)

The graphical user interface (GUI) provides a user-friendly front end for interacting with the medical tool. Developed using `Dash`, the GUI allows clinicians and other users to input patient data through intuitive forms, and receive model predictions along with informative visualizations. Currently, the web application supports predictions for cardiovascular death, but future development will extend the tool to additional outcomes, including atrial fibrillation (AF) itself.

The GUI includes input forms where users can enter relevant patient variables, such as age, weight, blood pressure, and lab results. A future additional feature will allow users to remove or add input fields as needed, enabling predictions even when some variables are missing.

Results are displayed through clear, interactive plots, highlighting the patient's data in the context of the training dataset. This visual feedback helps users interpret predictions at a glance. When a user submits patient data, the GUI sends a request to the backend API, retrieves the prediction and probability, and presents the results in real time.

The GUI is designed modularly: by separating it from the backend API, updates to the interface, visualizations, or layout can be made without affecting the underlying model or API logic.

This prototype design ensures that the tool is both accessible and informative, allowing clinical staff to efficiently leverage predictive insights while maintaining the flexibility to adapt and extend the interface in future iterations.

Figures 7.1 and 7.2 present the graphical user interface of the tool, showing how patient data is entered and processed through the API, with Figure 7.2 specifically illustrating the interactive prediction plots.

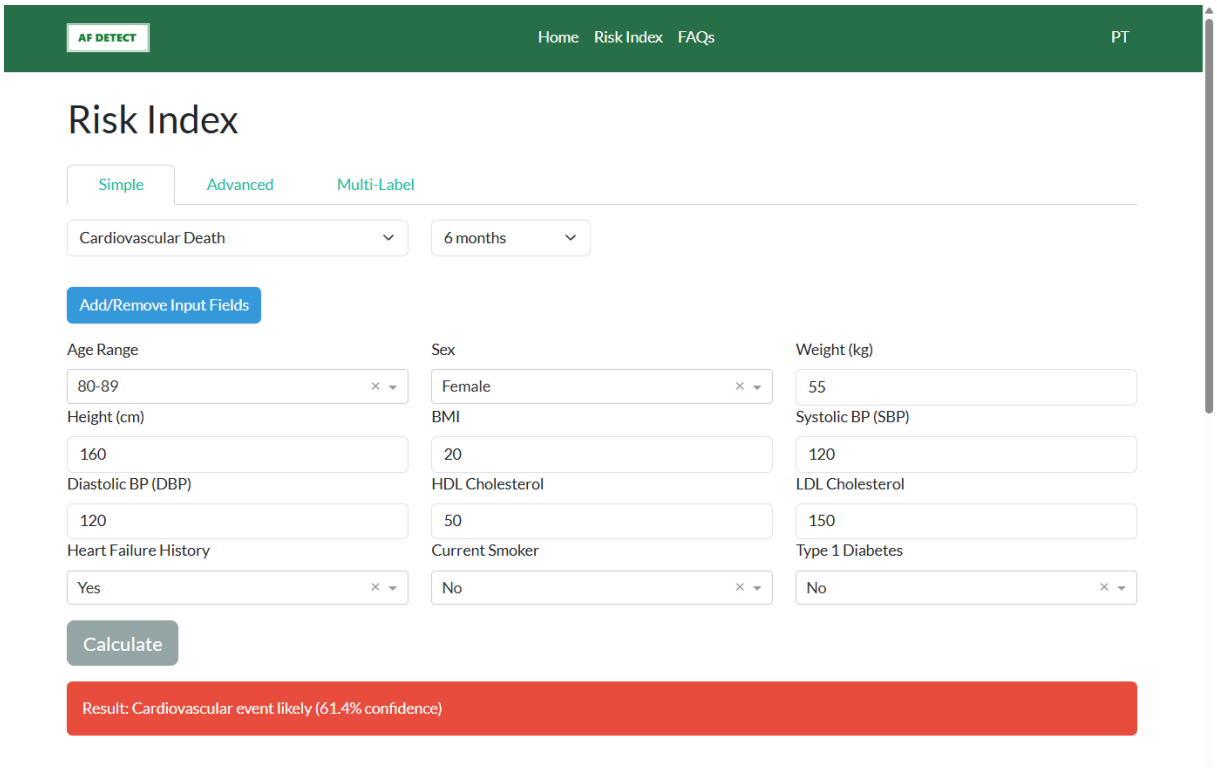

Figure 7.1: Prototype interface of the medical prediction tool predicting cardiovascular death at 6 months

