CATÓLICA
ESCOLA SUPERIOR DE BIOTECNOLOGIA

PORTO

CardioRiskAI: A Clinical Application for Rapid and Affordable Cardiovascular Risk Assessment using ECG Analysis and Machine Learning

by
Martim Gutierres Silva

September 2025

CATÓLICA
ESCOLA SUPERIOR DE BIOTECNOLOGIA

PORTO

CardioRiskAI: A Clinical Application for Rapid and Affordable Cardiovascular Risk Assessment using ECG Analysis and Machine Learning

Thesis presented to Escola Superior de Biotecnologia of the
Universidade Católica Portuguesa to fulfill the requirements of Master of Science degree in Biomedical Engineering

by
Martim Gutierres Silva

Supervisor: Professor Dr. Pedro Miguel de Luís Rodrigues

Co-supervisor: Researcher Dr. Vânia Guimarães

September 2025

Resumo

As doenças cardiovasculares (DCVs) constituem a principal causa de morte a nível global, embora a avaliação tradicional do risco seja frequentemente lenta e exigente em termos de recursos. Este trabalho investiga uma abordagem inovadora para uma avaliação rápida e acessível do risco cardiovascular, recorrendo a características de eletrocardiografia (ECG) cuidadosamente desenvolvidas e combinadas com um conjunto mínimo de dados clínicos (idade, índice de massa corporal, tabagismo e estado de diabetes). Ao aproveitar a crescente portabilidade e baixo custo da tecnologia de ECG, pretende-se disponibilizar uma alternativa economicamente viável aos métodos dependentes de análises laboratoriais, como o score de Framingham.

Foi desenvolvido um pipeline de processamento de ECG para extrair 102 características, utilizadas no treino de modelos de classificação e regressão com o score de Framingham como variável alvo. Os modelos demonstraram um desempenho excelente: a classificação binária entre doentes de baixo e alto risco alcançou uma precisão de 97,4% (AUC = 0,99); a comparação multiclasse entre grupos de baixo, moderado e alto risco atingiu 86.79% de precisão (AUC = 0.88); e a análise de regressão apresentou um erro quadrático médio entre 4 a 6 pontos percentuais entre todas as comparações. Para transpor estes resultados para a prática clínica, foi co-desenvolvida com cardiologistas uma aplicação clínica, integrando autenticação segura, gestão de doentes, inteligência artificial explicável e relatórios automatizados.

Os resultados indicam que um conjunto compacto de características derivadas do ECG, aliado a informação clínica básica, pode fornecer uma avaliação do risco cardiovascular rápida e economicamente eficiente. Esta abordagem tem o potencial de democratizar a estratificação de risco em contextos com recursos limitados e de acelerar a avaliação em situações de urgência. Caso seja validada a sua generalização, estes modelos poderão reduzir significativamente a dependência de análises laboratoriais extensas e dispendiosas na avaliação do risco cardiovascular.

Palavras-Chave: Aplicação Clínica, Doenças Cardiovasculares, Risco Cardiovascular, Suporte à Decisão Clínica, Eletrocardiograma, Score de Framingham, Aprendizagem Automática

Abstract

Cardiovascular diseases (CVDs) are the leading global cause of death, yet traditional risk assessment is often slow and resource-intensive. This work investigates a novel approach for rapid, accessible CVD risk assessment using engineered electrocardiography (ECG) features combined with minimal clinical data (age, BMI, smoking, and diabetes status). By leveraging increasingly portable and inexpensive ECG technology, we aim to provide a cost-effective alternative to lab-dependent methods like the Framingham score.

An ECG processing pipeline was developed to extract 102 features, which were used to train classification and regression models targeting the Framingham risk score. The models demonstrated excellent performance: binary classification distinguishing between low- and high-risk patients achieved 97.4% accuracy (AUC 0.99); the multiclass comparison among low-, moderate-, and high-risk groups reached 86.79% accuracy (AUC = 0.88); and regression analysis produced RMSE of between 4 to 6 percentage points. To translate these findings into practice, a clinical application was co-developed with cardiologists, integrating secure authentication, patient management, explainable AI, and automated reporting.

The results indicate that a compact set of ECG-derived features, coupled with basic clinical information, can provide rapid and cost-effective cardiovascular risk assessment. This approach holds the potential to democratize risk assessment in low-resource settings and accelerate urgent care evaluations. If validated for generalizability, these models could significantly reduce the reliance on extensive and costly blood laboratory tests for cardiovascular risk stratification.

Keywords: Cardiovascular Diseases, Cardiovascular Risk, Clinical Application, Clinical Decision Support, Electrocardiogram, Framingham Score, Machine Learning

Agradecimentos

Em primeiro lugar, gostaria de expressar a minha mais profunda gratidão ao Professor Pedro Rodrigues, que não só foi o meu supervisor nesta dissertação, como também foi a pessoa que despertou em mim o interesse por esta área de estudo através das suas aulas. Com a sua orientação tive o primeiro contato com o processo de investigação científica, a elaboração de um artigo e o subsequente processo de publicação. Foi com o Professor que comecei a minha carreira, na qual desenvolvi uma verdadeira paixão por investigação e *machine learning*. Foi muito mais do que um supervisor, esteve sempre disponível, atento ao trabalho e profundamente envolvido.

Agradeço igualmente à Fraunhofer AICOS Portugal pela excelente iniciativa do concurso '*Fraunhofer Portugal Challenge*', no qual participei em 2024 e tive a honra de vencer. Deste resultado nasceu esta colaboração, que me proporcionou a coorientação da Investigadora Vânia Guimarães, a quem agradeço, em conjunto com o Investigador Duarte Folgado.

Quero também reconhecer à Dra. Camila Leite da Universidade do Ceará, pela recolha dos dados utilizados nesta dissertação, bem como o Dr. Octávio Barbosa Neto, cardiologista, pela sua disponibilidade para as reuniões e pelo valioso feedback prestado ao longo do trabalho.

Num plano mais pessoal, agradeço aos meus pais, Goreti Pinho e Gomercindo Silva, pelo apoio e incentivo constantes, e sei o privilégio que é tê-los como pais. Agradeço igualmente aos meus amigos e à minha namorada pelo apoio essencial durante todo este percurso.

Este tema de dissertação tem ainda um significado especial para mim. Em ambos os lados da minha família, perdemos, demasiado cedo, familiares devido a eventos cardiovasculares. Este tipo de doenças não só constitui a principal causa de mortalidade a nível mundial, mas também é fonte de grande receio diário, pelas suas características súbitas e muitas vezes assintomáticas. Ao contrário de outras doenças que permitem um processo mais lento de aceitação, os eventos cardiovasculares trazem frequentemente choque, surpresa e sofrimento profundo às famílias que ficam, privadas para sempre de uma parte de si.

Este trabalho representa uma tentativa de acelerar, democratizar e aumentar a frequência da avaliação do risco cardiovascular. Foi com este objetivo que a presente dissertação procurou ir além da componente técnica de *machine learning* e tentou aproximar-se da aplicação clínica. Porque, no fim, esse deve ser o último propósito da investigação: transformar conhecimento em impacto real e positivo na vida das pessoas.

Martim Gutierres Silva

"Some people worry that artificial intelligence will make us feel inferior, but then, anybody in his right mind should have an inferiority complex every time he looks at a flower"

– Alan Kay

Outline

List of Figures

List of Tables

Abbreviations and Symbols

2FA	Two Factor Authentication
ABP	Arterial Blood Pressure
ACE	Angiotensin-Converting Enzyme inhibitors
AdaBoost	Adaptive Boost
AHA	American Heart Association
AI	Artificial Intelligence
ANN	Artificial Neural Network
APG	Acceleration Photoplethysmogram
ASCVD	Atherosclerotic Cardiovascular Disease
AUC	Area Under the Receiver Operating Characteristic Curve
AV	Atrioventricular
Bagg	Bagging
BMI	Body Mass Index
CAC	Coronary Artery Calcium
CANTOS	Canakinumab Anti-inflammatory Thrombosis Outcomes Study
CHD	Coronary Heart Disease
CKD	Chronic Kidney Disease
CSV	Comma-Separated Values
CVDs	Cardiovascular Diseases
C-index	Concordance index
DALYs	Disability-Adjusted Life Years
DAT	Data File
DC	Direct Current
DeTree	Decision Tree
DT	Decision Tree
DWT	Discrete Wavelet Transform
ECG	Electrocardiogram
ELM	Extreme Learning Machine
EMG	Electromyography

ExTree	Extra Trees
FHs	Framingham Heart Study
FN	False Negative
FP	False Positive
FRS	Framingham Risk Score
GauNB	Gaussian Naive Bayes
GBD	Global Burden of Disease
GPC	Gaussian Process Classifier
GradBoost	Gradient Boosting
GDPR	General Data Protection Regulation
Harrell's C	Concordance index
HCD	Human-Centered Design
HDL-C	High-Density Lipoprotein Cholesterol
HELV	Leonardo da Vinci State Hospital
HFpEF	Heart Failure with Preserved Ejection Fraction
HFrEF	Heart Failure with Reduced Ejection Fraction
HIPAA	Health Insurance Portability and Accountability Act
HR	Hazard Ratio
HUWC	Walter Cantídio University Hospital
IEEE	Institute of Electrical and Electronics Engineers
IHD	Ischemic Heart Disease
INFARMED	*Autoridade Nacional do Medicamento e Produtos de Saúde*
ISO	International Organization for Standardization
ISH	International Society of Hypertension
KNN	K-Nearest Neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
LDL-C	Low-Density Lipoprotein Cholesterol
LogReg	Logistic Regression
LR	Logistic Regression
LSVC	Linear Support Vector Classification
LVH	Left Ventricular Hypertrophy
MAT	MATLAB Data File
Max	Maximum
MDR	European Medical Device Regulation
Min	Minimum
ML	Machine Learning

MLP	Multilayer Perceptron
MRECCC	Minimal, Rapid, and Easily Communicated Clinical Context
mV	Millivolts
NA	Not Available
PDF	Portable Document Format
PROCAM	Prospective Cardiovascular Münster Study
QDA	Quadratic Discriminant Analysis
QRcode	Quick Response Code
RMS	Root Mean Square
RF	Random Forest
SA	Sinovial
SBP	Systolic Blood Pressure
SCORE	Systematic Coronary Risk Evaluation
SGLT2	Sodium–Glucose Cotransporter 2
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TC	Total Cholesterol
TN	True Negative
TOTP	Time-based One-Time Password
TP	True Positive
UK	United Kingdom
UML	Unified Modeling Language
WCAG	Web Content Accessibility Guidelines
WHO	World Health Organization
XAI	Explainable Artificial Intelligence
μV	Microvolts

CHAPTER 1

Introduction

The introduction presents the motivation for this work and situates it within the current clinical background. A systematic literature review, supported by a summary table, was conducted to identify research gaps in the field. These gaps guided the formulation of the research objectives and the overall structure of the dissertation.

1.1 Motivation

Cardiovascular diseases (CVDs) are an escalating global health crisis and the leading cause of mortality, morbidity, and disability worldwide. In 2021 alone, CVDs accounted for approximately 20.5 million deaths, representing nearly one third of all global fatalities [1]. This demonstrates a sharp increase of 69.4% from the 12.1 million deaths recorded in the year 1990 [1]. During the same period, the global population increased by only 49.1% [2], underscoring that the rise in cardiovascular related deaths has significantly outpaced population growth.

In 2022, approximately 775 million individuals worldwide were diagnosed with CVDs [3]. Ischemic heart disease and stroke bear the greatest burden. Collectively, these two conditions accounted for 85% of all cardiovascular related deaths globally [3]. Specifically, ischemic heart disease was responsible for approximately 9.2 million deaths, while stroke accounted for around 7.3 million deaths [3]. Additionally, the Global Burden of Disease (GBD) study highlights that CVDs had the highest global age-standardized disability-adjusted life years (DALYs) among all diseases, measuring at 5,078.4 per 100,000 population [3].

1.2 Background

To assess a patient's cardiovascular risk and guide critical clinical decisions such as medication prescriptions and surgical interventions, medical practitioners rely on cardiovascular risk scores [4]. These scores utilize weighted mathematical formulas incorporating clinical metrics usually including arterial blood pressure (ABP), systemic blood pressure (SBP), high-density lipoprotein (HDL-C), low-density lipoprotein (LDL-C), diabetes status, smoking status, age, and sex. One of the most widely utilized cardiovascular risk assessment tools globally is the Framingham Risk Score, originally developed by Wilson WF Peter *et al.* (1998) [5].

1.3 Systematic Literature Review

To build a solid foundation for the present study, it is essential to understand the current landscape of machine learning (ML) applications for cardiovascular risk prediction. To achieve this, the literature review followed established procedures for systematic reviews, ensuring that the process is rigorous, replicable, and capable of retrieving the same set of articles.

1.3.1 Research Questions

This literature review is guided by a series of key research questions, designed to capture the most relevant and impactful aspects of state-of-the-art approaches in this field:

RQ1. How has machine learning been used for cardiovascular risk prediction?

RQ2. What are the development trends in machine learning applied for cardiovascular risk prediction?

RQ3. How have recent studies utilized machine learning techniques to improve or innovate cardiovascular risk prediction beyond traditional risk scores?

RQ4. Which signal processing methods are most commonly used in preparing ECG data and what features are extracted for machine learning applications in cardiology?

RQ5. What are the current best performing machine learning models for cardiovascular risk prediction, and how is their performance evaluated?

RQ6. What are the limitations of prior studies on machine learning for cardiovascular risk prediction?

This set of research questions begins with RQ1, which explores the application of machine learning in cardiovascular risk prediction, examining whether researchers enhance existing clinical scores such as the Framingham Risk Score as ground truth labels or propose new predictive frameworks, and what types of features are commonly applied. RQ2 addresses whether there is a trend toward deep

learning models over classical machine learning approaches and wants to identify the features gaining prominence in these models to determine if a methodological shift is underway. RQ3 investigates the main innovations in recent studies, including technical contributions and efforts to improve the research in this field. RQ4 focuses on studies utilizing ECG signals, asking which signal processing techniques are used and what features are extracted. RQ5 evaluates what are the best performing machine learning models, feature selection methods, and normalization strategies, as well as how performance is assessed. Finally, RQ6 seeks to identify the key limitations found in current studies, helping to clarify the gaps that this dissertation aims to address through methodological and practical contributions.

1.3.2 Research Process and Criteria

To analyze current research in cardiovascular risk prediction using machine learning, this literature review follows the PRISMA methodology, which is widely regarded as the gold standard for conducting systematic reviews, meta-analyses, and literature reviews for its transparency and reproducibility.

The digital libraries selected for this review were chosen for their disciplinary strengths and broad coverage of relevant literature. Google Scholar was included for its broad indexing of scientific articles. Embase was selected for its focus on biomedical and life sciences literature, while IEEE Xplore was incorporated due to its strong representation of engineering and computer science publications. These databases are listed in 1.1.

Table 1.1: Databases used in the research process

Databases
Google Scholar
Embase
IEEE Xplore

To ensure the retrieval of relevant and comprehensive literature, a keyword strategy was developed to cover common variations in terminology related to ECG analysis, cardiovascular risk prediction, and machine learning. Boolean operators were used to construct a flexible and inclusive search string, represented in Table 1.2, to capture a broad yet focused range of studies.

The general search string structure followed the form $"A" \wedge "B" \vee "C"$, iterating through all possible combinations in each column and therefore ensuring coverage across the domains of machine learning (A) approaches, CVD risk (B) and clinical scoring systems (C).

Table 1.2: Strings used in the search process

A	B	C
Machine Learning	Cardiovascular Risk	Framingham Risk Score
Deep Learning	Heart Disease Risk	ASCVD Risk Estimator
Regression Model	Cardiac Risk	Qrisk
Statistical Model	Cardiovascular Event	SCORE
Predictive Model	Stroke Risk	PROCAM
ML		WHO/ISH

A rigorous set of inclusion and exclusion criteria was applied to ensure the quality and relevance of the selected studies. These criteria are outlined in Tables 1.3 and 1.4.

Table 1.3: Inclusion Criteria

Literature Review Inclusion Criteria
Peer-reviewed articles
Published in English
Full-text availability
Direct relevance to cardiovascular risk
Direct use of a type ML
Presents model evaluations

Table 1.4: Literature Review Exclusion Criteria

Exclusion Criteria
Duplicate articles
Editorials, workshop summaries, or non-peer-reviewed content
Titles or abstracts not aligned with selected keywords
Review papers, meta-analyses, or surveys
Exclusive focus on imaging modalities

These criteria facilitated the identification of primary research articles that directly address the proposed research questions, ensuring the literature review reflects the current state of the field and areas in need of further investigation.

1.3.3 PRISMA Method

Following the search, a total of 222 studies were identified using the string combinations in the databases mentioned above. Duplicate records were then removed, resulting in 161 studies. These articles were subsequently screened by title and abstract to determine whether they met the inclusion criteria outlined in Table 1.3 and did not fall under the exclusion criteria specified in Table 1.4. From this process, 31 studies were selected for full-text review, of which only 11 provided relevant insights for the present work.

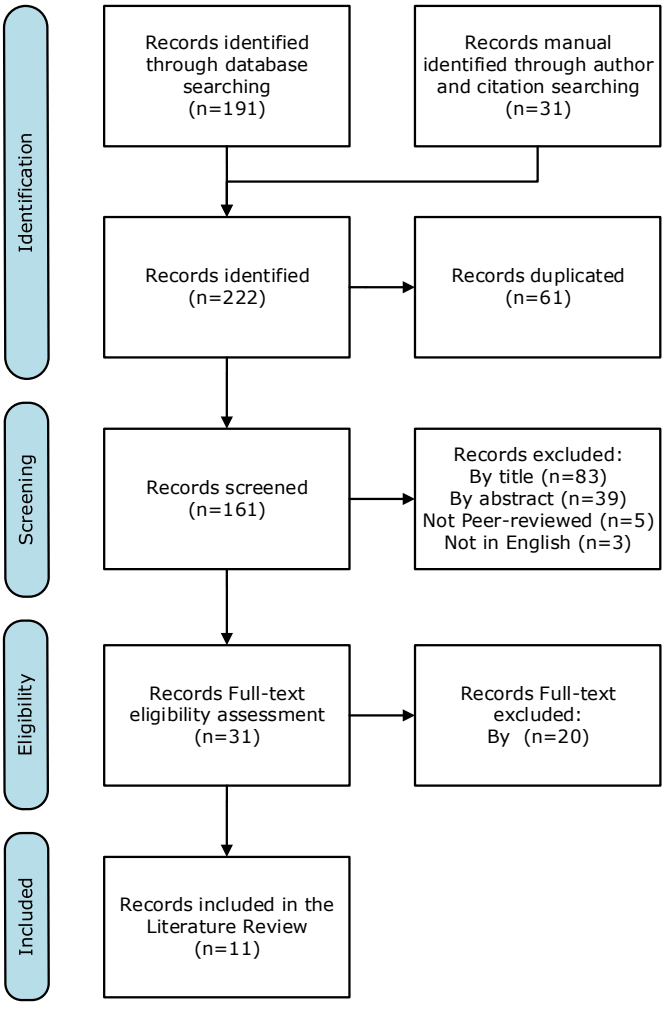


Figure 1.1: Literature Review PRISMA Method Flowchart

1.3.4 State-of-the-art Table and Discussion

To synthesize the evidence from the literature review, Table 1.5 was created. Below, it is discussed to answer the research questions posed previously.

Table 1.5: State-of-the-art for Cardiovascular Risk Prediction and Assessment Table

Ref.	Year	Features	Target	Dataset	Model	Validation	Evaluation
[6]	2021	Age, HDL-C, SBP, diabetes, hypertension, smoking	No CVD mortality/ CVD mortality	NWAHS, AusDiab MCCS	LR, RF, LDA, SVM	2-fold CV	90% AUROC, 85% SE, 84% SP
[7]	2024	Lipis, BP, smoking, drinking, age, sex, renal factors	No CVD occurrence/ CVD occurrence	Taiwan Biobank	LR, RF, NN, GB, BN	Hold-out	85% AUROC, 78% SE, 76% SP, 76% Acc, 56% F1
[8]	2016	Lipids, smoking, sex, BP, BMI, diabetes, hypertension	No CVD/ soft CVD/ hard CVD	BMES	LR, SVM	Hold-out	71% AUROC, 68% SE, 86% SP
[9]	2000	age, sex, BP, lipis, smoking, diabetes, LVH on ECG, BMI	No CVD occurrence/ CVD occurrence	INDIANA	LR, MLP, CART	Hold-out	78% AUROC, 76% Acc
[10]	2024	Age, sex, smoking, obisity, lipids, BP, renal, hemoglobin	No CVD occurrence/ CVD occurrence	Private from Hong Kong	XGB	Hold-out	Female: 76% Harrell's C Male: 74% Harrell's C
[11]	2023	Lipids, glycaemia, obesity, BP, renal factors	No CVD occurrence/ CVD occurrence	CHERRY	XGB, LASSO regression	5-fold CV	Ovr: 79%, Male: 77% Female: 81% AUROC
[12]	2022	203 risk factors	No CVD occurrence/ CVD occurrence	UK Biobank	LR, XGB	3-fold CV	76% AUROC
[13]	2019	473 risk factors	No CVD occurrence/ CVD occurrence	UK Biobank	15 Classifiers	10-fold CV	77% AUC
[14]	2021	77 risk factos, ASCVD risk	No CVD mortality/ CVD mortality	CAC Consortium	Ensemble boosting	10-fold CV	86% AUROC
[15]	2023	645 risk factors	No CVD occurrence/ CVD occurrence	UK Biobank	LGBM, RF, XGB, LR, SVM, KNN, ANN	LOOCV	76% AUROC
[16]	2016	7 APG features	Heathy/ At risk of CVD	Private from Malaysia	Extreme Learning Machine (ELM)	4-fold CV	86% Accuracy

HDL-C = High-Density Lipoprotein Cholesterol; SBP = Systolic Blood Pressure; BP = Blood Pressure; BMI = Body Mass Index; LVH = Left Ventricular Hypertrophy; APG = Acceleration Plethysmogram; LR = Logistic Regression; XGB = Extreme Gradient Boosting; RF = Random Forest; LDA = Linear Discriminant Analysis; SVM = Support Vector Machine; NN = Neural Network; GB = Gradient Boosting; BN = Bayesian Network; LGBM = Light Gradient Boosting Machine; KNN = K-Nearest Neighbors; ANN = Artificial Neural Network; LASSO = Least Absolute Shrinkage and Selection Operator; CV = Cross-validation; AUROC = Area Under the Receiver Operating Characteristic; SE = Sensitivity; SP = Specificity; Acc = Accuracy; F1 = F1-score.

Most studies relied primarily on traditional cardiovascular risk factors as predictors in their models [6–11]. A smaller number of works expanded the feature space substantially, incorporating numerous demographic, clinical, and lifestyle variables [12–15]. Only a single study was found to employ a physiological bio-signal for feature extraction [16].

The target variable varied across studies, some works used CVD occurrence as the target [7–13, 15], whereas others used mortality (death caused by a CVD) [6, 14]. Mortality outcomes often yield higher apparent discrimination but reflect a narrower slice of the disease burden. Only one study used risk as the target [16].

Most of the datasets used in the reviewed studies are private, institution specific or local cohorts with restricted availability, limiting external access and reproducibility. In contrast, only a small number of large resources are available through controlled access (application and fee), notably the UK Biobank, the Taiwan Biobank, and the CAC Consortium.

Validation in the reviewed studies was performed with common state-of-the-art classifiers, most frequently logistic regression (LR), support vector machines (SVM), random forests (RF), and gradient-boosted trees. Several works also assessed k-nearest neighbors (KNN), multilayer perceptrons (MLP)/artificial neural networks (ANNs), decision-tree (DT) variants, Bayesian networks, Least Absolute Shrinkage and Selection Operator (LASSO) regression, extreme learning machines (ELM), and ensemble boosting.

Performance was generally high, with reported Area Under the Receiver Operating Characteristic Curve (AUC) in the 0.70s [12, 13, 15], AUC around 0.75-0.85 [7, 10, 11, 14], and one study even reached 0.9 AUC [6]. When provided, accuracies were commonly 0.76-0.86, sensitivities around 0.68–0.85, and specificities 0.84-0.86. Some studies reported Harrell's C of roughly 0.74-0.76, and an occasional lower F1 score (0.56), likely reflecting class imbalance in the particular cohorts.

However, the addition of more risk factors does not necessarily improve performance. On the contrary, studies using very large feature sets, sometimes in the hundreds, often perform worse, on average, than those restricted to the established variables used in CVD risk scores. A likely reason is the curse of dimensionality, where, in high-dimensional spaces, data become sparse, distance measures lose discriminative power, and models are more prone to overfitting, especially with modest sample sizes.

Several consistent empirical regularities emerge across studies that cannot be concluded from the table but helped inform the design of our methodology. Age is invariably among the top predictors in feature importance analyses regardless of model family [7, 12, 13, 15]. Likewise, sex stratification often yields small but meaningful performance gains, both because baseline risk differs by sex and because the distribution of risk factors and their effects is sex specific, recent studies that reported sex specific Harrell's C illustrate this point [10][11]. These findings motivate modeling men and women separately whenever feasible.

As the datasets used by the studies in the table suggest, CVD prediction is fundamentally a data availability problem, developing a robust model requires datasets with a large number of variables and

longitudinal follow-up data collected several years later. Such datasets are rare, and even rarer when they include an ECG for each participant. This scarcity largely explains why most studies remain anchored in demographics, lipid profiles, blood pressure, diabetes status, smoking habits, and related laboratory measures, rather than incorporating ECG features. When biosignals are used, they are often restricted to coarse ECG surrogates, such as left ventricular hypertrophy (LVH) diagnosis in [9], or to alternative waveforms like the acceleration photoplethysmogram (APG) for "at risk" detection rather than for predicting hard events [16]. These choices reflect limitations in data availability rather than a definitive conclusion about the intrinsic predictive value of physiological signals or the features derived from them.

1.3.5 Research Gap

Recent efforts have focused on achieving incremental gains by incorporating more conventional risk factors into increasingly complex classifiers and regression models, rather than reducing the number of inputs or exploring new input modalities, such as physiological biosignals like ECG. This is largely due to the belief that physiological biosignals do not exhibit significant differences between risk groups. As shown in two studies reviewed here, physiological features were less statistically significant than clinical or lifestyle features. For this reason, most studies avoid including such features in their models, and certainly do not employ a methodology centered on them, as we propose to do in this dissertation.

When viewed collectively, the field appears to be pursuing incremental gains by aggregating more conventional risk factors into increasingly capable classifiers and regression models, rather than reducing the number of inputs or shifting modality, for example, to physiological bio-signals such as ECG. This is largely due to scarce data and the belief that physiological bio-signals do not exhibit significant differences between risk groups. In two studies that tested ECG features' statistical significance [11, 14], they were less statistically significant than clinical or lifestyle features. For this reason, most studies avoid including such features in their models, and certainly do not employ a methodology centered on them.

Despite cardiovascular risk scores' widespread adoption, traditional cardiovascular risk scores require numerous blood tests, which can be both time-consuming and costly, particularly in resource-limited settings. Recent advancements in AI have aimed to enhance cardiovascular risk prediction accuracy by incorporating additional features derived from even more extensive blood analyses. At the same time, ECG reading devices have become increasingly portable, cost-effective, and widely accessible [17, 18]

The literature suggests a clear gap that this dissertation seeks to address. The dominant research direction continues to layer traditional risk factors into slightly better learners, whereas few studies pursue metric parsimony, achieving competitive performance with fewer, cheaper, or more easily obtainable inputs. No studies have attempted a modal shift toward ECG features, and based on the current state-of-the-art analysis, it remains unclear which types of ECG features (linear, non-linear,

morphological) might perform best. As noted earlier, some prior reports even suggest that individual physiological features may not be sufficiently discriminative on their own [9]. However, this is precisely the hypothesis space we explore, whether carefully engineered ECG features, combined with minimal, rapid, and easily communicated clinical context (MRECCC), defined here as age, BMI, smoking status, and diabetes status, can assess risk, and whether this can be done in a way that supports clinician interpretability and adoption.

1.4 Objectives

Given the significant GBD of cardiovascular disease and the evident research gap, this dissertation investigates the potential of combining ECG signals with a minimal set of rapidly obtainable clinical variables, such as age, Body Mass Index (BMI), obesity status, and smoking history, for cardiovascular risk prediction. These variables were included because they can be collected without compromising the objective of rapid risk assessment, and because the literature consistently identifies them, particularly age, as highly influential predictors.

The first objective is to evaluate whether this framework contains sufficient information to accurately assess Framingham risk score strata. The second objective is to compare a range of machine learning methods to identify those most suitable for this task.

An additional aim is to examine the clinical viability of this approach, considering not only predictive accuracy but also usability and acceptance among cardiologists. This involves the development of a clinical application integrating the proposed risk assessment models, featuring a graphical user interface co-designed with clinicians and grounded in the principles of human-centered design.

To facilitate adoption, the Framingham risk score was selected as the prediction target. Although this choice was partly necessitated by the lack of longitudinal outcome data in available ECG datasets, it also provides the advantage of clinical familiarity and trust, which would likely be diminished for a novel score. While prior literature suggests potential benefits from sex-stratified modeling, the limited size of our dataset precluded such analyses.

1.5 Dissertation Structure

This dissertation is organized into five main chapters. Chapter 1 introduces the motivation for this research and situates it within the current clinical background. It also presents the systematic literature review, supported by a summary table, in order to identify research gaps in the field. These findings then guide the formulation of the dissertation's objectives.

Chapter 2 establishes the theoretical foundations of cardiovascular disease, covering epidemiology, pathophysiology, established risk scores, and the evaluation metrics used to assess those scores.

Chapter 3 presents the methodology used in the study. It begins with a description of the dataset, signal processing , and feature engineering. The machine learning section covers model validation

and evaluation. The chapter concludes with the design and development of the clinical application, including its structure, features, and components.

Chapter 4 discusses the results obtained from the ML models, as well as shows the clinical application. The findings are critically compared with existing literature, highlighting both consistencies and divergences from the state of the art.

Finally, Chapter 5 provides the conclusions of this research and outlines directions for future work. It reflects on the contributions of this dissertation to the field of cardiovascular risk assessment, the limitations of the current study, and potential pathways for extending this work in clinical and research contexts.

CHAPTER 2

Foundations of Cardiovascular Disease Risk Prediction

This chapter provides an overview of the main types of CVDs, including their epidemiology and pathophysiology. It then introduces the most widely used CVD risk scores, showing their parameters and methods of validation. Finally, the chapter discusses why ECG may contain valuable information that can inform machine learning models for CVD risk assessment.

Table 2.1 not only shows the main types of CVDS, but also the prevelent cases and deaths count in 2022, this data was taken by the study done by George A. Mensha *et al.* (2023) [3].

Table 2.1: Prevalence and Mortality of Major Cardiovascular Diseases in 2022 [3]

CVD type	**Cases**	**Deaths**
Ischemic heart disease	315,390,626	9,239,181
Ischemic stroke	86,661,746	3,542,299
Intracerebral hemorrhage	20,509,587	3,428,876
Subarachnoid hemorrhage	9,281,913	344,872
Hypertensive heart disease	13,052,641	1,353,074
Atrial fibrillation	55,414,434	362,381
Peripheral arterial disease	105,980,247	73,928
Rheumatic heart disease	46,358,651	386,947
Calcific aortic valve disease	13,551,699	146,199
Aortic aneurysm	NA	153,118

2.1 Epidemiology of Cardiovascular Disease

By the mid-twentieth century, coronary heart disease (CHD), the most common form of CVD, had emerged as a leading cause of death in industrialized countries. In response to this growing epidemic, the U.S. National Heart Institute launched the Framingham Heart Study (FHS) in 1948 [19], a landmark prospective cohort that enrolled 5,127 men and women aged 30 to 62 years from the town of Framingham, Massachusetts, all initially free of CHD. Participants underwent detailed examinations every two years, enabling researchers to track the incidence of CVD over time and to investigate potential determinants of disease onset.

Early follow-up reports from the study, published in the late 1950s, already pointed toward important trends [20]. They showed that CHD incidence was higher in older participants and in men, highlighting the importance of non-modifiable factors such as age and sex. At the same time, these reports identified measurable health attributes, including elevated blood pressure, high serum cholesterol, and excess body weight, that appeared to increase the likelihood of developing CVD.

William B. Kannel *et al.* (1961) [21] published a seminal paper entitled "*Factors of Risk in the Development of Coronary Heart Disease*", based on six-year follow-up data from the Framingham cohort. This publication is often credited with popularizing the term "risk factors" in the medical literature. The study provided the first quantitative evidence that hypertension, hypercholesterolemia, and left ventricular hypertrophy were major predictors of CHD incidence and mortality. Importantly, it demonstrated that these risks were not inevitable consequences of aging, but rather linked to specific, often modifiable, conditions and behaviors.

A crucial legacy of the Framingham study was the classification of risk factors into modifiable and non-modifiable categories. Hypertension, high serum cholesterol, cigarette smoking, and obesity were identified as modifiable risk factors, amenable to clinical intervention and lifestyle modification. Age and sex, on the other hand, were recognized as non-modifiable factors that nonetheless influenced baseline susceptibility. This framework has shaped preventive cardiology for decades, underscoring that while inherent traits cannot be altered, effective management of modifiable risks can substantially reduce the burden of cardiovascular disease.

Individual CVDs also have their specific risk factors. For example, ischemic heart disease (IHD) affects 5–8% of men and 2% of women over the age of 65, with men being six times more likely than women to develop clinically significant disease [22]. Ischemic stroke, which accounts for approximately 87% of all strokes, is similarly age-dependent, with incidence doubling every decade after the age of 55 [23]. Hemorrhagic forms of stroke, including intracerebral and subarachnoid hemorrhage, are less common but carry disproportionately high fatality rates [23]. Intracerebral hemorrhage occurs in 8–13% of strokes and increases nearly tenfold in individuals older than 85 years [24]. Subarachnoid hemorrhage has a lower incidence overall (9.1 per 100,000 annually) but demonstrates strong geographic variation, with higher prevalence in Japan and Finland [24].

Other CVD subtypes also reflect global disparities. Hypertensive heart disease remains highly prevalent in populations with poor blood pressure control, accounting for nearly half of all cases of heart failure in some cohorts [25]. Atrial fibrillation, the most common sustained arrhythmia, affected more than 52 million people worldwide in 2021, with a steady rise over the past two decades [26]. Rheumatic heart disease continues to dominate in low- and middle-income regions, particularly affecting children and young adults [27]. Degenerative and calcific valve diseases, by contrast, are strongly linked to aging and chronic comorbidities such as chronic kidney disease [27]. Aortic aneurysms, though less frequent, affect nearly 1% of adults between 30–79 years and occur four times more often in men [28].

In summary, the Framingham Heart Study transformed the understanding of cardiovascular disease by establishing the concept of modifiable and non-modifiable risk factors, laying the groundwork for modern preventive cardiology. Its pioneering findings demonstrated that the incidence of CVD is not solely a function of age or heredity but is powerfully shaped by controllable factors such as hypertension, cholesterol, smoking, and obesity. Subsequent research has expanded this framework across diverse CVD subtypes, revealing disease specific risk patterns and global disparities in prevalence.

2.2 Pathophysiology of Cardiovascular Diseases

James B. Herrick *et al.* (1912) [29] published a landmark description of coronary thrombosis as the underlying cause of myocardial infarction (MI), emphasizing that such events were not invariably fatal and could be diagnosed clinically with the aid of the ECG [29]. This shifted perception from sudden, unexplained death to a definable disease entity, further supported by subsequent clinicopathological correlations.

Advances in pathology also deepened the understanding of atherosclerosis. In 1913, Nikolai Anitschkow demonstrated experimentally that a cholesterol-rich diet produced arterial lesions in rabbits, implicating lipids in human vascular disease.

The mid to late twentieth century brought decisive advances in mechanistic and therapeutic knowledge. Marcus A. DeWood *et al.* (1980) [30] confirmed that acute MI was almost universally associated with thrombotic occlusion of a coronary artery, catalyzing the development of reperfusion therapies such as thrombolysis and percutaneous coronary intervention. Around the same time, the discovery of the LDL receptor by Michael S. Brown and J. L. Goldstein (1979) [31] established the molecular basis of cholesterol metabolism, providing the rationale for statin therapy, later validated in large randomized trials. Parallel work, like Frederick C. Novelho (1957) [32], in hypertension led to the acceptance of blood pressure control as a cornerstone of prevention, accelerated by the advent of effective diuretics in the 1950s and subsequent antihypertensive classes.

Further Physiological and molecular studies have clarified how major risk factors promote vascular injury. Elevated LDL cholesterol drives atherogenesis by entering the arterial wall [33], undergoing oxidation [34], and stimulating foam cell formation [35]. Hypertension causes endothelial dysfunction

through mechanical stress and activation of vasoactive mediators [36], while smoking and diabetes potentiate oxidative stress, inflammation, and thrombosis [37]. These converging mechanisms confirmed the centrality of modifiable risks in CVD pathophysiology and justified aggressive preventive interventions.

Heart failure understanding underwent a similar paradigm shift. Initially treated symptomatically with diuretics and digitalis, it came to be recognized as a systemic neurohormonal disorder, with chronic activation of the sympathetic nervous system and renin, angiotensin, aldosterone system driving disease progression [38]. This led to therapies such as ACE inhibitors, beta-blockers, and aldosterone antagonists, which were shown in landmark trials to improve survival [39]. More recently, agents like angiotensin receptor, neprilysin inhibitors and SGLT2 inhibitors have further improved outcomes, reflecting the integration of molecular cardiology with therapeutics [40].

In the twenty-first century, cardiovascular research has expanded into genomics, inflammation biology, and advanced imaging. Atherosclerosis is now understood as both a lipid-driven and immune-mediated process, as confirmed by anti-inflammatory interventions such as the CANTOS trial [41]. Imaging innovations allow identification of vulnerable plaques, while interventional cardiology has refined techniques with drug-eluting stents and minimally invasive surgery. Parallel advances in heart failure have differentiated subtypes such as HFrEF and HFpEF, with distinct pathophysiological pathways and tailored treatments [42].

2.3 Cardiovascular Risk Scores

The global rise of cardiovascular diseases, together with advances in epidemiological and pathological understanding stimulated the development of multivariable risk prediction models designed to quantify an individual's likelihood of future cardiovascular events. These risk scores combine several patient parameters into a single quantitative estimate of an individual's probability of developing CVD, or a specific type of CVD within a defined time horizon.

The development of these scores was driven by two main motivations. First, clinicians required practical tools to translate population-level findings into individualized risk assessment, enabling more precise targeting of preventive interventions. Second, health systems needed standardized methods to stratify patients according to risk, in order to guide treatment thresholds for antihypertensives, lipid-lowering drugs, and lifestyle interventions. Over time, multiple scores have been proposed and validated in diverse populations, each reflecting regional epidemiology and evolving understanding of CVD determinants.

Table 2.2 summarizes the most widely used cardiovascular risk scores,key predictor variables, and assessment time spans. In the present work, however, the analysis is restricted to the Framingham Risk Score (FRS).

The FHS also laid the foundation for the development of the FRS [5], it was the inaugural comprehensive cardiovascular risk prediction tool. Initially, it estimated 10-year risk of coronary heart

Table 2.2: Overview of major cardiovascular risk scores

Risk Score	Parameters	Outcome
Framingham [5]	Age, sex, SBP, TC, HDL-C, smoking, diabetes	10-year risk of CHD occurrence
QRISK3 [43]	Age, sex, SBP, HDL-C, BMI, smoking, diabetes, ethnicity + 13 parameters	10-year risk of cardiovascular events
SCORE [44]	Age, sex, SBP, TC, smoking	10-year risk of cardiovascular mortality
ASCVD [45]	Age, sex, SBP, TC, HDL-C, BP treatment, ethnicity, smoking, diabetes	10-year risk of CHD occurrence
WHO/ISH [46]	Age, sex, SBP, TC, smoking, diabetes	10-year risk of cardiovascular events
PROCAM [47]	Age, sex, SBP, LDL-C, HDL-C, BP, family history, smoking, diabetes	10-year risk of CHD occurrence

SCORE: Systematic Coronary Risk Evaluation; ASCVD: Atherosclerotic Cardiovascular Disease; WHO: World Health Organization; ISH: International Society of Hypertension; PROCAM: Prospective Cardiovascular Münster Study; SBP: Systolic Blood Pressure; LDL-C: Low-Density Lipoprotein Cholesterol; HDL-C: High-Density Lipoprotein Cholesterol; BP: Blood Pressure.

disease using age, sex, LDL-C, HDL-C, BP, diabetes, and smoking status. Successive updates were released in 2002 and 2008 [48], with the latter reflecting the recommendations of the Adult Treatment Panel III of the National Heart, Lung, and Blood Institute. The 2008 revision broadened age ranges, included hypertension treatment status and dyslipidemia, and reclassified diabetes as a coronary heart disease risk equivalent rather than an independent predictor [48].

2.4 Validation and Evaluation of Cardiovascular Risk Scores

Early FHS analyses used multivariable logistic regression [21]. However, the main FRS study employed Cox proportional hazards regression, allowing a more flexible treatment of follow-up time and censoring [5].

2.4.1 Cox Proportional Hazards Model

The Cox proportional hazards model, introduced by David Cox (1972) [49], is a semi-parametric survival analysis framework designed to estimate the hazard function $h(t)$, which represents the instantaneous risk of experiencing an event at time t, conditional on survival up to that time point. The model can be expressed as:

$$h(t) = h_0(t) \times \exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right), \tag{2.1}$$

where $h_0(t)$ is the baseline hazard function, β_i are regression coefficients, and x_i are covariates. The exponential term $\exp(\beta_i)$ defines the hazard ratio (HR), which quantifies the relative change in risk associated with a one-unit increase in the covariate x_i.