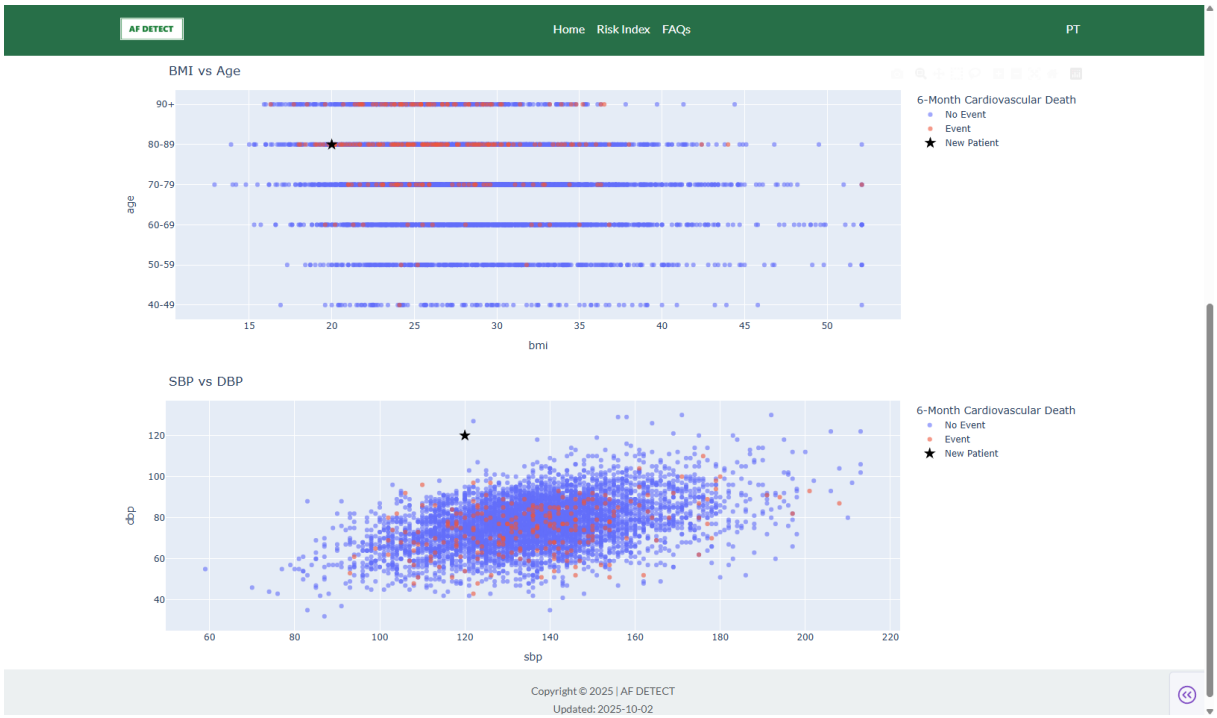

Figure 7.2: Interactive plots showing model predictions with training data and the new patient highlighted.

8

Conclusion

Contents

This concluding chapter summarizes the major results and insights of the conducted research, linking the models' findings with established medical knowledge while highlighting unexpected patterns influenced by comorbidities and medication use. The chapter also addresses system limitations and outlines directions for future research.

8.1 Concluding Remarks

This thesis addressed the challenges of prognosticating atrial fibrillation (AF)–related outcomes by pursuing two main objectives: developing predictive models of relevant clinical endpoints and designing a prototype clinical decision-support tool. In doing so, it sought to contribute to ongoing efforts to improve AF management in a real-world primary care setting.

The first objective centered on developing predictive models for key AF-associated endpoints. Machine learning algorithms consistently achieved strong performance, outperforming widely adopted

CHA_2DS_2-VASc for stroke/systemic embolism and the more advanced calculator GARFIELD-AF in stroke/systemic embolism and all-cause mortality, while also achieving robust results across other outcomes and prognostication time horizons.