The evaluation of CVD risk scores requires metrics that account for the time to event nature of cardiovascular outcomes. Unlike binary classification problems where a patient either has or does not have a condition, CVD risk models typically predict the probability of an event occurring within a specific time horizon. Because patients may be censored (lost to follow-up, or event free at the end of the study period), survival analysis based metrics are preferred over standard classification metrics such as accuracy.

2.4.2 Harrell's Concordance Index

The most widely used discrimination metric in this context is the Harrell's concordance index (C-index) [50]. It measures the model's ability to correctly rank patients according to risk. In practical terms, it is the probability that, for a randomly selected pair of patients, the one who experiences the event earlier had a higher predicted risk score.

Formally, let T_i denote the observed survival time for patient i, δ_i an event indicator ($\delta_i = 1$ if the event occurred, 0 if censored), and $\hat{r}_i$ the predicted risk score. Then the C-index is defined as:

$$C = \frac{\sum_{i<j} I(\hat{r}_i > \hat{r}_j) \cdot I(T_i < T_j, \delta_i = 1)}{\sum_{i<j} I(T_i < T_j, \delta_i = 1)} \tag{2.2}$$

where $I()$ is the indicator function. The numerator counts concordant pairs (predicted risk correctly matches observed outcome order), while the denominator counts all comparable pairs. A value of 0.5 corresponds to random ranking, while 1.0 indicates perfect concordance.

2.4.3 Time-dependent ROC and AUC

Another approach extends the receiver operating characteristic (ROC) framework to survival data. At a fixed time horizon t, one can define sensitivity (true-positive rate) and specificity (true-negative rate) for the predicted risks, and then compute a time-dependent AUC(t). This accounts for censoring through inverse probability weighting or Kaplan–Meier estimators [51].

$$\text{AUC}(t) = P(\hat{r}_i > \hat{r}_j \mid T_i \leq t, \delta_i = 1, T_j > t) \tag{2.3}$$

2.5 Electrocardiogram

The ECG is a non-invasive recording of the heart's electrical activity over time. The electrical impulses responsible for cardiac contraction originate in the sinoatrial (SA) node and propagate through the atria, the atrioventricular (AV) node, and the His–Purkinje system before depolarizing and repolarizing the ventricles [52].

At the body surface, this electrical activity can be captured as small voltage fluctuations, typically in the range of 0.5-4 mV, using electrodes placed on the skin. The standard ECG waveform consists of characteristic deflections, the P wave, QRS complex, and T wave, each corresponding to a distinct phase of atrial and ventricular depolarization or repolarization [53].

The clinical gold standard is the 12-lead ECG, in which electrodes are placed on the limbs and chest to provide multiple spatial projections of the heart's electrical field. Each lead represents the voltage difference between two electrode sites (bipolar) or between an electrode and a reference point (unipolar) [53].

For research and monitoring purposes, simplified configurations such as single-lead (1-channel) or three-lead ECGs are often used. These systems provide less spatial detail but are easier to acquire and can be sufficient for automated analysis tasks. The signals are typically sampled at frequencies ranging from 250 Hz to 1000 Hz, which provides enough temporal resolution.

The ECG provides a direct window into the electrophysiological and structural state of the heart. Many CVDs manifest through abnormalities in conduction, rhythm, or waveform morphology that can be detected in the ECG.

For instance, ischemic changes often appear as ST-segment deviations or T-wave abnormalities, reflecting either acute or chronic myocardial ischemia [29]. Disorders of rhythm, such as atrial fibrillation or ventricular tachyarrhythmias, are also primarily diagnosed based on characteristic ECG patterns. Structural alterations like left ventricular hypertrophy or conduction disturbances can be inferred from hypertrophy and remodeling markers, which include QRS voltage criteria, axis deviations, and the presence of bundle branch block [54]. Finally, the ECG also provides information about heart rate dynamics, including heart rate variability (HRV), they can offer indirect but clinically relevant insight into autonomic balance, which has been strongly associated with cardiovascular risk.

Beyond traditional interpretation, modern computational methods can extract high-dimensional features from ECG signals, such as wavelet coefficients, spectral components, or morphological markers, which may capture subtle patterns not visible to the naked eye. This opens the possibility of using ECG not only as a diagnostic tool but also as a predictive biomarker for long-term CVD risk stratification.

CHAPTER 3

Methodology

The methodology is divided into four main sections: (i) data collection and processing, which defines and explores the database and details signal preprocessing (normalization, filtering, windowing); (ii) feature extraction and engineering, including feature compression and selection; (iii) machine learning, covering model validation, classification, and regression; and (iv) clinical application, describing the development of the user interface.

Below, in Figure 3.1, there is a flowchart summarizing the methodology used.

3.1 Database

This dissertation utilizes a curated private dataset from 53 patients, each of whom underwent the blood tests and clinical assessments required to calculate the Framingham Risk Score (FRS). In addition to these clinical measurements, the dataset includes single-channel (1-lead) ECG recordings for each patient, with a 1000 Hz sampling rate and recording durations ranging between 5 and 6 minutes. The dataset was collected in Brazil and provided by a consortium of Brazilian cardiologists from the Walter Cantídio University Hospital (HUWC) and the Leonardo da Vinci State Hospital (HELV). All patients provided written informed consent, and the study protocol was approved by the ethics committee under number 45360915.1.1001.5262. This research specifically utilizes only the ECG recordings and their corresponding FRS values. The FRS values are categorized into three risk stages: Stage 1 denotes a low 10-year cardiovascular risk (0–9%), Stage 2 indicates a moderate risk (10–15%), and Stage 3 represents a high risk (16–100%). The database is complete, without missing data, and is statistically summarized in Table 3.1 and Table 3.2. As shown in the tables, the database is balanced in terms of stage distribution but not in terms of gender distribution, being composed mainly of female patients.

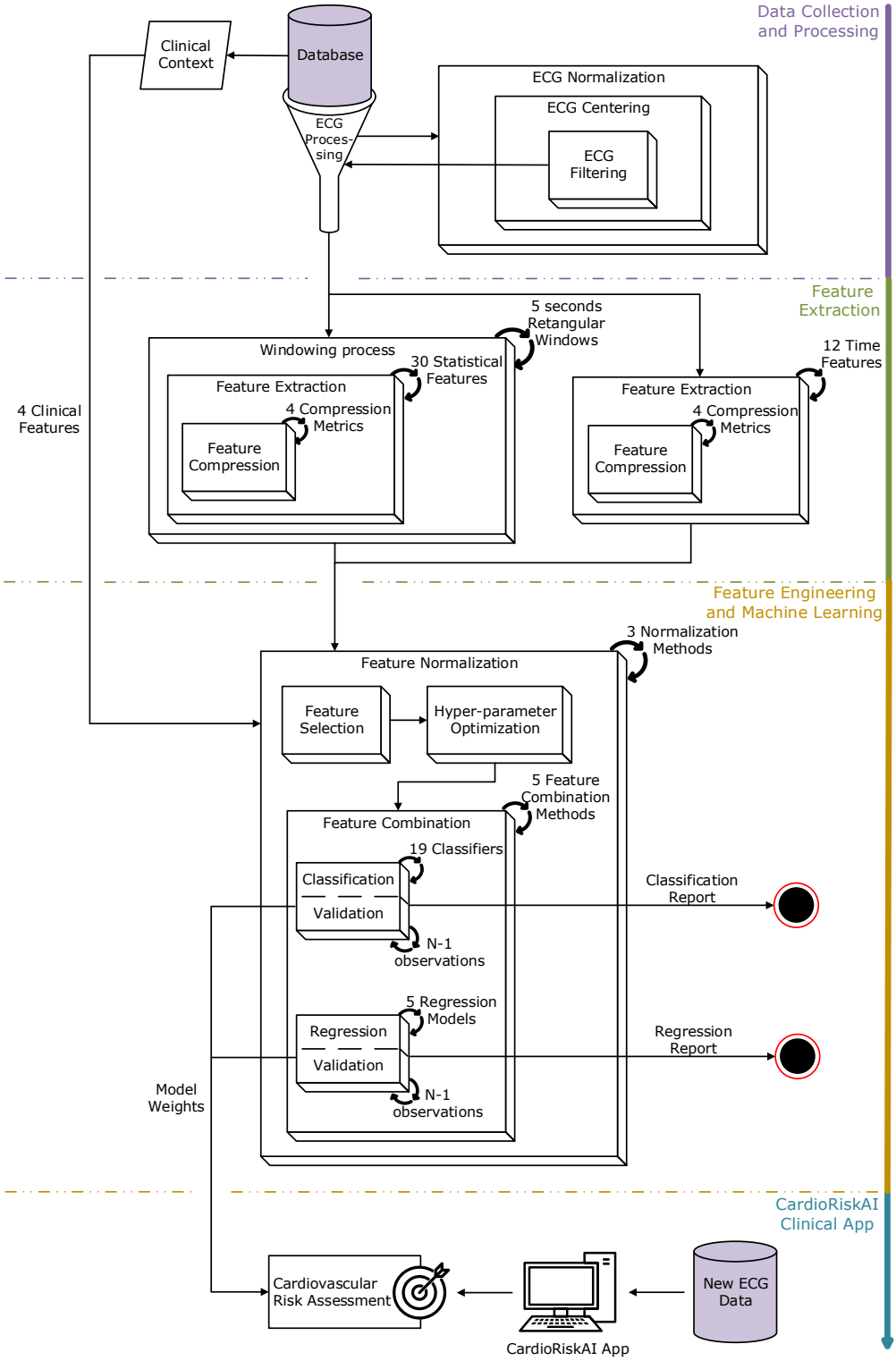


Figure 3.1: Methodology Summary Flowchart

Table 3.1: Database Statistics

Category	Count
Total	53
Number of Females	46
Number of Males	7
Mean Age (years)	38
Framingham Score 1 Count	22
Framingham Score 2 Count	14
Framingham Score 3 Count	17

Table 3.2: Database Statistics by Stages

Category	Stage 1	Stage 2	Stage 3
Number of Patients	22	14	17
Number of Females	21	10	15
Number of Males	1	4	2
Mean Age (years)	35.91	48.71	54.59

3.2 Signal Processing

In this section, we describe the signal processing methods employed to preprocess the raw ECG data.

3.2.1 Normalization

The initial preprocessing step in this pipeline is raw ECG signal normalization, a technique used to mitigate amplitude variations between patients.. These amplitude differences can originate from diverse sources such as variability in recording conditions including differing electrode impedances and amplifier gains as well as physiological variations across patients, notably differences in body composition, chest wall thickness, and heart positioning. If unaddressed, these amplitude variations can introduce artificial inter-patient variability that does not correlate with cardiac health, bias machine learning algorithms toward signals with inherently higher amplitudes, diminish the effectiveness of subsequent filtering processes, and yield inconsistent outcomes in feature extraction steps.

The Root Mean Square (RMS) normalization was applied to the ECG signal $x[n]$ of length N. The RMS value is computed as:

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x[i]^2}, \tag{3.1}$$

and each point of the ECG signal is then normalized by this value, yielding the normalized signal:

$$x_{\text{RMS}}[n] = \frac{x[n]}{\text{RMS}}. \tag{3.2}$$

This normalization method was specifically selected due to several advantages. Unlike min-max or z-score normalization techniques, RMS normalization inherently preserves the relative amplitudes of ECG waveforms, such as the critical P, QRS, and T components. It scales the signals based on their energy content, providing a physiologically meaningful form of amplitude adjustment. Additionally, RMS normalization demonstrates robust performance in the presence of occasional spikes or transient artifacts, unlike peak-based normalization methods which are highly sensitive to such outliers. Lastly, performing RMS normalization as an initial step establishes consistent amplitude levels across signals, thereby facilitating optimal performance in subsequent filtering and feature extraction processes.

3.2.2 Centralization

Centralization is a critical preprocessing step in ECG signal analysis, primarily employed to remove the Direct Current (DC) offset, a constant vertical shift in the signal that can arise from instrumentation or electrode imbalance. This offset can vary substantially between different recordings, potentially biasing amplitude dependent features.

The centralization process involves adjusting each discrete ECG sample by subtracting the mean value of the entire signal. Specifically, given a discrete ECG signal $x[n]$ of length N, the centralized signal $x_c[n]$ is obtained as follows:

$$x_c[n] = x[n] - \mu_x \tag{3.3}$$

where μ_x represents the mean of the signal and is calculated by:

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x[i] \tag{3.4}$$

3.2.3 Filtering

Following normalization and centralization, the ECG signals undergo bandpass filtering. Despite these preprocessing steps, ECG recordings may still contain undesirable frequency components, including baseline wander and high-frequency noise.

Low frequency artifacts typically range between 0 and 0.5 Hz [55]. Examples of such artifacts are baseline wander, primarily caused by respiration, patient movements, and electrode skin impedance variations. Additionally, low frequency noise can include DC drift resulting from slow electrode polarization and amplifier temperature fluctuations, as well as postural changes that may mimic genuine cardiac waveforms, potentially leading to false detections. Similarly, high frequency artifacts, generally present above 50 Hz, arise from muscle activity (EMG), powerline interference, and electronic noise [55].

To mitigate these artifacts, a third-order Butterworth bandpass filter was implemented, employing conservative cutoff frequencies of 0.5 Hz (high pass) and 35 Hz (low pass). The filter was designed and applied at the dataset's sampling frequency of 1000 Hz. The continuous-time transfer function of this bandpass filter was constructed as the product of a high-pass component $H_{HP}(s)$ and a low-pass component $H_{LP}(s)$:

$$H(s) = H_{HP}(s) \cdot H_{LP}(s) \tag{3.5}$$

where the individual components are defined as:

$$H_{HP}(s) = \frac{\left(\frac{s}{\omega_1}\right)^3}{\left(1+\frac{s}{\omega_1}\right)^3}, \quad H_{LP}(s) = \frac{1}{\left(1+\frac{s}{\omega_2}\right)^3}, \tag{3.6}$$

with the angular cutoff frequencies:

$$\omega_1 = 2\pi \cdot 0.5\,\text{rad/s}, \quad \omega_2 = 2\pi \cdot 35\,\text{rad/s}. \tag{3.7}$$

For practical implementation in a digital system, the analog filter was converted to a discrete-time filter using the bilinear transformation, yielding the transfer function:

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_6 z^{-6}}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_6 z^{-6}}. \tag{3.8}$$

The digital transfer function $H(z)$ is implemented computationally through the difference equation:

$$y[n] = \sum_{k=0}^{6} b_k x[n-k] - \sum_{j=1}^{6} a_j y[n-j], \tag{3.9}$$

Where $y[n]$ represents the current output sample, $x[n-k]$ are the input samples delayed by k time steps, and the coefficients $\{b_k\}$ and $\{a_j\}$ are derived from the bilinear transformation. This recursive formulation shows that each output is a weighted combination of the current and six previous input samples, along with six past output values [56].

To preserve the temporal fidelity and morphological features of the ECG signals, a zero-phase filtering technique was employed. This approach involves applying the filter twice first in the forward

direction, and then again in reverse which cancels out any phase shifts introduced by the filter. Mathematically, this bidirectional filtering results in an effective transfer function:

$$H_{\text{eff}}(z) = |H(z)|^2 \tag{3.10}$$

As a result, the overall phase response is neutralized to exactly zero at all frequencies, ensuring that the locations and shapes of key ECG components remain unaffected while the magnitude response becomes the square of the original filter's magnitude response.

The selected bandwidth of 0.5 to 35 Hz is intentionally narrower compared to the conventional diagnostic ECG frequency range of 0.5 to 50 Hz. This decision aligns with recent findings in the literature, suggesting that lowering the upper cutoff frequency of the bandpass filter effectively improves the overall quality and clinical utility of ECG signals [57].

Finally, adopting a third-order Butterworth filter provided a balance between achieving a sufficiently sharp transition (approximately 18 dB per octave roll-off) and maintaining signal integrity, verified through inspection of its frequency response, ensuring -3 dB attenuation at 0.5 and 35 Hz, zero phase condition, symmetric impulse response, and absence of filtering artifacts.

Figures 3.2 and 3.3 compare the raw ECG signal with its preprocessed counterpart, which has been normalized, centered, and filtered, respectively.

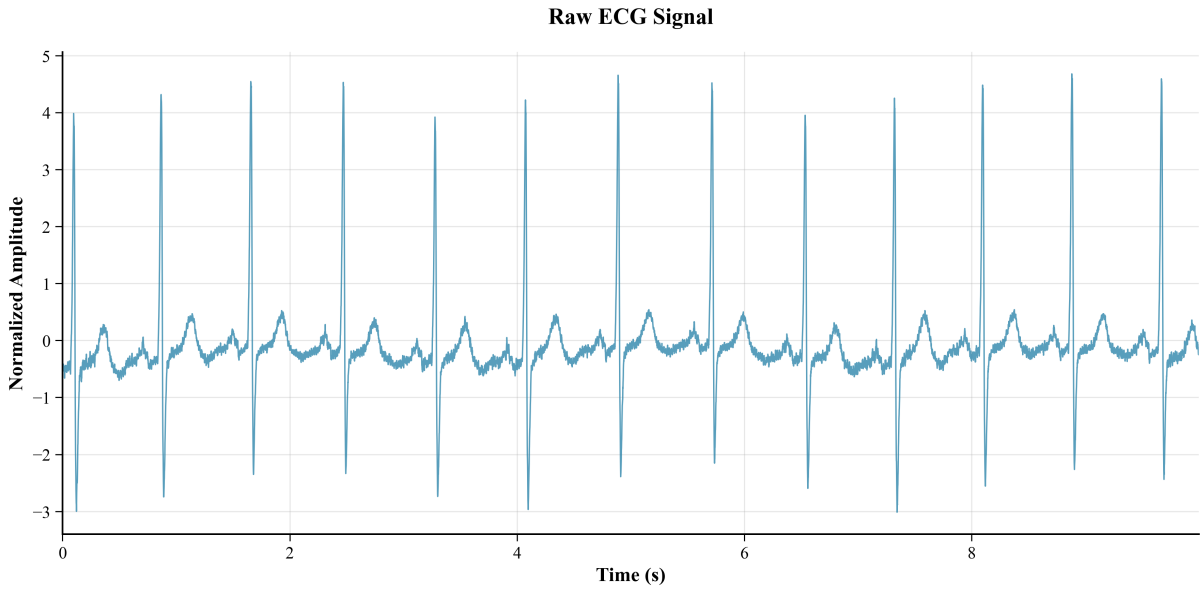


Figure 3.2: Plot of the Raw ECG signal

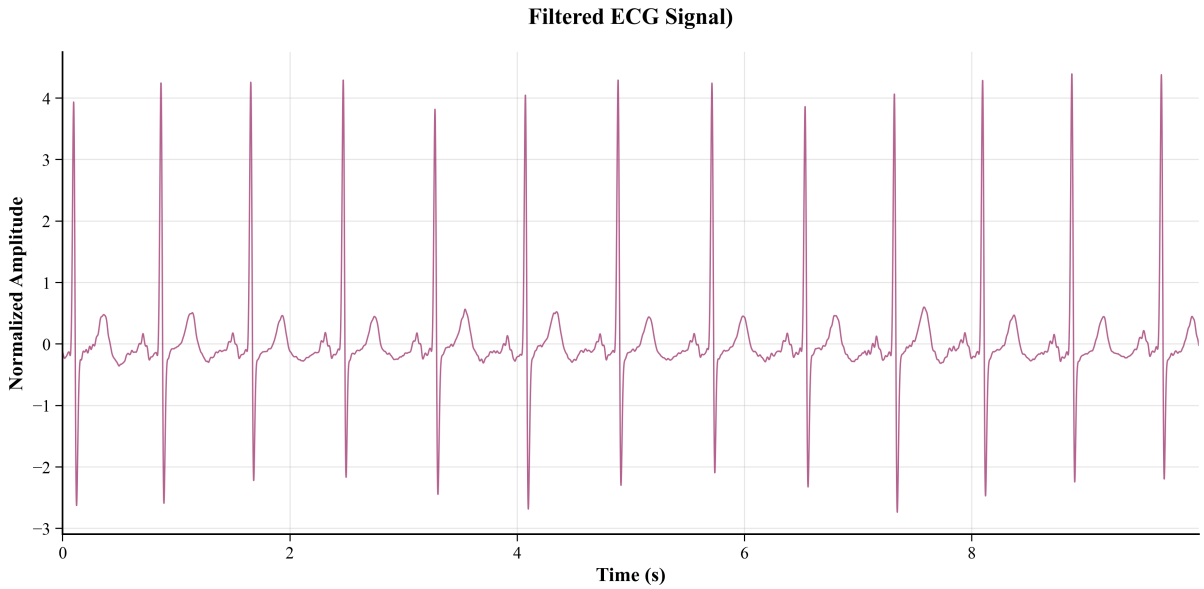


Figure 3.3: Plot of the filtered ECG signal

3.2.4 Windowing

Temporal segmentation, commonly referred to as windowing, is a crucial methodological step in the analysis of ECG recordings. Although ECG data inherently capture continuous cardiac activity typically lasting between 5 and 6 minutes in this dataset direct analysis of entire recordings presents multiple challenges. These include diminished feature stability over extended segments and reduced temporal resolution, making it difficult to track dynamic physiological changes effectively [58].

To address these issues, this study adopts a fixed-duration, non-overlapping windowing approach utilizing a rectangular window function. This method is formally defined through the following parameter. Firstly, the window length L is calculated as:

$$L = T_w \cdot f_s = 5\,\text{s} \times 1000\,\text{Hz} = 5000\,\text{samples}, \tag{3.11}$$

where T_w denotes the window duration in seconds and f_s represents the sampling frequency. A non-overlapping step size equal to the window length is employed, given by:

$$\Delta = L = 5000\,\text{samples}. \tag{3.12}$$

A rectangular (uniform) window function $w[n]$ is applied to the signal, mathematically expressed as:

$$y[n] = x[n] \cdot w[n] \tag{3.13}$$

where:

$$w[n] = \begin{cases} 1, & 0 \leq n < L \\ 0, & \text{otherwise} \end{cases} \tag{3.14}$$

The rectangular windowing approach was selected due to its ability to preserve the original morphology of the ECG signal [59]. Unlike tapered window functions, for example, Hamming or Hann, the rectangular window does not introduce amplitude modulation or edge attenuation, thus maintaining the diagnostic features of the P, QRS, and T waves. These properties are particularly advantageous when planning to extract ECG delineation features, where the accuracy of waveform amplitudes and durations is essential.

A window duration of five seconds was chosen as the minimum length that remains consistent with good scientific practice. This duration provides a sufficient number of samples to ensure reliable convergence in frequency domain feature calculations. From a physiological standpoint, a five second window typically captures approximately five to eight complete cardiac cycles at resting heart rates.

3.3 Feature Extraction

The features listed in Table 3.3 are organized into eight functional groups based on the type of signal characteristic they quantify. Statistical features (f_1–f_6) capture basic distributional properties of the signal such as central tendency, dispersion, and energy. Fractal dimension features (f_7–f_8) quantify geometric complexity using self-similarity metrics [60]. Entropy features (f_9–f_{14}) assess the irregularity, unpredictability, and chaotic informational content of the signal [61]. Complexity features (f_{15}–f_{18}) focus on dynamic and structural properties including pattern diversity and nonlinear behavior [62]. Spectral features (f_{19}–f_{23}) describe frequency domain properties and power distribution [63]. Recurrence Quantification Analysis (RQA) features (f_{24}–f_{27}) are derived from recurrence plots and capture repetitive dynamics [64]. Phase space features (f_{28}–f_{30}) reflect geometric properties of the reconstructed trajectory in phase space [65]. Morphological features (f_{31}–f_{36}) characterize key waveform components such as durations and shapes of the P, QRS, and T waves. Finally, Heart Rate Variability (HRV) features (f_{37}–f_{42}) quantify temporal variability in beat-to-beat intervals.

Table 3.3: Summary of all Feature Extraction Metrics

ID	Feature Name	Equation	Description
f_1	Mean (μ)	$\frac{1}{N}\sum_{n=0}^{N-1} x[n]$	Average signal amplitude

ID	Feature Name	Equation	Description
f_2	Standard Deviation (σ)	$\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}(x[n]-\mu)^2}$	Variation around the mean
f_3	Range	$\max(x[n]) - \min(x[n])$	Difference between max and min values
f_4	Skewness	$\frac{\frac{1}{N}\sum_{n=0}^{N-1}(x[n]-\mu)^3}{\sigma^3}$	Asymmetry of the signal distribution
f_5	Kurtosis	$\frac{\frac{1}{N}\sum_{n=0}^{N-1}(x[n]-\mu)^4}{\sigma^4} - 3$	Peakedness of the signal distribution
f_6	Energy (E)	$\sum_{n=0}^{N-1}\lvert x[n]\rvert^2$	Total energy of the signal
f_7	Higuchi Fractal Dimension (HFD)	$-\frac{d\log_{10}(L(k))}{d\log_{10}(k)}$	Fractal complexity based on curve length
f_8	Katz Fractal Dimension (KFD)	$\frac{\log_{10}(L)}{\log_{10}(d/L+1)}$	Geometric measure of signal complexity
f_9	Shannon Entropy (H)	$-\sum_i p_i \log_2(p_i)$	Signal uncertainty measure
f_{10}	Approximate Entropy (ApEn)	$\text{ApEn}(m,r) = \phi(m) - \phi(m+1)$	Regularity and unpredictability of fluctuations
f_{11}	Sample Entropy (SamEn)	$\text{SampEn}(m,r) = -\log\left(\frac{A}{B}\right)$	Improved entropy measure over ApEn
f_{12}	Fuzzy Entropy (FuzzyEn)	$\text{FuzzyEn}(m,r) = \log\left(\frac{\phi^m(r)}{\phi^{m+1}(r)}\right)$	Entropy using fuzzy logic membership
f_{13}	Permutation Entropy (PE)	$-\sum_\pi p(\pi)\log(p(\pi))$	Entropy based on ordinal pattern distribution
f_{14}	Conditional Entropy (H)	$H(X_{m+1}\lvert X_m) = H(X_{m+1}) - H(X_m)$	Entropy conditioned on previous value
f_{15}	Lempel-Ziv Complexity (LZC)	$\frac{c(n)}{n/\log_2(n)}$	Number of distinct patterns in signal
f_{16}	Largest Lyapunov Exponent (λ)	$\lim_{t\to\infty}\frac{1}{t}\log\left(\frac{\lvert\delta(t)\rvert}{\lvert\delta(0)\rvert}\right)$	Rate of divergence of nearby trajectories
f_{17}	Correlation Dimension (D_2)	$\lim_{r\to 0}\frac{\log C(r)}{\log r}$	Dimensionality of attractor in phase space

ID	Feature Name	Equation	Description
f_{18}	Detrended Fluctuation Analysis (F)	$F(n) \propto n^{\alpha}$	Self-similarity of signal fluctuations
f_{19}	Power Spectral Density (P_{xx})	$P_{xx}(f) = \frac{1}{LU} \sum \lvert X_i(f) \rvert^2$	Frequency power distribution (Welch)
f_{20}	Spectral Entropy (H_s)	$-\sum_f \frac{P(f)}{\sum P(f)} \log_2 \left(\frac{P(f)}{\sum P(f)} \right)$	Frequency-domain uncertainty measure
f_{21}	Dominant Frequency (f_{dom})	$\arg\max_f P_{xx}(f)$	Frequency with highest power
f_{22}	Total Power (P_{total})	$\sum_f P_{xx}(f)$	Sum of power across all frequencies
f_{23}	Spectral Edge Frequency (f_{edge})	$f_{edge} : \sum_{i=0}^{f} P(i) = 0.95 \sum_i P(i)$	Frequency below which 95% of power resides
f_{24}	Recurrence Rate (RR)	$\frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j}$	Recurrence plot density
f_{25}	Determinism (Det)	$\frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{i,j} R_{i,j}}$	Fraction of points forming diagonal lines
f_{26}	Laminarity (Lam)	$\frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{i,j} R_{i,j}}$	Fraction of points forming vertical lines
f_{27}	Trapping Time (TT)	$\frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=v_{min}}^{N} P(v)}$	Mean vertical line length in recurrence plot
f_{28}	Phase Space Area (PSA)	$\pi \sigma_x \sigma_y$	Elliptical area in phase space
f_{29}	Centroid	$(\bar{x}, \bar{y}) = \left(\frac{1}{N} \sum x_i, \frac{1}{N} \sum x_{i+\tau} \right)$	Center of mass of phase trajectory
f_{30}	Orbit Stability (S)	$\frac{1}{N} \sum \sqrt{(x_i - \bar{x})^2 + (x_{i+\tau} - \bar{y})^2}$	Distance of points from phase center
f_{31}	QRS Duration	$(S_{peak} - Q_{peak}) \cdot \frac{1000}{f_s}$	Duration of QRS complex
f_{32}	PR Interval	$(Q_{peak} - P_{peak}) \cdot \frac{1000}{f_s}$	Onset of P to onset of Q
f_{33}	QT corrected Interval (QTc)	$\frac{\mathrm{QT}}{\sqrt{RR}}$	QT interval corrected for heart rate

ID	Feature Name	Equation	Description
f_{34}	Wave Area (WA)	$\int (x(t) - \text{baseline})dt$	ECG wave area using trapezoidal integration
f_{35}	Wave Symmetry (WS)	$\left\lvert \frac{\text{slope}_{asc}}{\text{slope}_{desc}} \right\rvert$	Ratio of wave ascent and descent slopes
f_{36}	P Wave Dispersion	$\max(P_{duration}) - \min(P_{duration})$	Max difference in P wave duration
f_{37}	RMSSD	$\sqrt{\frac{1}{N-1}\sum(\Delta RR_i)^2}$	Root mean square of RR intervals
f_{38}	pNN50	$\frac{\#\{\lvert\Delta RR_i\rvert > 50\text{ms}\}}{N-1} \times 100\%$	Proportion of RR intervals differing by $>$50 ms
f_{39}	SDANN	$\text{SD}(\overline{RR}_i)$	SD of mean RR intervals in segments
f_{40}	Triangular Index (TI)	$\frac{N}{\max(\text{histogram})}$	HRV triangular index
f_{41}	SD1	$\frac{\text{SD}(\Delta RR)}{\sqrt{2}}$	Short axis of Poincaré ellipse
f_{42}	SD2	$\sqrt{2 \cdot \text{Var}(RR) - \text{SD1}^2}$	Long axis of Poincaré ellipse

Features f_1 through f_{30} were extracted using the sliding window approach described in the Signal Processing section. In this method, the ECG signal is segmented into windows, and each window is independently analyzed to compute those features.

In contrast, features f_{31} through f_{42} consist of morphological and beat-to-beat characteristics such as wave durations, intervals, and heart rate variability (HRV) metrics. These features rely on accurately detecting fiducial points, onsets, peaks, and offsets of the P, QRS, and T waves, and require a continuous view of the cardiac cycle. Extracting these features from windowed segments would risk cutting through waves, leading to incomplete or unreliable measurements.

To address this, we opted to extract morphological and HRV features from the entire ECG recording, analyzing as many complete waveforms as could be identified. This was made possible through the application of an ECG delineation algorithm, which detects wave boundaries and allows for the calculation of clinically relevant intervals and variability metrics. The next section provides details on the delineation algorithm used.

3.4 ECG Delineation

Initially, to delineate the ECG signal, a more straightforward approach was explored using the `NeuroKit2` Python library [66], which is widely adopted for ECG signal delineation in the scientific literature and community. While this method showed reliable performance in detecting R peaks and estimating RR intervals, it failed to provide accurate delineation for other waveform components, particularly the P and T waves. Inaccuracies in locating the onsets and offsets of these waves significantly limited its applicability for precise morphological analysis.

Given these limitations, a more robust delineation method based on wavelet transforms was implemented, following the methodology proposed by Martínez et al. (2004) [67]. This algorithm leverages the multi-resolution properties of the discrete wavelet transform (DWT) to identify ECG fiducial points. Using a quadratic spline wavelet, the signal is decomposed into multiple scales while preserving temporal resolution through the algorithm.

3.4.1 Multiband Decomposition

The DWT of a finite-energy ECG signal corresponds to its decomposition into a set of basis functions derived from prototype sequences and their time-shifted versions [68]. This makes it an optimal tool for analyzing the time-frequency characteristics of the ECG, as it can capture both sharp transients (QRS complexes) and lower-frequency components (P and T waves) within a unified framework.

This expansion is implemented through an octave-band critically decimated filter bank [68, 69]. The m^{th} subband is confined to the normalized frequency range:

$$W_m = \begin{cases} \left[0, \pi/2^S\right], & m = 0, \\ \left[\pi/2^{S-m+1}, \pi/2^{S-m}\right], & m = 1,2,\ldots,S, \end{cases} \tag{3.15}$$

where S is the decomposition level, $S+1$ is the number of subbands, and π is the normalized angular frequency (equivalent to half the sampling rate).

The DWT employs a scaling analysis function $\tilde{\phi}_1(n)$ and a wavelet analysis function $\tilde{\psi}_1(n)$, which are the impulse responses of the half-band low-pass and high-pass filters, respectively [69]. Defining the recursive relations:

$$\tilde{\phi}_{i+1}(n) = \tilde{\phi}_i(n/2) * \tilde{\phi}_1(n), \tag{3.16}$$

$$\tilde{\psi}_{i+1}(n) = \tilde{\phi}_i(n) * \tilde{\psi}_1(n/2^i), \tag{3.17}$$

the equivalent analysis filter of the mth subband is given by:

$$h_m(n) = \begin{cases} \tilde{\phi}_S(n), & m = 0, \\ \tilde{\psi}_{S+1-m}(n), & m = 1,2,\ldots,S. \end{cases} \tag{3.18}$$

The m^{th} subband signal, $m = 0, 1, \ldots, S$, is therefore obtained as:

$$x_m(n) = \begin{cases} \sum_{k=-\infty}^{\infty} x(k) h_m(2^S n - k), & m = 0, \\ \sum_{k=-\infty}^{\infty} x(k) h_m(2^{S-m+1} n - k), & m = 1, 2, \ldots, S. \end{cases} \tag{3.19}$$

In the present work, the ECG was decomposed up to the fifth level ($S = 5$) using a quadratic spline wavelet, consistent with the delineation algorithm of Martínez et al. [67]. This multiband decomposition allowed simultaneous emphasis of high-frequency QRS components and low-frequency P/T wave components, thus enhancing the robustness of fiducial point detection in noisy clinical recordings.

The delineation proceeds in four main stages. First, QRS complexes are detected by tracing modulus maxima across scales and identifying zero crossings between opposite-sign extrema. Once located, individual QRS waves are delineated, with onsets and ends determined by thresholding relative to wavelet amplitudes. Next, T waves are identified within a search window after each QRS, using local maxima at specific scales to capture both monophasic and biphasic morphologies; their boundaries are refined using amplitude- and slope-based criteria. Similarly, P waves are searched in the pre-QRS interval, where their lower amplitude makes delineation more challenging; nevertheless, the multiscale approach helps discriminate them from baseline wander and noise.

The wavelet-based delineation approach demonstrated visually high accuracy in identifying the complete morphology of each heartbeat, as shown in Figure 3.4, which presents the delineated ECG output for one patient, displaying the first 10 seconds of the recording. This enabled the extraction of high-quality features such as QRS duration, PR interval, QTc interval, wave symmetry, and beat-to-beat variability metrics. This delineation step was essential for capturing subtle morphological and temporal characteristics that are critical for downstream classification tasks.

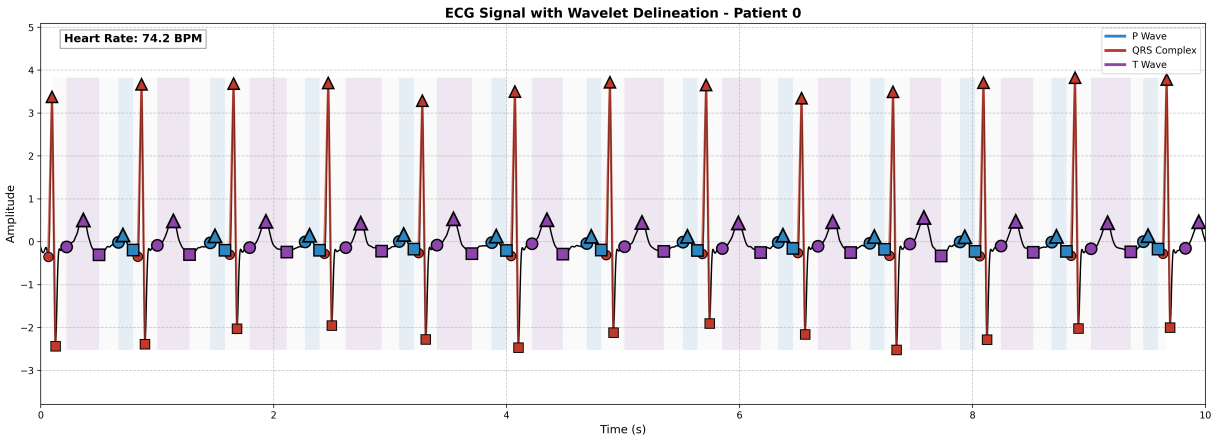


Figure 3.4: Plot of the Delineated ECG signal

3.5 Feature Compression

After feature extraction, the statistical features computed from windows (as opposed to the delineation and MRECCC features) represent each patient as a matrix of size $W \times F$, where W denotes the number of windows per patient and F the number of statistical features extracted from each window. Each element corresponds to the value of a specific feature computed over a specific window segment.

This window-based strategy allows for the capture of localized, time-varying patterns in the signal that would be lost if features were extracted from the entire signal in a single pass. Extracting a feature from each window reveals not just an average behavior, but also the variability and dynamics of that feature across time. For instance, computing a feature like entropy on the entire signal would provide no insight into how that feature fluctuates or spikes in specific regions of interest [70].

However, the resulting high dimensional feature set introduces significant challenges, such as increased computational complexity, risk of model overfitting, redundancy among correlated features, and reduced generalizability. Classical machine learning algorithms typically perform best when the input data representation is compact, informative, and minimally redundant. High dimensional datasets often lead to the so called "curse of dimensionality," characterized by exponential increases in computational requirements, decreased model interpretability, and susceptibility to noise and irrelevant information [70].

To address these challenges, feature compression methods were applied. These methods not only reduce dimensionality but also preserve essential information about the distribution and variation of each feature across time [70]. The following compression metrics were used:

- **Mean (μ)**: captures the central tendency of the feature values across the analysis period.
- **Standard Deviation (σ)**: quantifies variability or dispersion, capturing the consistency of feature values across time windows or along the entire signal.
- **Minimum (min)**: identifies the lowest observed feature value, highlighting extreme behavior or events in the signal.
- **Maximum (max)**: identifies the highest observed feature value, similarly highlighting significant events or peak occurrences.

Each compression metric was applied to all windows for each feature, resulting in four summary values per patient: mean, minimum, maximum, and standard deviation. This approach preserves key statistical characteristics while substantially reducing feature dimensionality.

3.6 Feature Engineering

Some statistical techniques were applied to the extracted and compressed features. These techniques should be performed exclusively on the training sets and subsequently applied to both the training

and test sets to avoid data leakage. A more detailed description of their application to the data splits will be provided after the presentation of the chosen model validation.

3.6.1 Feature Normalization

The extracted ECG features display considerable variability in both their numerical ranges and measurement units. For example, time-domain features are typically expressed in physical units such as microvolts (μV), millivolts (mV), or seconds (s). In contrast, frequency-domain and entropy-based features are obtained through mathematical transformations, resulting in values that lack direct physical units and are not inherently comparable to one another. This disparity complicates model training, as features with larger numerical values may disproportionately influence learning algorithms, potentially overshadowing the contributions of other relevant variables. To address this issue and ensure that all features contribute more equally to the model, three normalization techniques were explored: Min-Max normalization, mean normalization, and standardization.

3.6.1.1 Min-Max Normalization

Min-Max normalization, also known as rescaling, is a simple yet effective method that linearly transforms a feature's values to fit within a predefined interval, typically the range $[0, 1]$. This is achieved using the equation:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3.20}$$

where x is the original feature value, and $x_{\min}$, $x_{\max}$ are the minimum and maximum values for that feature. Min-Max normalization is highly sensitive to outliers, which can distort scaling for most observations, and requires storing $x_{\min}$ and $x_{\max}$ to ensure consistent transformation of test data.

3.6.1.2 Mean Normalization

A second approach considered is mean normalization, which centers the data by subtracting the mean and scales it using the range, as defined by:

$$x_{\text{norm}} = \frac{x - \mu_x}{x_{\max} - x_{\min}} \tag{3.21}$$

Here, μ_x denotes the feature mean. This normalization centers data around zero while preserving the original range; for uniformly distributed data, normalized values typically fall within $[-0.5, 0.5]$. Mean normalization maintains relative distances between observations and yields an expected value of zero for the transformed feature:

$$\mathbb{E}[x_{\text{norm}}] = 0 \tag{3.22}$$

and ensures that the scaled feature spans a unit range:

$$\text{Range}(x_{\text{norm}}) = 1 \tag{3.23}$$

This property can be useful in methods that rely on mean-centered features but benefit from bounded variance.

3.6.1.3 Standardization

The third and most statistically rigorous approach employed is standardization, also referred to as z-score normalization. This method transforms each feature to have a mean of zero and a standard deviation of one, as expressed by:

$$x_{\text{norm}} = \frac{x - \mu_x}{\sigma_x} \tag{3.24}$$

where μ_x and σ_x denote the mean and standard deviation of the feature, respectively. Standardization is particularly well-suited for features that are approximately normally distributed, and it is commonly assumed by many machine learning algorithms, including principal component analysis (PCA), support vector machines (SVMs), and logistic regression. After standardization, the normalized feature satisfies the following statistical properties:

$$\mathbb{E}[x_{\text{norm}}] = 0 \tag{3.25}$$

$$\text{Var}(x_{\text{norm}}) = 1 \tag{3.26}$$

$$\sigma_{x_{\text{norm}}} = 1 \tag{3.27}$$

This approach is robust to outliers compared to Min-Max scaling and effective across diverse ECG-derived features, including time-domain HRV measures (RMSSD, SDNN, SDANN), spectral power, and higher-order statistics (skewness, kurtosis). All three normalization strategies were implemented and evaluated during preprocessing; final selections were based on performance across feature categories and their impact on model interpretability and classification accuracy.