Compared with previously developed machine-learning models, our ischemic stroke predictions showed lower discrimination, with AUCs of 0.60–0.65. Prior studies have typically reported AUCs above 0.70 [134], although one study achieved similarly modest performance and attributed this to the limited number of stroke events, an issue that also affects our dataset. Notably, that study also focused on anticoagulated patients, as does our dataset, which may contribute to the reduced ability to predict stroke events [132]. For all-cause mortality and cardiovascular mortality, our results align with existing literature, with AUCs around 0.78 [131, 132]. In contrast, the proposed models for predicting heart failure hospitalization, inpatient visits, and acute coronary syndrome in patients with atrial fibrillation appear to have no directly comparable studies. These outcomes therefore represent novel contributions to the field.

The developed models integrated multiple diagnostic and risk assessment dimensions, demographics, clinical examinations, comorbidities, behavioral risk factors, drug regimens, and the temporal characteristics of these events, thereby offering a more holistic approach to patient risk stratification. For all outcomes, incorporating temporal features improved model performance by capturing information such as the timing of diagnosis, medication prescriptions, clinical examination dates, and primary or secondary care visits.

Across the models, several recurring patterns emerged in the dataset. Height, although not a traditional risk factor, consistently appeared inversely correlated with risk across outcomes—most notably in stroke. This finding is thought to reflect early-life social and physical conditions that increase stroke and cardiovascular risk later in life.

The obesity paradox was another common theme: patients with lower weight and BMI tended to have a higher risk across outcomes. This phenomenon is likely driven by factors such as malnutrition, frailty, and the inability of BMI and weight to distinguish between fat and lean mass.

Lipid patterns also showed unexpected trends. Low LDL and low total cholesterol were often associated with increased risk, which can likely be explained by the widespread use of statins in the cohort. By lowering LDL and total cholesterol, statins may alter the way these biomarkers are interpreted as risk factors.

Systolic and diastolic blood pressure also gave interesting insights: a high SBP increased risk, as expected by clinical knowledge, but diastolic blood pressure, a less important risk factor, usually appeared with more importance by the model, with lower diastolic blood pressure increasing outcome risks. These findings likely reflect the heavy use of blood pressure–lowering medications in the population and their consequences. Moreover, hypertension itself—a well-known cardiovascular risk factor—rarely emerged

as important in the models, likely because 81% of patients carried a hypertension diagnosis and 91% were prescribed antihypertensive medications, which dampens its predictive signal.

Similarly, some established comorbidities, such as diabetes, did not consistently emerge as predictors, while insulin use did, again reflecting how treatment patterns shape the models' interpretation of risk. Other well-recognized comorbidities—including heart failure, cancer, COPD, and valvular heart disease—were more consistently highlighted, but not as much as medications like digoxin or loop diuretics.

Among clinical examinations, eGFR frequently appeared as a relevant predictor. Sex was associated with risk only in heart failure hospitalization, while age emerged as a predictor in several outcomes but not in the most unbalanced ones (stroke and acute coronary syndrome). This may suggest that additional data are needed, or that age-related risk was largely captured by age-related clinical variables such as kidney function and blood pressure.

Overall, these results highlight patterns that may seem unexpected but largely reflect the influence of extensive medication use within the cohort. Importantly, these are specific patterns observed in an AF population, but they likely mirror similar trends in other highly medicated cardiovascular populations with significant comorbidity burden. Examining how traditional risk factors behave in the presence of such treatments provides valuable insights and contributes to a better understanding of their interplay in shaping cardiovascular risk.

Despite the dominance of the temporal models, the interpretation of the temporal variables remains uncertain and warrants further investigation. Many of these variables appear to primarily reflect their underlying binary status—for instance, the "nitrates time" variable likely captures nitrate use rather than a true temporal pattern. Other time-related variables linked to clinical measurements are more difficult to interpret, as it is unclear whether they are interacting with the corresponding measurement. Moreover, assigning predictive importance to variables that reflect patterns of healthcare utilization may be conceptually problematic, as such patterns could represent biases in care delivery rather than true pathophysiological risk factors.