3.6.2 Feature Selection

Feature selection is a critical step in the development of robust predictive models, particularly given the large number of ECG derived features generated in this study. To mitigate feature redundancy and prevent model degradation caused by irrelevant inputs, which can negatively impact performance, this methodology integrates correlation analysis and statistical significance testing. The goal is to isolate a subset of discriminative features that are most informative for predicting the FRS. To visually assess group differences and data distributions, violin plots were used, as illustrated in Figure A.1.

3.6.2.1 Correlation Analysis

Initially, pairwise correlation analysis was performed to identify redundant features. Instead of Pearson's correlation, which assumes linearity and normality, Spearman's rank correlation coefficient (ρ) was employed due to the non-Gaussian distribution of many ECG features [71].

Spearman's rank correlation is defined as:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{3.28}$$

where d_i is the difference between the ranks of the i-th observation in the two variables being compared, and n is the number of paired observations.

Pairs of features exhibiting correlation $|\rho| > 0.9$ were considered redundant. For each such pair, the feature with the higher average absolute correlation with all other features was excluded. This ensures that the retained feature contributes more uniquely to the overall feature space. Figure A.3 presents the correlation matrix of all features.

3.6.2.2 Normality Assessment

After reducing redundancy through correlation analysis, the next step involved assessing the normality of each remaining feature using the Shapiro-Wilk test [72]. This evaluation is needed to guide the future selection of the statistical test used. The Shapiro-Wilk test is defined as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.29}$$

where $x_{(i)}$ denotes ordered sample values, a_i are constants derived from covariances of order statistics, and $\bar{x}$ is the sample mean. The null hypothesis for this test assumes normality, and rejection indicates a significant deviation from a normal distribution. None of the features followed a normal distribution ($p < 0.05$), justifying the use of non-parametric methods for subsequent analyses. A visual assessment of the normality of some features is presented in Figure A.2.

3.6.2.3 Statistical Significance Testing

Given the lack of normality, the statistical significance across different FRS categories was evaluated using the Kruskal-Wallis test, which is suitable for comparing non-normal distributions [73]. The Kruskal-Wallis test is calculated as:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1) \tag{3.30}$$

where R_j is the sum of ranks for the j^{th} group, n_j is the sample size of group j, k is the number of groups, and N is the total sample size.

3.6.2.4 Multiple Comparisons Correction

To rigorously control for multiple comparisons, a necessity given the large number of features tested and the increased likelihood of observing statistically significant results merely by chance, a Bonferroni correction was applied. This adjustment counters the effects of random variation and reduces the risk of false positive discoveries [74]. Given our initial set of 102 candidate features, the adjusted significance threshold ($\alpha_{\text{corrected}}$) was computed as:

$$\alpha_{\text{corrected}} = \frac{\alpha}{m} = \frac{0.05}{102} \approx 0.00049 \tag{3.31}$$

where $\alpha = 0.05$ represents the original significance level, and m is the total number of features under investigation. Figure 3.5 shows the 18 features that passed this stringent significance criterion and were retained for final model training and validation.

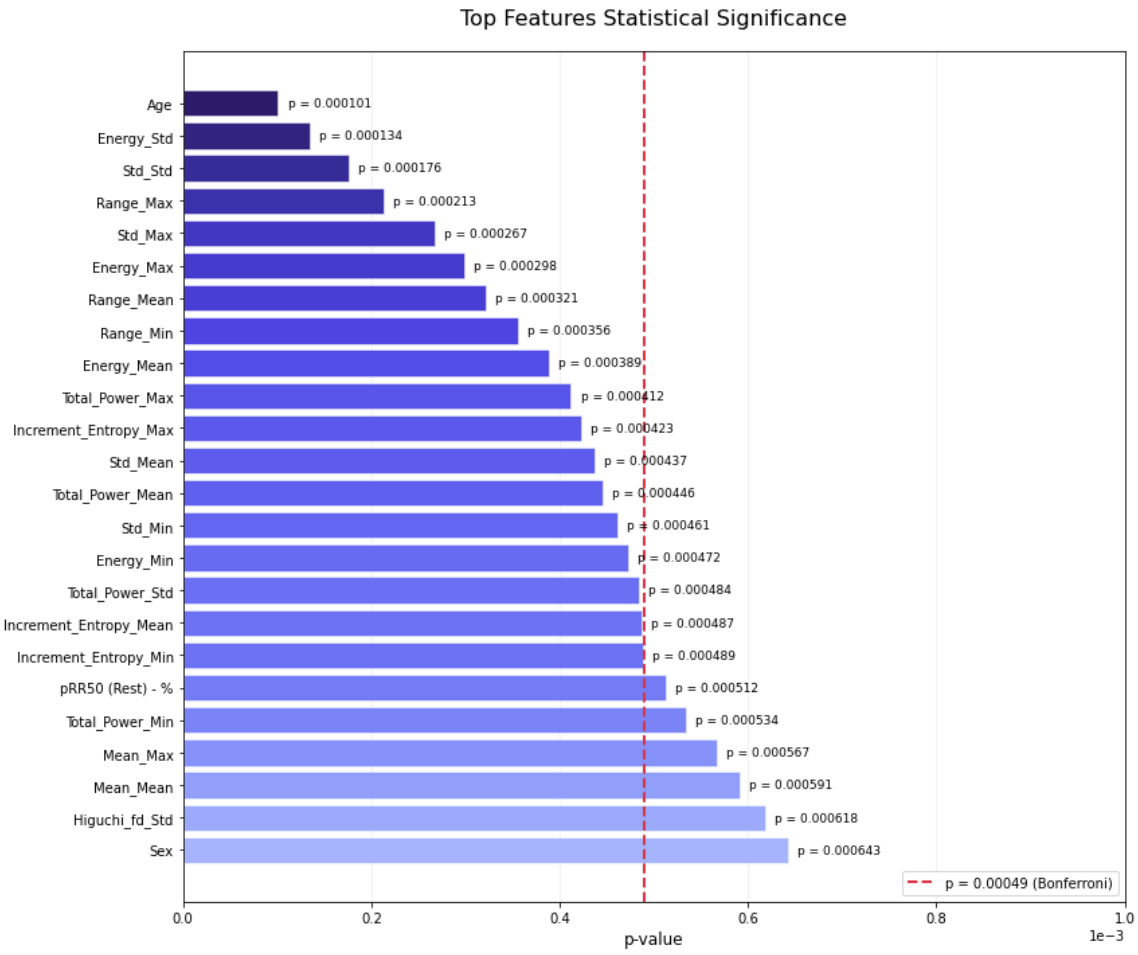


Figure 3.5: Top Features P-values with Bonferroni Correction Threshold

3.6.3 Feature Combination

Due to the large dimensionality of the ECG-derived feature sets, a computationally efficient feature selection approach was implemented during model training. Rather than exhaustively evaluating all possible combinations of features, an automated selection strategy was employed. This significantly reduces computational complexity from an exponential $O(2^n - 1)$, where n is the number of features, to a manageable linear complexity of $O(k \times n)$, with k representing the number of feature selection methods evaluated [70].

The exhaustive search for optimal feature subsets is computationally prohibitive due to its combinatorial nature, as the number of possible subsets is defined by:

$$\text{Total combinations} = \sum_{i=1}^{n} \binom{n}{i} = 2^n - 1 \tag{3.32}$$

When there are 102 features ($n = 102$), yields approximately:

$$\text{Total combinations} = 2^{102} - 1 \approx 5.07 \times 10^{30} \tag{3.33}$$

To circumvent this, our methodology evaluates each number of features sequentially, once per statistical selection criterion, resulting in:

$$\text{Total evaluations} = k \times n \tag{3.34}$$

When there are 5 selection methods ($k = 5$) and 102 features ($n = 102$), this approach yields only:

$$\text{Total evaluations} = 5 \times 102 = 510 \tag{3.35}$$

Four statistical methods were systematically employed to guide the feature combination during training. Each feature subset size from 1 to n was evaluated once per statistical method. Features identified as more statistically significant and discriminative according to the test were retained.

3.6.3.1 ANOVA F-test (F-classification)

The ANOVA F-test evaluates the discriminative power of a feature based on the ratio between inter-class and intra-class variances [75], defined by:

$$F_j = \frac{\frac{\text{SSB}_j}{k-1}}{\frac{\text{SSW}_j}{N-k}} \tag{3.36}$$

with the between-group and within-group variances computed as:

$$\text{SSB}_j = \sum_{i=1}^{k} n_i(\bar{x}_{ij} - \bar{x}_j)^2, \quad \text{SSW}_j = \sum_{i=1}^{k}\sum_{l=1}^{n_i}(x_{ijl} - \bar{x}_{ij})^2 \tag{3.37}$$

where n_i is the number of samples in class i, $\bar{x}_{ij}$ the mean of feature j for class i, $\bar{x}_j$ the overall mean, and N the total number of samples.

3.6.3.2 Mutual Information

Mutual information quantifies the dependency between a feature X and the target variable Y [76], capturing both linear and nonlinear relationships:

$$I(X;Y) = \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) = H(Y) - H(Y|X) \tag{3.38}$$

where $H(Y)$ is the entropy of the target variable, and $H(Y|X)$ the conditional entropy of Y given X, calculated as:

$$H(Y) = -\sum_{i=1}^{k} p(y_i)\log_2 p(y_i), \quad H(Y|X) = -\sum_x p(x)\sum_{i=1}^{k} p(y_i|x)\log_2 p(y_i|x) \tag{3.39}$$

3.6.3.3 Chi-Squared Test

The Chi-Squared test assesses whether observed class frequencies for each feature significantly deviate from expected frequencies [77], with:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad E_{ij} = \frac{R_i C_j}{N} \tag{3.40}$$

Here, O_{ij} and E_{ij} are the observed and expected frequencies, respectively, R_i and C_j are marginal totals, and N the grand total. Since this test requires non-negative values, features were scaled via Min-Max normalization:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{3.41}$$

3.6.3.4 False Discovery Rate

To manage the risk of false positives due to multiple testing, a variante of the False Discovery Rate (FDR), the Benjamini-Hochberg procedure was implemented [78]. Features are ranked by their p-values ($p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$), and the largest index i satisfying the condition below is identified:

$$p_{(i)} \leq \frac{i}{m}\alpha \tag{3.42}$$

Hypotheses corresponding to $p_{(1)}, \ldots, p_{(i)}$ are then selected. An adaptive significance level α was used:

$$\alpha_{\text{adaptive}} = \min\left(0.2, \max\left(0.05, \frac{i}{50}\right)\right) \tag{3.43}$$

The expected False Discovery Proportion (FDP) is bounded by:

$$E[\text{FDP}] = E\left[\frac{V}{R}\right] \leq \alpha \tag{3.44}$$

where V represents false discoveries and R total discoveries.

3.6.3.5 Family-Wise Error Rate

The Family-Wise Error Rate (FWER) method aims to strictly control Type I error probability across multiple tests:

$$\text{FWER} = P(V \geq 1) \leq \alpha \tag{3.45}$$

The Holm-Bonferroni correction was applied [79], adjusting significance thresholds adaptively:

$$p_{(i)} \leq \frac{\alpha}{m - i + 1}, \quad \alpha_{\text{adaptive}} = \min\left(0.3, \max\left(0.05, \frac{i}{30}\right)\right) \tag{3.46}$$

3.7 Machine Learning

3.7.1 Model Validation

To rigorously evaluate the predictive performance and generalizability of the proposed model, Leave-One-Out Cross-Validation (LOOCV) was employed. LOOCV is an exhaustive validation technique and constitutes a special k-fold cross-validation, in which the number of folds, k, equals the number of observations, n. Consequently, each observation in the dataset is used exactly once as the test set, while the remaining $n-1$ observations comprise the training set.This concept was introduced by Allen, D. M. (1974) applied to linear regression [80], and popularized by Stone, M. (1974) [81].

Formally, given a dataset $D = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the LOOCV procedure involves iterating through n steps, each defined as follows:

$$\text{Training Set } (i) : D_{-i} = D \setminus \{(x_i, y_i)\}, \quad \text{Test Set } (i) : \{(x_i, y_i)\} \tag{3.47}$$

where D_{-i} denotes the dataset excluding observation i.

Following the completion of all n iterations, the LOOCV procedure aggregates the individual predictions to compute chosen performance metrics.

LOOCV was selected as the validation strategy because it represents the state-of-the-art method for evaluating predictive models trained on very small datasets, such as the one used in this study. This approach maximizes the training data, while reducing model bias, as each observation in the dataset is utilized as test data.

In the implementation of the proposed framework, feature normalization and feature combination were nested within the inner loop and applied consistently across the outer loop. However, due to the considerable computational burden, feature selection and hyperparameter optimization could not be nested in the same manner. To reduce the probability of data leakage while still enabling these procedures, a hold-out split of 60-40 was adopted, where these techniques were applied only to the 60% training portion of the data. While this approach can introduce a minimal degree of leakage, we consider it practically negligible. This would be overly restrictive, preventing the use of state-of-the-art ML techniques.

3.7.2 Classification

Different classifiers were considered in order to capture the best perspective on the data, ranging from linear models to more complex nonlinear approaches. Each classifier is characterized by its type, conceptual foundation, and the loss function it optimizes during training.

The objective was to assign each observation to one of three predefined cardiovascular risk categories: low, moderate, or high risk. Table 3.4 summarizes the set of classifiers employed in this work.

Table 3.4: Summary of Classification Algorithms

Classifier	**Type**	**Concept**	**Loss Function**
Adaptive Boosting (AdaBoost)	Ensemble (Boosting)	Sequentially trains weak learners, focusing on misclassified samples [82]	Exponential loss
Bagging Classifier (Bagg)	Ensemble (Bagging)	Aggregates predictions from multiple base estimators trained on bootstrapped samples [83]	Depends on base estimator

Classifier	**Type**	**Concept**	**Loss Function**
Decision Tree Classifier (DeTree)	Tree-based	Recursively partitions feature space to maximize information gain [84]	Gini impurity Entropy
Extra Trees Classifier (ExTree)	Ensemble (Randomized Trees)	Random splits for features and thresholds; variance reduction [85]	Gini impurity Entropy
Gaussian Naive Bayes (GauNB)	Probabilistic (Bayesian)	Assumes feature independence given class; models Gaussian likelihoods [70]	Log-loss, Naive Bayes likelihood
Gaussian Process Classifier (GPC)	Probabilistic (Bayesian)	Uses GP for classification with Laplace approximation for inference [86]	Negative log marginal likelihood
Gradient Boosting (GradBoost)	Ensemble (Boosting)	Iteratively fits residuals to correct previous errors using gradient descent [87]	Deviance (log-loss)
K-Nearest Neighbors (KNN)	Instance-based	Classifies based on majority vote among k nearest samples [88]	No explicit loss non-parametric
Linear Discriminant Analysis (LDA)	Linear (Generative)	Projects data to maximize class separability using shared covariance [89]	Log-likelihood ratio
Linear Support Vector Classification (LSVC)	Margin-based	Finds a linear decision boundary with maximum margin [90]	Hinge loss
Logistic Regression (LogReg)	Linear (Discriminative)	Models log-odds of class membership using a linear function [91]	Log-loss (cross-entropy)

Classifier	Type	Concept	Loss Function
Logistic Regression CV (LogRegCV)	Linear (Discriminative)	Logistic regression with automatic hyperparameter tuning via CV [91]	Log-loss (cross-entropy)
Multi-Layer Perceptron (MLP)	Neural Network	Deep feed-forward neural network using backpropagation [92]	Cross-entropy loss
Quadratic Discriminant Analysis (QDA)	Generative	Like LDA but allows class-specific covariance matrices [93]	Log-likelihood ratio
Random Forest (RF)	Ensemble (Bagging)	Aggregates predictions from multiple random decision trees [94]	Gini impurity Entropy
Stochastic Gradient Descent (SGD)	Linear (Optimization-based)	Fits linear models using iterative stochastic gradient descent	Hinge, log, or modified Huber loss
Support Vector Machines (SVM)	Margin-based	Finds optimal hyperplane using kernel trick to maximize margin [95]	Hinge loss

The classifiers listed in Table 3.4 differ not only in model family and underlying assumptions but also in the functions they optimize during training. Because the choice of loss function governs how classification errors are penalized, it directly shapes decision boundaries, calibration, and robustness to class imbalances. Table 3.5 summarizes the loss functions associated with the models used in this work, providing their formulations for reference.

Table 3.5: Summary of Loss Functions Used in Classification Algorithms

Loss Function	**Equation**
Exponential Loss [82]	$\mathscr{L}_{\exp}(y, f(x)) = \exp(-y \cdot f(x)), \quad y \in \{-1, +1\}$
Gini Impurity [84]	$\text{Gini}(D) = 1 - \sum_{k=1}^{K} p_k^2$
Entropy [61]	$\text{Entropy}(D) = -\sum_{k=1}^{K} p_k \log_2(p_k)$
Log-Loss / Cross-Entropy (Binary) [91]	$\mathscr{L}_{\log}(y, p) = -[y \log(p) + (1-y)\log(1-p)]$
Log-Loss / Cross-Entropy (Multiclass) [96]	$\mathscr{L}_{\log}(y, \mathbf{p}) = -\sum_{k=1}^{K} y_k \log(p_k)$
Hinge Loss [90]	$\mathscr{L}_{\text{hinge}}(y, f(x)) = \max(0, 1 - y \cdot f(x))$
Squared Hinge Loss [97]	$\mathscr{L}_{\text{sqhinge}}(y, f(x)) = \max(0, 1 - y \cdot f(x))^2$
Modified Huber Loss [98]	$\mathscr{L}_{\text{huber}}(y, f(x)) = \begin{cases} \max(0, 1 - y \cdot f(x))^2 & \text{if } y \cdot f(x) \geq -1 \\ -4y \cdot f(x) & \text{if } y \cdot f(x) < -1 \end{cases}$
Deviance Loss [99]	$\mathscr{L}_{\text{deviance}}(y, p) = -2[y \log(p) + (1-y)\log(1-p)]$
Log-Likelihood [100]	$\mathscr{L}_{\log} = -\sum_{i=1}^{n} \log P(y_i \mid x_i)$
Gaussian Process Log-Posterior [101]	$\mathscr{L}_{\text{GP}} = -\log P(y \mid f) - \log P(f \mid X)$

3.7.3 Regression

In addition to classification, regression models were employed to predict the continuous cardiovascular risk percentage point estimates. Unlike categorical assignment into discrete risk groups, these models output a continuous probability of risk occurrence, expressed in percentage points. The theoretical prediction domain full range is from 0 to 100%.

Table 3.6 summarizes the regression models implemented in this work, outlining their type, concept, and loss functions.

Table 3.6: Summary of Regression Algorithms

Regressor	**Type**	**Concept**	**Loss Function**
Linear Regression [102]	Linear (Parametric)	Fits a linear function to minimize squared residuals between predictions and targets	Mean Squared Error (MSE): $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$
Polynomial Regression [103]	Non-linear (Feature Engineering)	Extends linear regression by creating polynomial features up to degree d	Mean Squared Error (MSE): $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$
Ridge Regression [104]	Linear (Regularized)	Linear regression with L2 penalty to prevent overfitting; shrinks coefficients uniformly	MSE + L2 penalty: $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\alpha\sum_{j=1}^{p}\beta_j^2$
LASSO Regression [105]	Linear (Regularized)	Linear regression with L1 penalty; performs feature selection by zeroing coefficients	MSE + L1 penalty: $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\alpha\sum_{j=1}^{p}\lvert\beta_j\rvert$

Regressor	Type	Concept	Loss Function
Elastic Net Regression [106]	Linear (Regularized)	Combines L1 and L2 penalties; balances feature selection and coefficient shrinkage	MSE + L1 + L2 penalty: $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\alpha\rho\sum_{j=1}^{p}\lvert\beta_j\rvert+\alpha(1-\rho)\sum_{j=1}^{p}\beta_j^2$

3.7.4 Hyper-parameter Optimization

Hyperparameter optimization is a step in some machine learning workflows aimed at identifying the best-performing model configurations to enhance performance. Given the variety and complexity of classifiers employed in this study, a structured and rigorous hyperparameter optimization strategy was adopted using Grid Search [107].

Grid Search systematically evaluates classifier performance across predefined, discrete sets of hyperparameter values. The search spaces for each classifier were selected considering the small dataset size and reasonable computational complexity, ensuring an exhaustive yet computationally feasible search. Cross-validation (5-fold) was employed to ensure robust performance estimation [107].

For classifiers benefiting significantly from input feature scaling (SVC, Logistic Regression, SGD, and Multi-layer Perceptron), the optimization was performed within a pipeline including a standard scaler transformation.

3.7.4.1 Hyperparameter Spaces

Carefully defined hyperparameter spaces or grids were developed individually for each classifier, based on literature recommendations and preliminary experiments. For instance, the hyperparameters optimized for the Support Vector Classifier (SVC) included regularization parameter C, kernel type, gamma values, and polynomial degree, while ensemble methods like Random Forest and Gradient Boosting were optimized over parameters such as tree depth, number of estimators, and sampling methods [107].

For example, the hyperparameter grid for the SVC was defined as:

$$C \in \{0.1, 1, 10, 100\}, \tag{3.48}$$

$$\text{kernel} \in \{\text{rbf, poly, sigmoid}\}, \tag{3.49}$$

$$\gamma \in \{\text{scale, auto}, 0.001, 0.01, 0.1, 1\}, \tag{3.50}$$

$$\text{degree} \in \{2,3,4\}. \tag{3.51}$$

Similarly, the hyperparameter search space for Random Forest included:

$$n_{\text{estimators}} \in [10,200], \tag{3.52}$$

$$\text{max\_depth} \in [3,20], \tag{3.53}$$

$$\text{min\_samples\_split} \in [2,20], \tag{3.54}$$

$$\text{min\_samples\_leaf} \in [1,10], \tag{3.55}$$

$$\text{max\_features} \in \{\text{sqrt, log2, None}\}. \tag{3.56}$$

Each hyperparameter search was performed using parallel computation to enhance efficiency, leveraging multiple CPU cores through the `n_jobs=-1` parameter setting.

3.7.4.2 Hyperparameters Evaluation

Model performance during hyperparameter optimization was evaluated using accuracy as the scoring metric due to its interpretability and suitability given the balanced nature of the dataset. Table 3.7 shows the optimal hyperparameters selected by each optimization method and subsequently used for model training and final validation.

Table 3.7: Optimized Hyperparameters for Classification Algorithms

Classifier (Abbreviation)	**Optimized Hyperparameters**
Adaptive Boosting (AdaBoost)	`n_estimators`=100, `learning_rate`=0.5, `algorithm`=SAMME.R
Bagging Classifier (Bagg)	`n_estimators`=50, `max_samples`=0.8, `bootstrap`=True
Decision Tree Classifier (DeTree)	`max_depth`=10, `criterion`=gini, `min_samples_split`=4
Extra Trees Classifier (ExTree)	`n_estimators`=100, `max_depth`=None, `max_features`=sqrt
Gaussian Naive Bayes (GauNB)	Default (no hyperparameters optimized)
Gaussian Process (GauPro)	`max_iter_predict`=200, `warm_start`=True
Gradient Boosting (GradBoost)	`n_estimators`=100, `learning_rate`=0.1, `max_depth`=3
K-Nearest Neighbors (KNN)	`n_neighbors`=7, `weights`=distance, `algorithm`=kd_tree
Linear Discriminant Analysis (LDA)	`solver`=svd, `shrinkage`=auto
Linear Support Vector Classification (LSVC)	`C`=1.0, `loss`=squared_hinge, `penalty`=l2
Logistic Regression (LogReg)	`C`=1.0, `solver`=lbfgs, `penalty`=l2
Logistic Regression CV (LogRegCV)	`cv`=5, `penalty`=l2, `solver`=lbfgs
Multi-Layer Perceptron (MLP)	`hidden_layer_sizes`=(100,50), `activation`=relu, `solver`=adam
One-vs-Rest (OvsR)	Base estimator=LinearSVC (`C`=0.5, `penalty`=l2)
Quadratic Discriminant Analysis (QDA)	Default (no hyperparameters optimized)
Random Forest (RF)	`n_estimators`=200, `max_depth`=10, `max_features`=sqrt
Stochastic Gradient Descent (SGD)	`loss`=modified_huber, `penalty`=elasticnet, `alpha`=0.001
Support Vector Machines (SVM)	`C`=10, `kernel`=rbf, `gamma`=scale

3.7.5 Model Evaluation

The performance of the cardiovascular risk classification models developed in this study was evaluated using standard classification metrics derived from confusion matrices of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values. Given the multi-class nature of the classification problem, both binary and multi-class evaluation methodologies were applied. Cardiovascular risk stages were encoded as 1 for low, 2 for moderate, and 3 for high risk.

3.7.5.1 Evaluation Metrics

The following standard classification metrics were calculated to quantitatively evaluate model performance:

Accuracy [70]: The proportion of correct predictions relative to the total number of predictions made:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.57}$$

Precision [108]: The proportion of true positive predictions relative to all positive predictions made by the model:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3.58}$$

Recall (Sensitivity) [108]: The proportion of actual positive instances that were correctly identified by the model:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3.59}$$

Specificity: The proportion of actual negative instances correctly identified as negative:

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{3.60}$$

F1-Score [109]: The harmonic mean of precision and recall, providing a balanced measure of the model's performance in terms of precision and recall:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.61}$$

3.7.6 Multi-class comparisons

The AllvsAll comparison was calculated by OnevsRest classifications using a macro-averaging strategy. Macro-averaging calculates the desired metric independently for each class and then computes an unweighted average, assigning equal importance to each class irrespective of class distribution imbalance:

$$\text{Macro-Average} = \frac{\text{Metric}_{\text{Class1}} + \text{Metric}_{\text{Class2}} + \text{Metric}_{\text{Class3}}}{3} \tag{3.62}$$

It was concluded that the 2-vs-1&3 comparison was not interpretable due to the ambiguous definition of the positive and negative classes; therefore, we did not report precision and recall metrics when this comparison was present.

3.8 Clinical Application

The methodology adopted for the development of CardioRiskAI followed a structured, iterative process rooted in Human-Centered Design (HCD) principles, and was strengthened through several meetings with physicians, where their feedback directly informed successive refinements of the application. CardioRiskAI is a web-based application designed to support cardiovascular risk assessment through ECG analysis and the calculation of the Framingham risk score, by incorporating the models described in the previous sections. From the beginning, the system's development was continuously refined based on feedback from medical professionals, ensuring both clinical validity and user centered usability. The application was developed in accordance with ISO 9241-210 and ISO 9241-971 guidelines, adheres to WCAG 2.1 AA accessibility standards, and tries to ensure HIPAA compliance.

3.8.1 Visual Design

A custom logo, as shown in Figure 3.6 incorporating elements of cardiovascular anatomy, electric signals, and ML, was created to function as both a brand identifier and a trust marker for users.

Figure 3.6: CardioRiskAI Logo

The interface employed a predominantly gray and blue color scheme, reflecting established norms in clinical software, where high-contrast, light backgrounds facilitate accurate data interpretation and

improve visual clarity. The interface was designed to be simple, with low cognitive load and minimal complexity.

3.8.2 Structure and Features

To illustrate the interaction between users and the system, a UML (Unified Modeling Language) use case diagram was developed. This diagram provides a structured representation of the primary functionalities of CardioRiskAI, highlighting the relationships between physicians and patients and the core modules of the application. Figure 3.7 presents the UML use case diagram for CardioRiskAI.

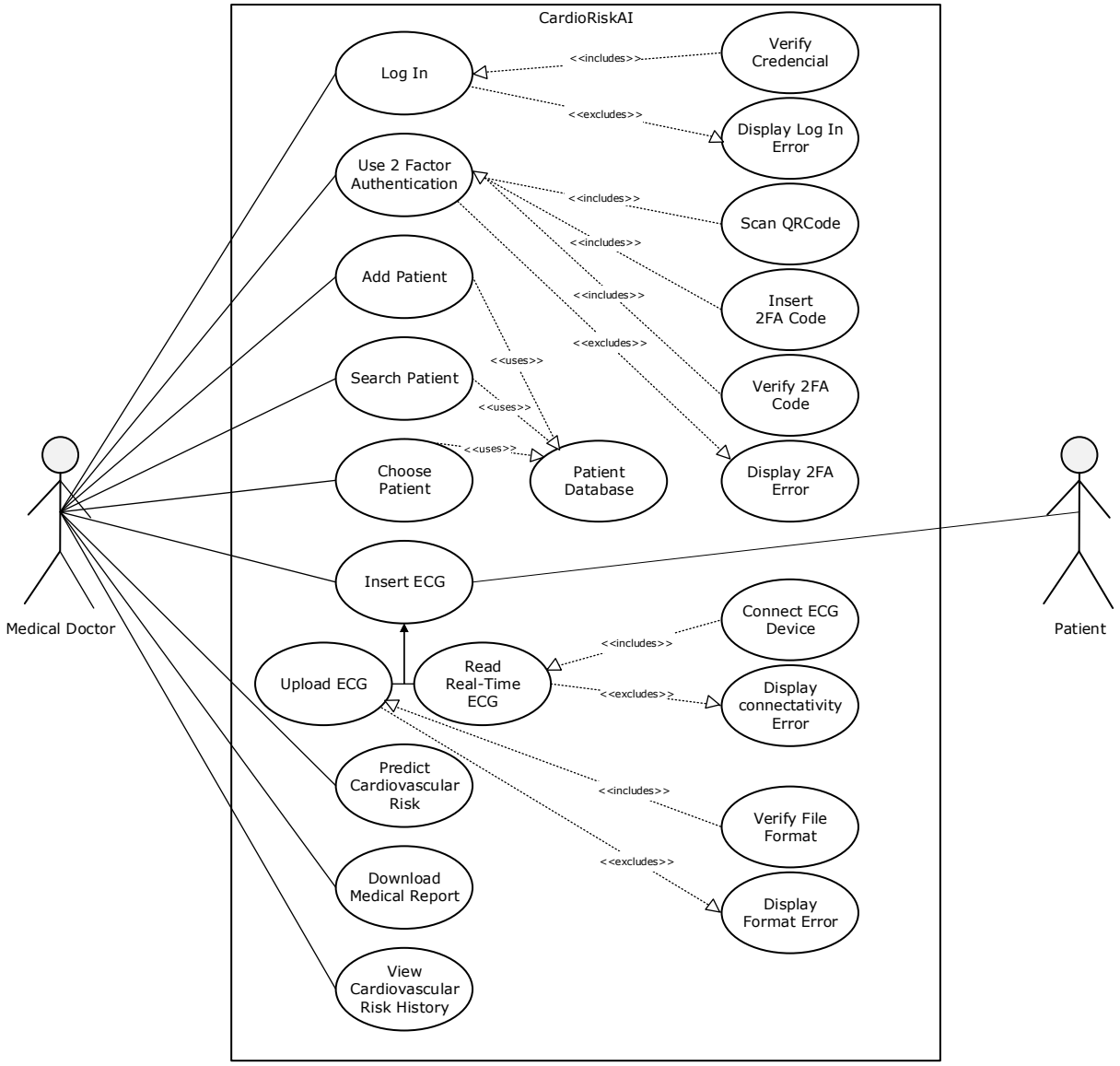


Figure 3.7: CardioRIskAI UML Use Case Diagram

3.8.2.1 Log In and Two Factor Authentication

The application implements a robust three-tier authentication system designed to meet healthcare data security standards and regulatory compliance requirements. The login page presents a professional, clinical aesthetic with the CardioRiskAI branding prominently displayed alongside the institutional logo. A dual-column layout separates branding elements from functional components, creating an intuitive user experience that reinforces the medical nature of the application.

Upon successful username and password verification, users are guided through a comprehensive two-factor authentication (2FA) setup process. The system generates unique TOTP (Time-based One-Time Password) secrets and presents them as scannable QR codes. For accessibility, users may manually enter the secret key into their authenticator applications. Real-time code validation ensures secure account activation, and support is provided for industry-standard applications such as Google Authenticator, Microsoft Authenticator, and Authy. Figure 4.1 shows the log in interface, while Figure 4.2 shows the 2FA interface.

3.8.2.2 Patient Database

The patient database interface serves as the central hub for clinical workflow management. Patient cards display each patient's full name, sex, and age in a scannable format. Users can search for patients using the search bar and add new patients, who are instantly included in the database.

When the intended patient is found, users should click on the "View" button to access the patient's clinical profile. This page is shown in Figure 4.3.

3.8.2.3 Patient Clinical Profile

The patient clinical profile represents the core clinical workspace where cardiovascular risk assessment occurs. A tabbed navigation system structures this interface into three distinct areas: the Risk Prediction tab, the Conditions and Treatments tab, where it displays medical history, and the Documents tab, where clinical documents and reports are kept.

3.8.2.4 Cardiovascular Risk Detection

The risk detection interface provides two distinct operational modes, each tailored to specific clinical scenarios: ECG data upload or real-time ECG data acquisition.

The file upload mode supports workflows in which ECG data is captured using conventional clinical equipment and subsequently analyzed for cardiovascular risk assessment (Figure 4.4). This module accepts multiple file formats (CSV, MAT, DAT), automatically verifies file integrity, and applies preprocessing and signal processing techniques. Once validated, the data is passed to the predictive models, which are stored as serialized weight files in the application's backend. These models are loaded into memory at runtime, not requiring re-training or recompilation.

The real-time mode is faster and more suitable for urgent scenarios, and was integrated and tested with the Plux OpenSignals ECG acquisition device (4.5). Clinicians can adjust sampling rates, configure multi-lead options, and monitor signal quality in real time. This mode enables immediate risk assessment, supporting emergency cardiovascular evaluation. The acquired data is processed and passed through the predictive models in the same way as uploaded data. To ensure consistency and avoid complications with streaming inputs, the analysis is performed only after the full data has been collected. The connection and real-time data acquisition were implemented using the publicly available Python libraries for the Plux OpenSignals device, specifically `pyplux` for device communication and signal streaming, and `pyserial` for managing the Bluetooth serial interface.

Clinicians can select which predictive approach to apply: a regression model or a classification model. Each model is presented alongside its performance metrics, allowing physicians to make an informed choice based on clinical needs. When the regression model is selected, the output is a continuous value expressed as a risk percentage. In contrast, the classification model outputs a discrete risk category: low risk ($<$10%), moderate risk (10–20%), and high risk ($>$20%).

3.8.2.5 Explainable AI

Explainable AI (XAI) is an emerging concept in deep learning, aiming to provide transparency in models that are often perceived as 'black boxes'. In clinical contexts, this lack of interpretability is particularly problematic, as physicians prefer to understand the rationale behind algorithmic predictions rather than relying on opaque outputs. In response to clinician feedback, we designed a customized form of explainability. Specifically, the model was evaluated on each sliding window of the ECG signal, and the windows producing the highest risk score were flagged. These flagged signal segments are then displayed within the application, enabling physicians to visualize which parts of the ECG contributed most strongly to the model's high risk score.

3.8.2.6 Clinical Report

The integrated reporting system transforms risk assessment data into comprehensive clinical reports. Reports include patient information, FRS detected, dedicated annotation spaces for physicians, ECG section flagged by the explainable AI algorithm, and authentication fields for signature and licensing verification. The clinical report is shown in Figures 4.7 and 4.8.

Technically, the system uses `ReportLab` for PDF generation, ensuring professional medical standards and compatibility with electronic health record systems.

3.8.2.7 Risk History Tracking

The application maintains longitudinal patient risk histories, enabling trend analysis and treatment monitoring. Interactive charts display risk evolution over time, with color coded zones to support clinical interpretation.

3.8.3 Privacy and Ethics

Privacy and security were addressed through the implementation of 2FA mechanisms, including the use of TOTP. While these features ensure strong protection against unauthorized access, the clinical deployment of the application would additionally require a formal informed consent process. Such a consent form would outline the scope of data collection, its intended use in cardiovascular risk assessment, the rights of participants, and the safeguards in place to guarantee confidentiality and compliance with applicable regulations. This aligns with the principles of the General Data Protection Regulation (GDPR), which applies across the European Union, including Portugal.

In terms of regulatory compliance, the pathway to clinical implementation in Portugal involves oversight by INFARMED (*Autoridade Nacional do Medicamento e Produtos de Saúde*), the national authority responsible for the regulation of medicines and health products, including medical devices. For the application to be introduced as a clinical decision-support tool, it would need to be classified and certified as a medical device under the European Medical Device Regulation (MDR 2017/745). This process includes: (i) demonstrating conformity with essential safety and performance requirements, (ii) conducting a risk analysis of the software as a medical device, (iii) implementing a quality management system, and (iv) producing technical documentation for evaluation. Once conformity is established, the application may obtain CE marking, which is a prerequisite for clinical use in Portugal.

3.8.4 App Financial Viability

A dedicated financial report was conducted to evaluate the economic feasibility of the proposed application. The financial model considered two main commercialization strategies: an individual package for a single doctor and a group package designed for practices of up to ten doctors. Within this framework, the analysis incorporated projected sales revenues alongside material, service, human resource, and investment costs.

The results of the report indicated that the initial capital required to launch the business was relatively low, making entry into the market accessible. Furthermore, the analysis showed that the percentage of market adoption necessary to reach profitability was realistic and attainable within the identified target segments. These findings support the financial viability of the application and provide a solid foundation for potential scaling in a clinical context.

3.8.5 Market Readiness

Although the financial analysis demonstrated feasibility, the transition to market implementation requires additional steps. Specifically, the application would need a secure cloud-based server pipeline. In parallel, the application must undergo certification as a medical device through INFARMED. Achieving this certification under the European Medical Device Regulation (MDR 2017/745) is a prerequisite for clinical deployment.

CHAPTER 4

Results and Discussion

From the feature selection process, it becomes evident why cardiovascular risk assessment is not typically performed by physicians through direct visual inspection of the ECG signal, unlike the diagnosis of many other cardiac conditions. Our analysis did not reveal statistically significant differences between groups when using traditional morphological wave features, shown in Figure **??**. Instead, the most significant differences emerged from linear and non-linear ECG features. This suggests that the variations relevant for cardiovascular risk assessment are too subtle to be detected by the human eye, and therefore cannot be reliably identified by physicians. These findings corroborate long-standing clinical practice, where risk assessment has traditionally relied on established scoring systems and biomarkers rather than visual ECG interpretation.

Our findings align with the state of the art in confirming that age is by far the feature with the most significant differences between groups, shown in Figure **??**. However, in contrast to existing literature, the linear and non-linear ECG features selected for inclusion in our model exhibited lower p-values than lifestyle and demographic features such as BMI, smoking status, and obesity. While previous studies typically report ECG-derived features as being among the least significant, with lifestyle and demographic variables ranking higher, our results suggest greater discriminative power for ECG-based features within this dataset.

4.1 Classification and Regression Models

Table 4.1 shows clear, graded separability across the three pairwise comparisons. The Low-High task delivers the strongest performance, reaching 97.44% accuracy, 94.44% precision, 100.00% recall, an F1-score of 97.14%, and an AUC of 0.99 with a Random Forest trained on four ANOVA-selected features. The Low-Moderate task is also strong with 91.67% accuracy, 92.31% precision, 85.71% recall, F1-score of 88.89%, and AUC of 0.93 using Bagging with four FDR-selected features. By

Table 4.1: Best Result for the Binary Classification

Comparison	Acc	Pre	Rec	F1	AUC	Classifier	# Feat.	Selector
Low-Mod	91.67	92.31	85.71	88.89	0.93	Bagg	4	FDR
Low-High	97.44	94.44	100.00	97.14	0.99	RF	4	ANOVA
Mod-High	77.42	81.25	76.47	78.79	0.78	DT	6	ANOVA

contrast, Moderate-High is notably harder, with 77.42% accuracy, 81.25% precision, 76.47% recall, F1-score of 78.79%, and AUC of 0.78, the best model here is a Decision Tree using six features chosen by ANOVA.

Taken together, these results reinforce the well-established view that cardiovascular risk lies on a continuum rather than in discrete categories. Risk varies gradually; individuals near decision cutoffs often resemble one another, so the labels "low," "moderate," and "high" reflect thresholds chosen by clinical convention rather than natural discontinuities in physiology. This pattern helps explain why the Low–High comparison achieved stronger performance than analyses that include the Moderate class. Importantly, this does not render regression unnecessary or inferior. When classification cleanly separates groups with very high performance, while regression metrics though not directly comparable are less satisfactory, there is a case that classification may be more clinically actionable for screening and triage. We will revisit this question after all models have been presented.

Taken together, these results reinforce the well-established view that cardiovascular risk lies on a continuum rather than in discrete categories. Risk varies gradually: individuals near the decision cut-offs often resemble one another, so the labels "low", "moderate", and "high" reflect thresholds chosen by clinical convention rather than natural discontinuities in the underlying physiology. This helps explain why the Low-High comparison achieved stronger performance than the comparisons that include the Moderate class. Despite this being a fundamental regression problem, this does not render classification unnecessary or inferior. When classification cleanly separates groups with very high performance, while regression metrics, though not directly comparable, are less satisfactory, there is a case that classification may be more clinically actionable for screening and triage. We will revisit this question and try to draw conclusions about this topic after all models have been presented.

The near-perfect separation between Low and High suggests that individuals at the extremes occupy largely non-overlapping regions of feature space, whereas the drop in performance for Moderate and High implies substantial overlap. Clinically, this suggests that the moderate group is heterogeneous and transitional.

Moreover, a small, parsimonious feature set is highly informative: achieving at least 90% accuracy in two tasks with only four features shows that a compact ECG signal features paired with MRECC is sufficient to classify most patients, which supports deployability, quickness, and interpretability

through reduced data burden and dimentinal space.