The second objective was accomplished through the design and development of a prototype clinical decision-support tool. This prototype integrated the machine learning models via a backend API and provided a user-oriented interface for data entry and visualization of predictions. While preliminary, the tool demonstrates the feasibility of translating complex predictive modeling into practical clinical support.

The contributions of this work lie in demonstrating the value of machine learning approaches and temporal stances for AF risk prediction, uncovering meaningful population-specific patterns, and building the foundation for a practical clinical tool. With further refinement and validation, these advances have the potential to enhance clinical decision-making, improve patient stratification, and ultimately support better outcomes for individuals affected by atrial fibrillation.

8.2 System Limitations and Future Work

While the present study provides valuable insights and demonstrates the feasibility of the proposed system, several limitations should be acknowledged, and opportunities for future work remain.

A major limitation is the restricted access to clinical records, as this was the first data iteration. A more comprehensive integration of Electronic Health Records (EHRs) would enable the use of richer datasets and allow for more accurate and generalizable models. Important variables were missing or incomplete. For example, the type of AF was not available; yet this distinction is clinically relevant not only for AF prediction itself but also for predicting its outcomes. Other variables absent from the dataset but included in GARFIELD-AF and classical risk calculators include pulse, dementia, anticoagulant treatment type (NOAC vs. VKA), significant cardiac murmur, and ethnicity. Moreover, AF patients are also at risk of bleeding events, but the dataset lacked detailed information on bleeding history, liver health, and alcohol consumption, which are important predictors.

Additionally, this work focused on phase 1 due to data limitations. To enable phase 2, future datasets should include both AF patients and a representative control group of individuals without AF, allowing for the development of models for AF risk prediction. Progressing to phase 3 would also require access to electrocardiograms (ECGs). Whether stored as raw signals or as pre-extracted features, ECGs could provide complementary predictive information—such as PR interval, left ventricular hypertrophy, and left atrial enlargement—thereby supporting the development of more robust models and enabling comparative analyses to better characterize AF-specific patterns.

Modeling limitations also exist. The present study primarily employed models compatible with the current longitudinal data structure but not specifically tailored to it. Future work could explore deep learning architectures designed for longitudinal data represented as variable–timestamp pairs. Such approaches are better suited to capture temporal dynamics and irregular sampling patterns, which may improve the accuracy of patient trajectory modeling.

Feature engineering could also be improved by defining clinically meaningful transformations or risk-aware thresholds for laboratory measures such as eGFR or creatinine, which may not only enhance model performance but also offer deeper insights into relevant risk factors.

Several methodological extensions are also possible. Multi-output learning did not perform optimally, likely due to underrepresented outcome combinations. Nevertheless, certain outcome pairs (e.g., cardiovascular death and all-cause death) could be revisited, as these may offer more balanced and clinically interpretable associations. Threshold tuning is another avenue of improvement: identifying optimal points on the ROC curve to maximize F1, F2, or other relevant metrics could enhance the clinical utility of the models.

Finally, the prototype medical tool interface requires further development. Iterative design with clinicians will be essential to refine its usability, expand available options, and ensure that the system inte-

grates seamlessly into clinical workflows.

Bibliography

[1] 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.

[2] G. Y. Lip, R. Nieuwlaat, R. Pisters, D. A. Lane, and H. J. Crijns, "Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation," *Chest*, vol. 137, no. 2, pp. 263–272, 2010.

[3] 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.

[4] B. P. Krijthe, A. Kunst, E. J. Benjamin, G. Y. Lip, O. H. Franco, A. Hofman, J. C. Witteman, B. H. Stricker, and J. Heeringa, "Projections on the number of individuals with atrial fibrillation in the european union, from 2000 to 2060," *European heart journal*, vol. 34, no. 35, pp. 2746–2751, 2013.

[5] M. Gouveia, J. Costa, J. Alarcão, M. Augusto, D. Caldeira, L. Pinheiro, A. V. Carneiro, and M. Borges, "Burden of disease and cost of illness of atrial fibrillation in portugal," *Revista Portuguesa de Cardiologia (English Edition)*, vol. 34, no. 1, pp. 1–11, 2015.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

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

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

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

[17] 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.

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

[19] 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.

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

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

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

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

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

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

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

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

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

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

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

[56] 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.

[57] 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.

[58] 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.