Finally, ensembling reduces variance when boundaries are complex: Random Forests and Bagging outperform others on the easier tasks, pointing to more stable decision surfaces, while the Decision Tree's advantage in the Moderate-High task suggests that lower capacity models with stronger regularization may be better suited to that overlap-prone boundary. Regarding feature selection, ANOVA outperformed the other statistical methods in two of the three comparisons, this may suggest that ANOVA-based selection can be a strong choice for binary cardiovascular risk classification.

Table 4.2 presents the multiclass comparisons, excluding precision, recall, and F1-score, as outlined in the methodology. These metrics were omitted due to the lack of interpretability arising from the definition of positive and negative classes in the Mod–Low&High comparison.

Table 4.2: Best Result for the Multiclass Cassification

Comparison	**Accuracy**				**AUC**	**Classifier**	**# Feat.**	**Selector**
	Overall	**Low**	**Mod**	**High**				
Low-Mod&High	96.23	-	-	-	0.96	Bagg	4	MI
Mod-Low&High	79.24	-	-	-	0.80	Bagg	4	FWER
High-Low&Mod	84.91	-	-	-	0.88	Bagg	6	FDR
Low-Mod-High	86.79	96.23	79.24	84.91	0.88	-	-	-

The strongest result appears in the Low-Mod&High configuration with 96.23% accuracy, reflecting that Low cases are easiest to separate when contrasted with the combined Moderate and High. Followed up for the 84.81% accuracy of the High-Low&Mod. In the Mod-Low&High setting the accuracy drops to 79.24% despite strong Low and High figures, underscoring how the Moderate degrades performance.

The Low–Mod–High (All-vs-All) comparison represents the most clinically relevant scenario. Using macro-averaging, this analysis achieved an accuracy of 86.79% and an AUC of 0.88. These results indicate that the model can reliably differentiate between the groups, in a manner consistent with both research standards and clinical acceptance.

Bagging dominated performance across the three One-vs-Rest comparisons. This supports the conclusion that, in this case, Bagging classifiers achieved superior results in the multiclass setting.

Similar to the binary comparisons, the models in the multiclass setting relied on between four and six features. Notably, four features were sufficient for the Low–Mod&High comparison, highlighting

both parsimony and potential deployability. However, no definitive conclusions can be drawn regarding feature selector preference.

The multiclass results reinforce a continuum of cardiovascular risk: extremes are separable, the Moderate boundary is the challenge, and small, carefully chosen feature sets, especially when paired with ensembling, can deliver strong accuracy and discrimination.

Table 4.3 presents the best results for all comparisons using regression. The metrics were normalized back to the original scale of the target variable, such that RMSE and MAE correspond to the true interval errors of the model. As previously noted, the theoretical range of CVD risk spans from 0 to 100%. In practice, however, it is uncommon for individuals to exceed a predicted risk of 60%.

Table 4.3: Best results for all comparisons using regression

Comparison	**RMSE**	**MSE**	**MAE**	R^2	**MAPE**	**Regressor**	**# Feat.**	**Selector**
Low–Mod	5.12	26.21	3.89	0.66	0.57	LASSO	12	FWER
Low–High	3.85	14.82	2.75	0.78	0.42	Ridge	7	ANOVA
Mod–High	4.47	19.98	3.21	0.72	0.49	LASSO	8	MI
Low–Mod&High	4.03	16.24	2.94	0.76	0.44	Elastic	9	ANOVA
Mod–Low&High	6.02	36.24	4.48	0.58	0.63	Poly	9	ANOVA
High–Low&Mod	4.65	21.62	3.40	0.69	0.52	Poly	7	FDR
Low–Mod–High	5.38	28.94	4.05	0.64	0.59	Poly	10	CS

The best performing comparison was once again the Low-High, achieving an RMSE of 3.85 and a MAE of 2.75. In contrast, the Mod–Low&High comparison yielded the weakest performance, with an RMSE of 6.02 and a MAE of 4.48. Notably, the best and worst comparisons were consistent with those observed in the classification task.

However, in the classification task the Mod-High comparison performed worse than the Low–Mod case, this pattern was not observed in regression. Specifically, Low-Mod yielded an RMSE of 5.12, while Mod-High achieved a lower RMSE of 4.47.

Polynomial regression appears repeatedly among the multiclass comparisons, indicating that nonlinear modeling helps capture structure within the data. In contrast, LASSO achieved stronger performance in the binary comparisons.

The regressors relied on between seven and twelve predictors, with no monotonic relationship observed between feature count and accuracy. In comparison with classification, regression appeared to require higher dimensionality to achieve stable performance, suggesting that capturing continuous risk estimates demands a broader set of input features.

4.2 Clinical Application

The following figures present the results of the graphical interface developed for CardioRiskAI. This interface was iteratively refined through feedback meetings with physicians, which served both to validate the implemented features and to adjust the placement of user interface elements.

As illustrated in Figure 4.1, CardioRiskAI begins with the Log In screen. The user credentials are first verified, and upon successful authentication the system redirects to the 2FA screen (Figure 4.2). At this stage, a QR code is scanned with an authenticator application (Google Authenticator, Microsoft Authenticator, or Authy). The authenticator generates a six-digit TOTP only if the account is correctly linked to the user's device. Once the TOTP is entered and verified, access is granted to the patient database screen, shown in Figure 4.3.

Within the patient database, the user can either search for an existing patient or add a new record to the system. Once the desired patient has been located, selecting the *View* button to the right of the entry opens the corresponding patient profile, as illustrated in Figure 4.4.

The default screen on the patients profile is the CVD risk assessment through ECG upload. There is a drag and drop space as well as a button to browse the files on the users computer. However, you can choose to click on the real-time ECG where you can make the CVD assessment using a Bluethooth ECG device, needed UI elements to connect and start and stop recording are implemented (Figure 4.5). As a prototype the validation of these technique was done using the Plux OpenSignals ECG acquisition device.

Using either of the methods after the upload or recording the app automatically starts assessing the CVD risk score by applying all the techniques in the methodology and calling the pretrained models mentioned above. The average time to compute is 5 seconds. The risk assessment result screen is shown in Figure 4.6.

Beneath the result screen shown in Figure 4.6, a *Download Medical Report* button is available. This feature generates an automated report that includes patient information, the risk assessment outcome, a section for physician annotations to allow the inclusion of relevant clinical notes, a plot highlighting the regions flagged by the explainable AI algorithm as higher risk, and signature fields for formal validation. An example of the generated report is presented in Figures 4.7 and 4.8.

Overall, the physicians expressed enthusiasm for the project, describing the application as "simple," "intuitive," and "well done." They also noted that they were "looking forward to using the application in their daily clinical practice." The study was characterized as "very interesting" and "innovative," and the physicians emphasized that research of this kind, particularly studies with a strong clinical

focus or a direct clinical component, is both needed and welcomed within the cardiovascular medical community.

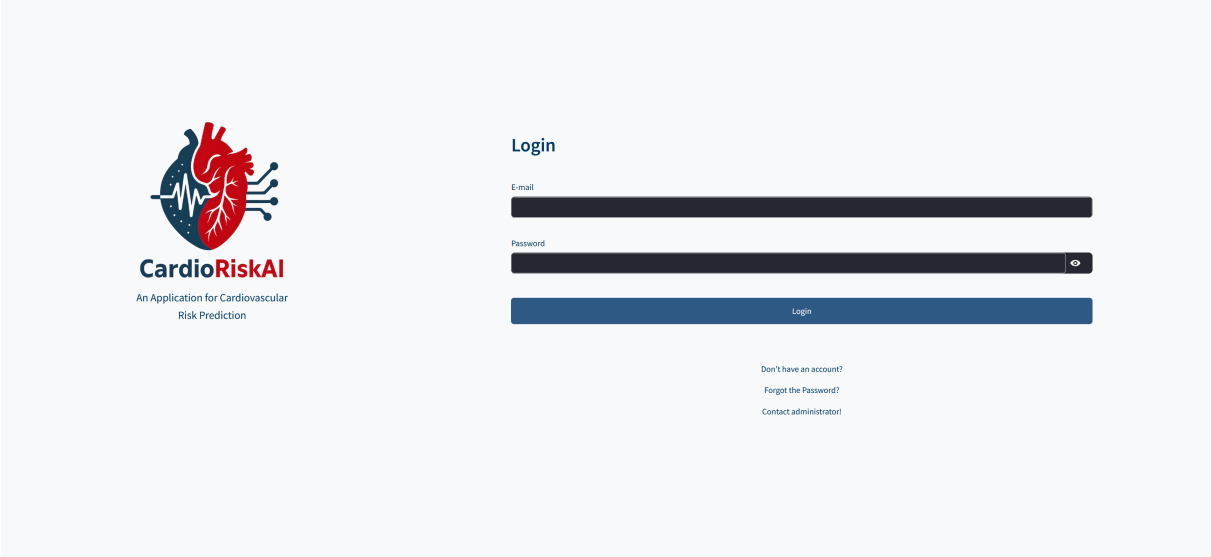


Figure 4.1: Log In Screen

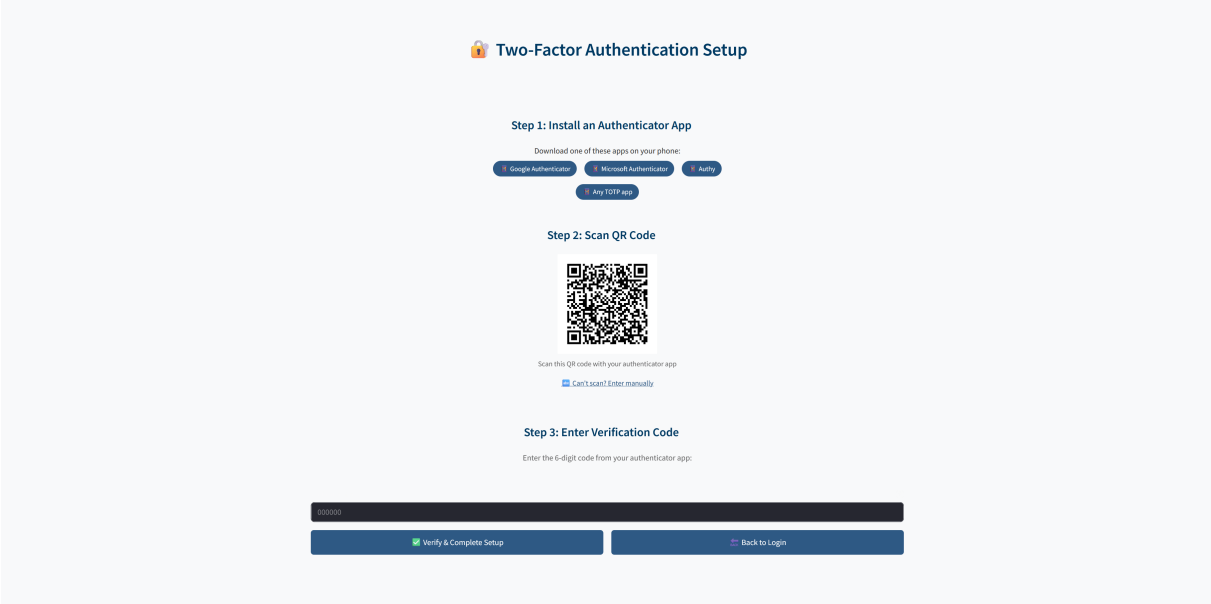


Figure 4.2: Two Factor Authentication Screen

Figure 4.3: Patient Database Screen

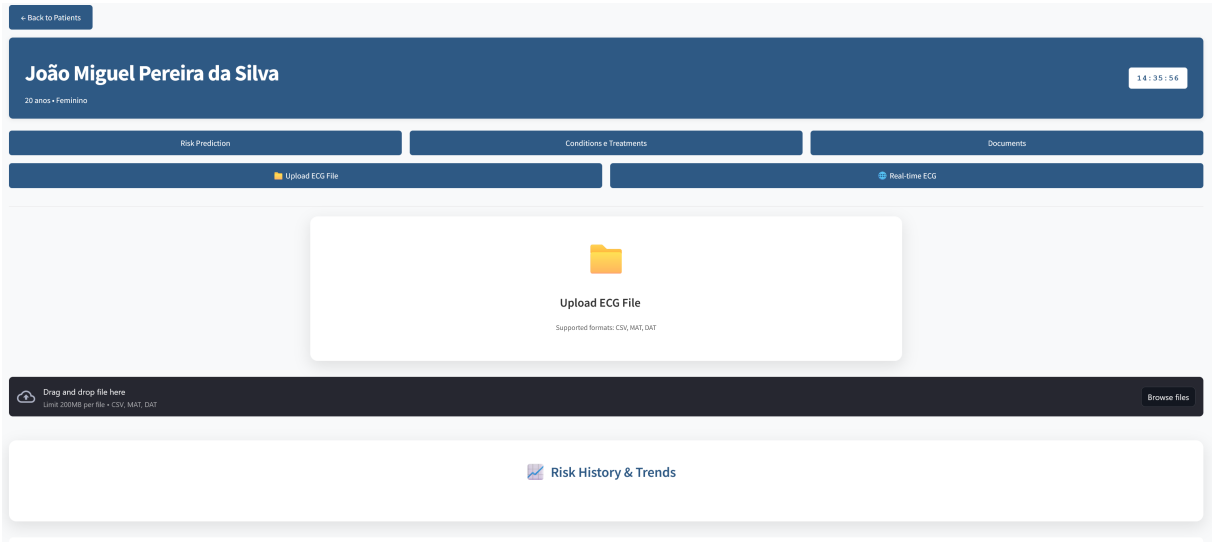


Figure 4.4: ECG Upload Screen

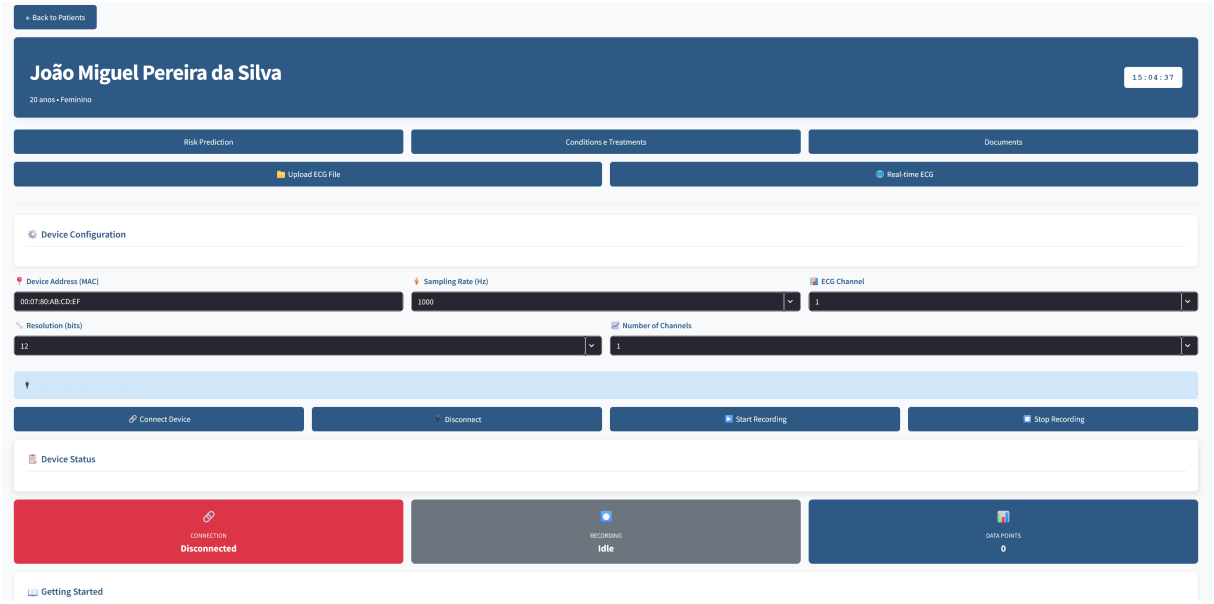


Figure 4.5: ECG Real Time Reading Screen

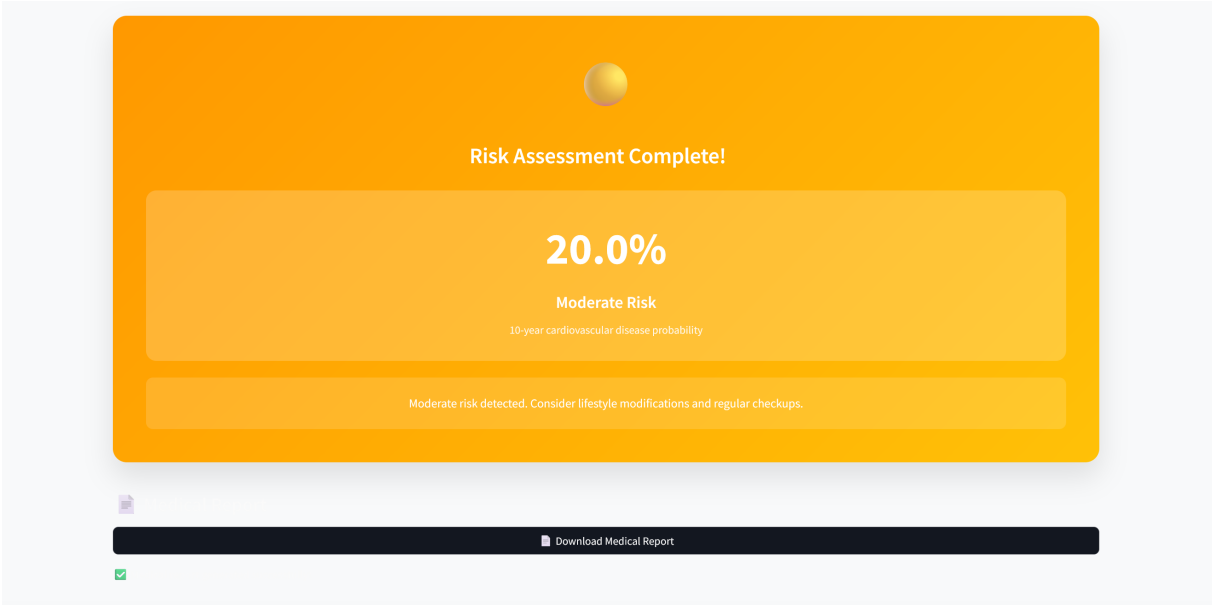
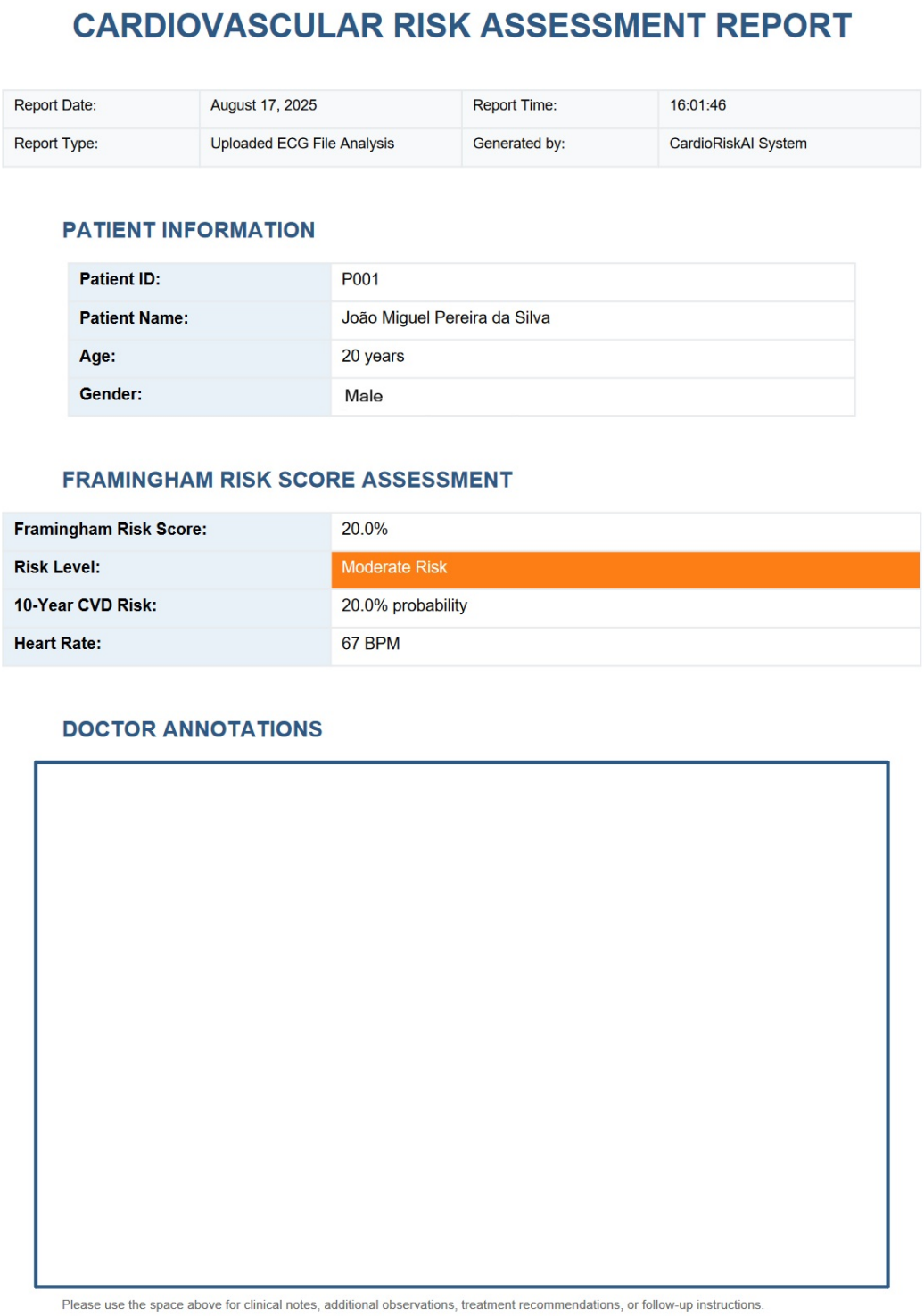


Figure 4.6: Cardiovascular Risk Assessment Screen



CARDIOVASCULAR RISK ASSESSMENT REPORT

Report Date:	August 17, 2025	Report Time:	16:01:46
Report Type:	Uploaded ECG File Analysis	Generated by:	CardioRiskAI System

PATIENT INFORMATION

Patient ID:	P001
Patient Name:	João Miguel Pereira da Silva
Age:	20 years
Gender:	Male

FRAMINGHAM RISK SCORE ASSESSMENT

Framingham Risk Score:	20.0%
Risk Level:	Moderate Risk
10-Year CVD Risk:	20.0% probability
Heart Rate:	67 BPM

DOCTOR ANNOTATIONS

Please use the space above for clinical notes, additional observations, treatment recommendations, or follow-up instructions.