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

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

[61] 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.

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

[63] 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.

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

[65] 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.

[66] 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.

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

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

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

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

[71] 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.

[72] 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.

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

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

[75] 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.

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

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

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

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] 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.

[112] 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.

[113] M. Salvi, M. R. Acharya, S. Seoni, O. Faust, R.-S. Tan, P. D. Barua, S. García, F. Molinari, and U. R. Acharya, "Artificial intelligence for atrial fibrillation detection, prediction, and treatment: A systematic review of the last decade (2013–2023)," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 3, p. e1530, 2024.

[114] 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.

[115] 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.

[116] 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.

[117] S. Khurshid, U. Kartoun, J. M. Ashburner, L. Trinquart, A. Philippakis, A. V. Khera, P. T. Ellinor, K. Ng, and S. A. Lubitz, "Performance of atrial fibrillation risk prediction models in over 4 million individuals," *Circulation: Arrhythmia and Electrophysiology*, vol. 14, no. 1, p. e008997, 2021.

[118] 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.

[119] 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.

[120] 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.

[121] 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.

[122] 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.

[123] P. Gallego, V. Roldán, J. M. Torregrosa, J. Gálvez, M. Valdés, V. Vicente, F. Marín, and G. Y. Lip, "Relation of the has-bled bleeding risk score to major bleeding, cardiovascular events, and mortality in anticoagulated patients with atrial fibrillation," *Circulation: Arrhythmia and Electrophysiology*, vol. 5, no. 2, pp. 312–318, 2012.

[124] Z. Hijazi, J. Lindbäck, J. H. Alexander, M. Hanna, C. Held, E. M. Hylek, R. D. Lopes, J. Oldgren, A. Siegbahn, R. A. Stewart *et al.*, "The abc (age, biomarkers, clinical history) stroke risk score: a biomarker-based risk score for predicting stroke in atrial fibrillation," *European heart journal*, vol. 37, no. 20, pp. 1582–1590, 2016.

[125] Z. Hijazi, J. Oldgren, J. Lindbäck, J. H. Alexander, S. J. Connolly, J. W. Eikelboom, M. D. Ezekowitz, C. Held, E. M. Hylek, R. D. Lopes *et al.*, "The novel biomarker-based abc (age, biomarkers, clinical history)-bleeding risk score for patients with atrial fibrillation: a derivation and validation study," *The Lancet*, vol. 387, no. 10035, pp. 2302–2311, 2016.

[126] ——, "A biomarker-based risk score to predict death in patients with atrial fibrillation: the abc (age, biomarkers, clinical history) death risk score," *European heart journal*, vol. 39, no. 6, pp. 477–485, 2018.

[127] A. P. Benz, Z. Hijazi, J. Lindbäck, S. J. Connolly, J. W. Eikelboom, J. Oldgren, A. Siegbahn, and L. Wallentin, "Biomarker-based risk prediction with the abc-af scores in patients with atrial fibrillation not receiving oral anticoagulation," *Circulation*, vol. 143, no. 19, pp. 1863–1873, 2021.

[128] J.-P. Bassand, P. N. Apenteng, D. Atar, A. J. Camm, F. Cools, R. Corbalan, D. A. Fitzmaurice, K. A. Fox, S. Goto, S. Haas *et al.*, "Garfield-af: a worldwide prospective registry of patients with atrial fibrillation at risk of stroke," *Future cardiology*, vol. 17, no. 1, pp. 19–38, 2021.

[129] H. Nishi, N. Oishi, H. Ogawa, K. Natsue, K. Doi, O. Kawakami, T. Aoki, S. Fukuda, M. Akao, T. Tsukahara *et al.*, "Predicting cerebral infarction in patients with atrial fibrillation using machine learning: The fushimi af registry," *Journal of Cerebral Blood Flow & Metabolism*, vol. 42, no. 5, pp. 746–756, 2022.

[130] Y. Hamatani, H. Nishi, M. Iguchi, M. Esato, H. Tsuji, H. Wada, K. Hasegawa, H. Ogawa, M. Abe, S. Fukuda *et al.*, "Machine learning risk prediction for incident heart failure in patients with atrial fibrillation," *JACC: Asia*, vol. 2, no. 6, pp. 706–716, 2022.

[131] A. Bisson, Y. Lemrini, G. F. Romiti, M. Proietti, D. Angoulvant, S. Bentounes, W. El-Bouri, G. Y. Lip, and L. Fauchier, "Prediction of early death after atrial fibrillation diagnosis using a machine learning approach: a french nationwide cohort study," *American heart journal*, vol. 265, pp. 191–202, 2023.

[132] A. Bernardini, L. Bindini, E. Antonucci, M. Berteotti, B. Giusti, S. Testa, G. Palareti, D. Poli, P. Frasconi, and R. Marcucci, "Machine learning approach for prediction of outcomes in anticoagulated patients with atrial fibrillation," *International Journal of Cardiology*, vol. 407, p. 132088, 2024.

[133] Y. Chen, Y. Gue, P. Calvert, D. Gupta, G. McDowell, J. L. Azariah, N. Namboodiri, T. Bucci, A. Jabir, H. F. Tse *et al.*, "Predicting stroke in asian patients with atrial fibrillation using machine learning: a report from the kerala-af registry, with external validation in the aphrs-af registry," *Current problems in cardiology*, vol. 49, no. 4, p. 102456, 2024.

[134] B. Goh and S. M. Bhaskar, "Evaluating machine learning models for stroke prognosis and prediction in atrial fibrillation patients: a comprehensive meta-analysis," *Diagnostics*, vol. 14, no. 21, p. 2391, 2024.

[135] G. Hindricks, T. Potpara, N. Dagres, E. Arbelo, J. J. Bax, C. Blomström-Lundqvist, G. Boriani, M. Castella, G.-A. Dan, P. E. Dilaveris *et al.*, "2020 esc guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the european association for cardio-thoracic surgery (eacts) the task force for the diagnosis and management of atrial fibrillation of the european society of cardiology (esc) developed with the special contribution of the european heart rhythm association (ehra) of the esc," *European heart journal*, vol. 42, no. 5, pp. 373–498, 2021.

[136] I. Njølstad, E. Arnesen, and P. G. Lund-Larsen, "Body height, cardiovascular risk factors, and risk of stroke in middle-aged men and women: a 14-year follow-up of the finnmark study," *Circulation*, vol. 94, no. 11, pp. 2877–2882, 1996.

[137] S. Suzuki, "" cholesterol paradox" in atrial fibrillation," *Circulation Journal*, vol. 75, no. 12, pp. 2749–2750, 2011.

A

Extended Data Visualizations

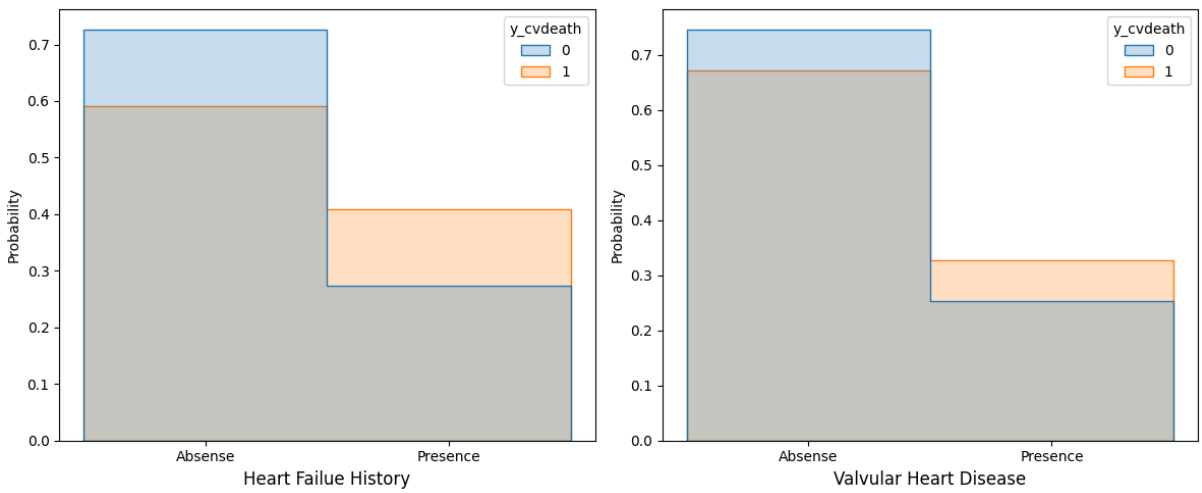

Figure A.1: Barplot of heart failure history and valvular heart disease stratified by cardiovascular death

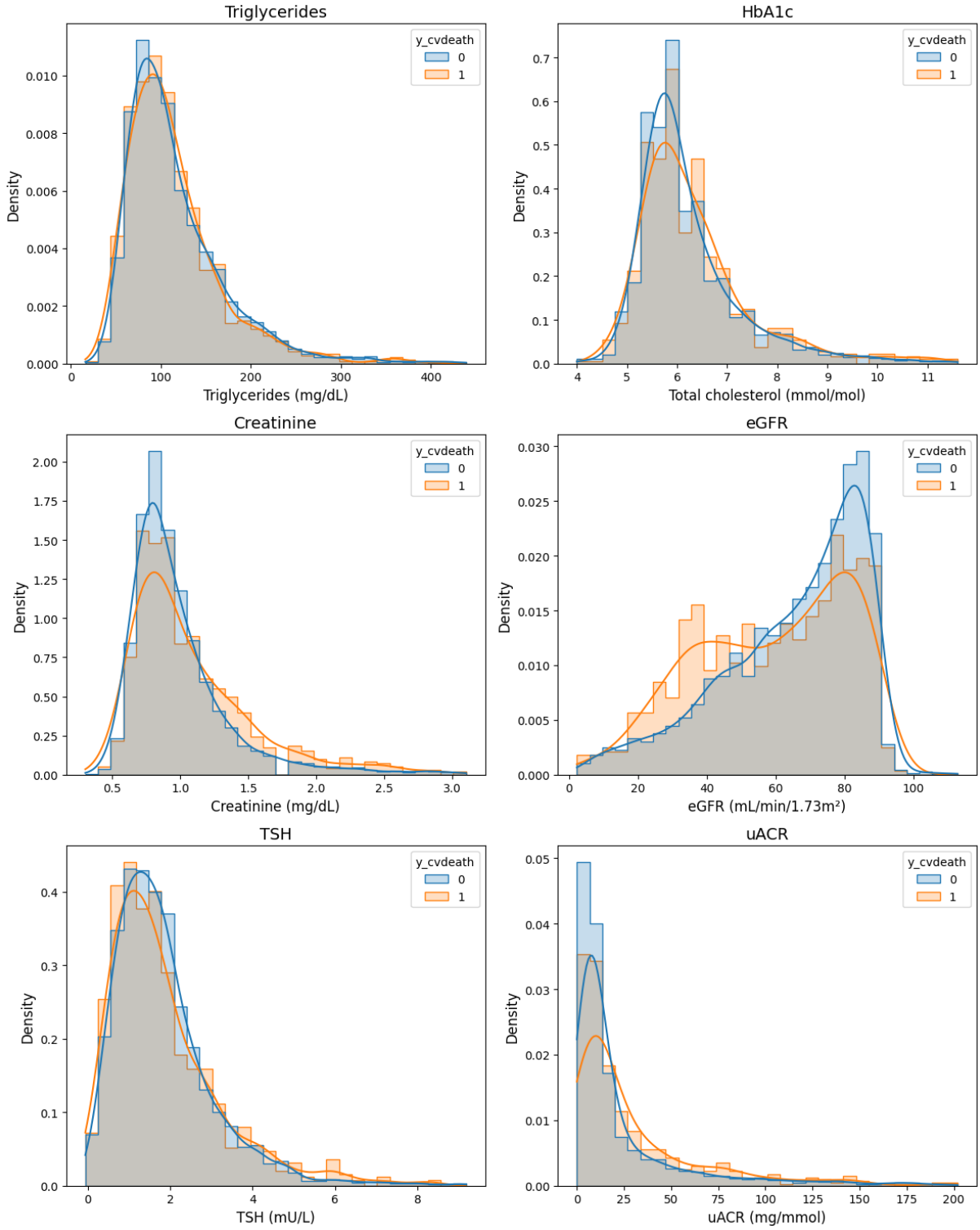

Figure A.2: Histograms of triglycerides, HbA1c, creatinine, eGFR, TSH, and uACR variables stratified by cardiovascular death

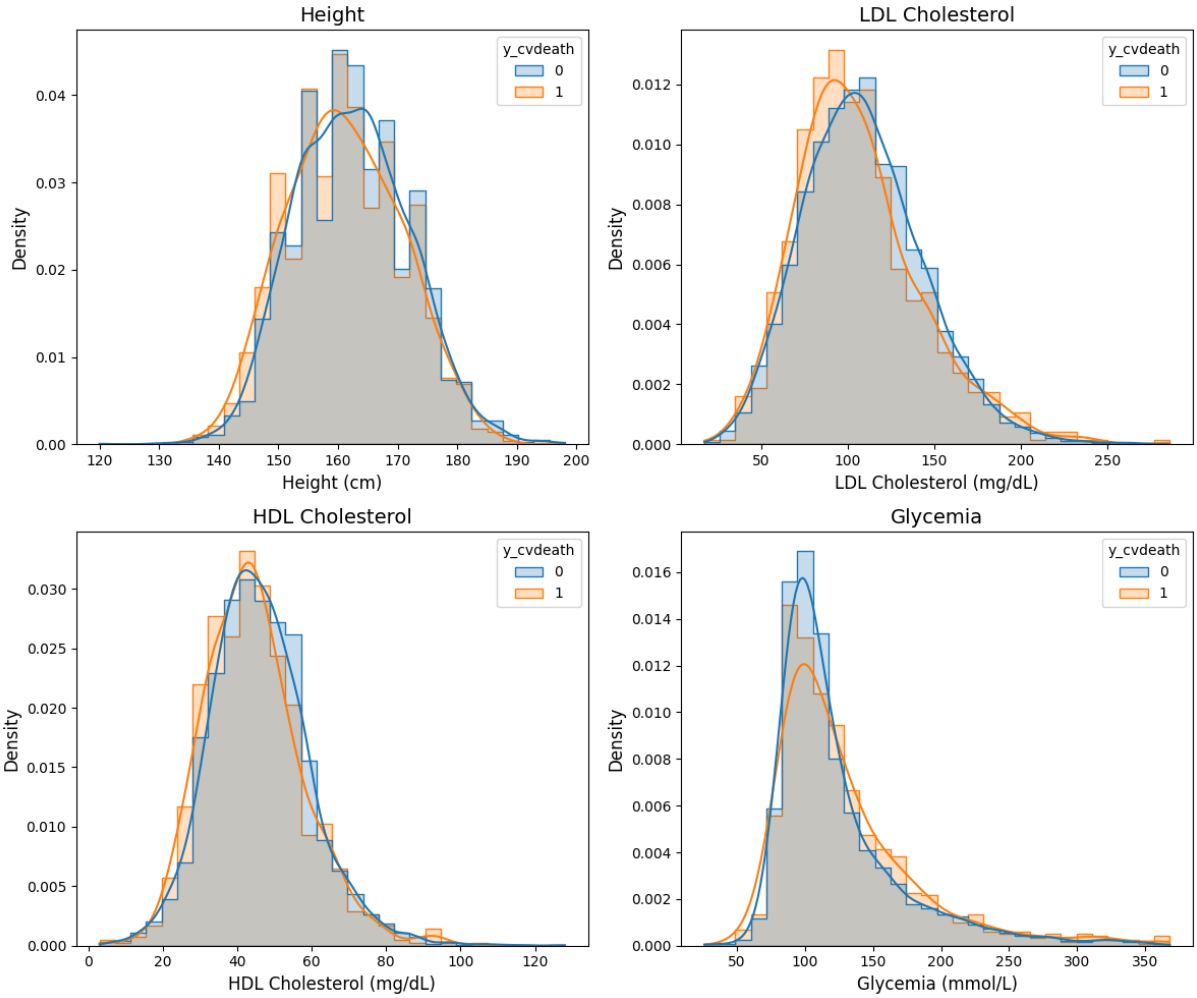

Figure A.3: Histograms of height, LDL and HDL cholesterol, and glycemia variables stratified by cardiovascular death