Figure 4.7: First Page of the Clinical Report

ECG ANALYSIS

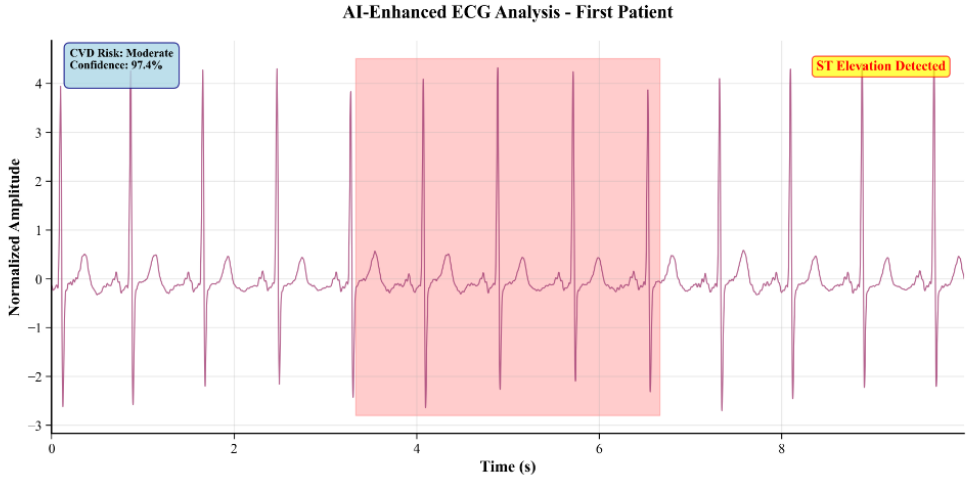


Figure 1: ECG signal analysis showing cardiac rhythm and morphology patterns used for risk assessment.

PHYSICIAN SIGNATURE

Physician Name: _____ Date: _____

Signature: _____ License: _____

This report was generated by CardioRiskAI - Cardiovascular Risk Prediction System
Generated on August 17, 2025 at 16:01:46
For clinical use only. Please consult with a healthcare professional for medical decisions.

Figure 4.8: Second Page of the Clinical Report

CHAPTER 5

Conclusion and Future Work

Cardiovascular diseases (CVDs) remain the leading global cause of mortality and morbidity. Risk assessment in clinical practice is typically performed using conventional cardiovascular risk scores, such as the Framingham model, which rely on a combination of blood biomarkers, clinical measurements, and lifestyle factors. While widely adopted, these tools require multiple laboratory tests, may delay timely risk assessment, and are resource intensive, particularly in low-resource settings. In contrast, electrocardiography (ECG) is becoming inexpensive, portable, and widely accessible, yet has historically been regarded as insufficiently discriminative for cardiovascular risk assessment.

This dissertation investigated whether carefully engineered ECG features, combined with a minimal, rapid, and easily communicated clinical context (MRECCC: age, body mass index, smoking status, and diabetes status), could enable accurate and interpretable assessment of the Framingham Risk Score. To strengthen the clinical component, an application named CardioRiskAI was also developed, providing a simple graphical interface through which physicians can apply the proposed risk assessment models.

Results demonstrate that binary classification achieved excellent discrimination between low and high risk groups, with an accuracy of 97.4% and an AUC of 0.99. Low-moderate comparisons also yielded high performance, whereas moderate-high comparisons remained challenging, reflecting the transitional and heterogeneous nature of the moderate risk group. Multiclass classification exhibited similar patterns, with the Low-Moderate-High (all-vs-all) configuration, the most clinically relevant, achieving an overall accuracy of 86.79%. Regression analysis produced RMSE of between 4 to 6 percentage points, with pratical intervals concentrated within 0-60%.

These findings suggest that a compact, interpretable set of ECG-derived features, when combined with MRECCC, can provide rapid and cost-effective cardiovascular risk assessment. Such an approach has the potential to democratize risk assessment in resource-limited settings, and improve urgent care risk assessment speed in all setting. Finally, if these models can be shown to generalize effectively,

they raise the question of whether extensive blood laboratory testing remains necessary, given that comparable risk assessment may be achievable without it.

The clinical application received highly positive feedback from physicians, who emphasized that they were "looking forward to using the application in their daily clinical practice."

Future methodological improvements should focus on systematically evaluating ECG signal delineation using error metrics and accuracy measures, particularly by verifying whether wave boundaries are detected within clinically acceptable intervals. This validation step is essential to confirm whether the finding that ECG derived features were not statistically significant is truly robust. The explainable AI algorithm may also be improved, although such developments often constitute a separate line of research and thus could not be explored in depth within the scope of this dissertation.

Future work should additionally examine model generalization and incorporate risk scores from diverse demographic groups as reference values. These methodologies should be tested on large publicly available datasets containing the necessary clinical and ECG information, such as the UK Biobank and the Taiwan Biobank.

APPENDIX A

Appendix

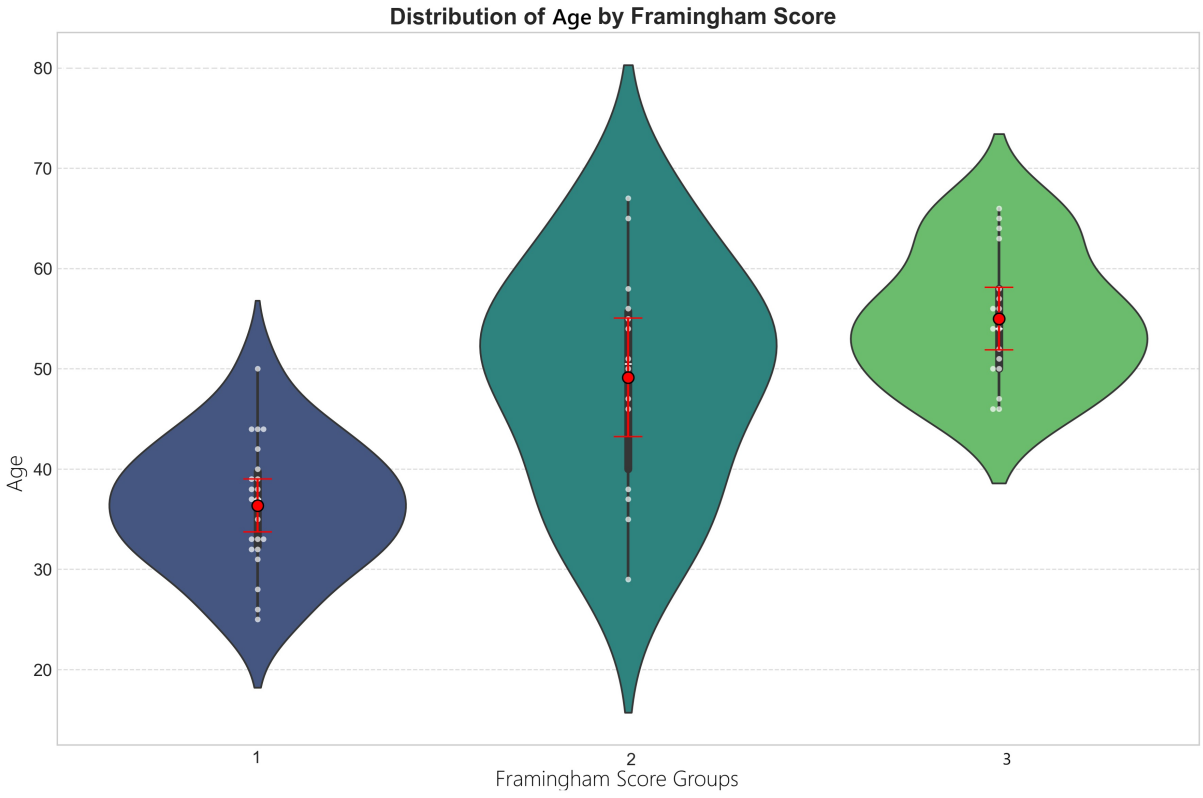


Figure A.1: Age Feature Violin Plot

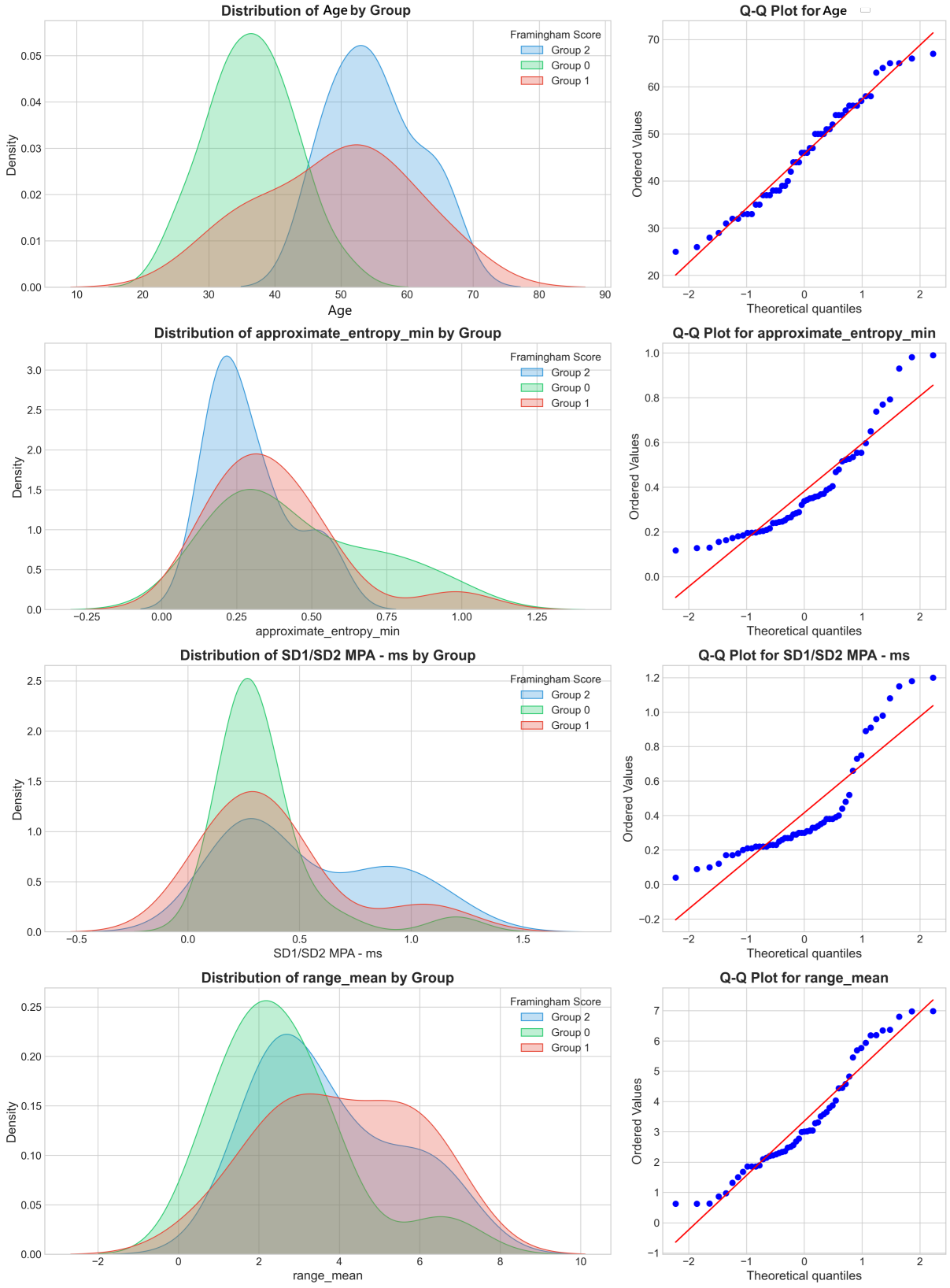


Figure A.2: Normality Visualization by Feature Distributions per Group

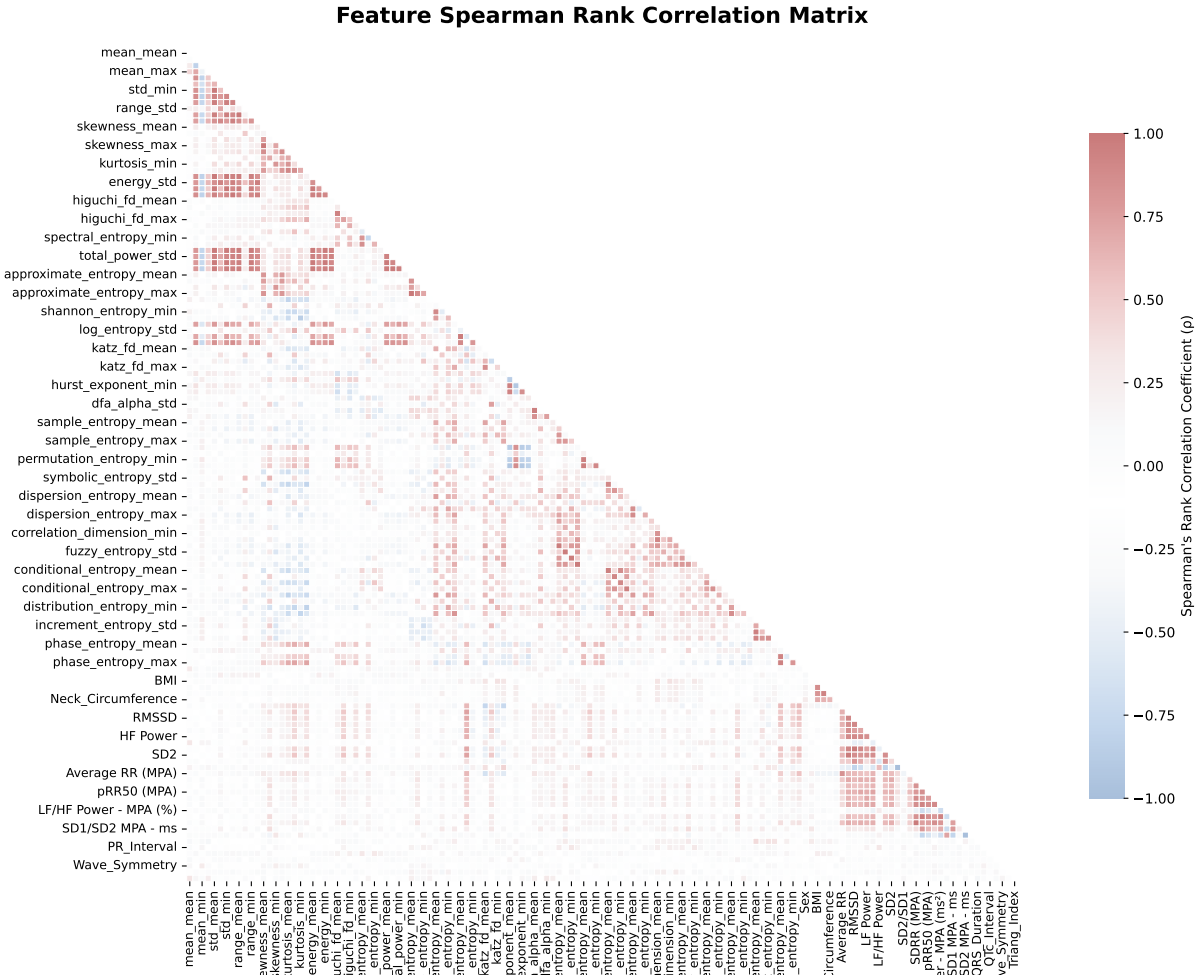


Figure A.3: Feature Correlation Matrix

Bibliography

[1] M. Di Cesare, P. Perel, S. Taylor, C. Kabudula, et al. "The Heart of the World". In: *Global Heart* 19.1 (2024). ISSN: 2211-8179. DOI: 10.5334/gh.1288.

[2] H. Ritchie, L. Rodés-Guirao, E. Mathieu, M. Gerber, E. Ortiz-Ospina, J. Hasell, and M. Roser. "Population Growth". In: *Our World in Data* (2023). https://ourworldindata.org/population-growth.

[3] G. A. Mensah, V. Fuster, C. J. Murray, G. A. Roth, et al. "Global Burden of Cardiovascular Diseases and Risks, 1990-2022". In: *Journal of the American College of Cardiology* 82.25 (Dec. 2023), 2350–2473. ISSN: 0735-1097. DOI: 10.1016/j.jacc.2023.11.007.

[4] I. M. Graham. "The importance of total cardiovascular risk assessment in clinical practice". In: *European Journal of General Practice* 12.4 (Jan. 2006), 148–155. ISSN: 1751-1402. DOI: 10.1080/13814780600976282.

[5] P. W. F. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel. "Prediction of Coronary Heart Disease Using Risk Factor Categories". In: *Circulation* 97.18 (May 1998), 1837–1847. ISSN: 1524-4539. DOI: 10.1161/01.cir.97.18.1837.

[6] S. Sajeev, S. Champion, A. Beleigoli, D. Chew, et al. "Predicting Australian Adults at High Risk of Cardiovascular Disease Mortality Using Standard Risk Factors and Machine Learning". In: *International Journal of Environmental Research and Public Health* 18.6 (Mar. 2021), p. 3187. ISSN: 1660-4601. DOI: 10.3390/ijerph18063187.

[7] C.-H. Cheng, B.-J. Lee, O. N. Nfor, C.-H. Hsiao, Y.-C. Huang, and Y.-P. Liaw. "Using machine learning-based algorithms to construct cardiovascular risk prediction models for Taiwanese adults based on traditional and novel risk factors". In: *BMC Medical Informatics and Decision Making* 24.1 (July 2024). ISSN: 1472-6947. DOI: 10.1186/s12911-024-02603-2.

[8] P. Unnikrishnan, D. K. Kumar, S. Poosapadi Arjunan, H. Kumar, P. Mitchell, and R. Kawasaki. "Development of Health Parameter Model for Risk Prediction of CVD Using SVM". In: *Computational and Mathematical Methods in Medicine* 2016 (2016), 1–7. ISSN: 1748-6718. DOI: 10.1155/2016/3016245.

[9] I. Colombet, A. Ruelland, G. Chatellier, F. Gueyffier, P. Degoulet, and M.-C. Jaulent. "Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression". In: *Proceedings of the AMIA Symposium*. 2000, p. 156.

[10] W. Dong, E. Y. F. Wan, D. Y. T. Fong, K. C. Tan, W. W. Tsui, E. M. Hui, K. H. Chan, C. S. C. Fung, and C. L. K. Lam. "Development and validation of 10-year risk prediction models of cardiovascular disease in Chinese type 2 diabetes mellitus patients in primary care using interpretable machine learning-based methods". In: *Diabetes, Obesity and Metabolism* 26.9 (July 2024), 3969–3987. ISSN: 1463-1326. DOI: 10.1111/dom.15745.

[11] C. Li, X. Liu, P. Shen, Y. Sun, T. Zhou, W. Chen, Q. Chen, H. Lin, X. Tang, and P. Gao. "Improving cardiovascular risk prediction through machine learning modelling of irregularly repeated electronic health records". In: *European Heart Journal - Digital Health* 5.1 (Oct. 2023), 30–40. ISSN: 2634-3916. DOI: 10.1093/ehjdh/ztad058.

[12] A. Kesar, A. Baluch, O. Barber, H. Hoffmann, M. Jovanovic, D. Renz, B. L. Stopak, P. Wicks, and S. Gilbert. "Actionable absolute risk prediction of atherosclerotic cardiovascular disease based on the UK Biobank". In: *PLOS ONE* 17.2 (Feb. 2022). Ed. by T. R. Gadekallu, e0263940. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0263940.

[13] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, and M. van der Schaar. "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423, 604 UK Biobank participants". In: *PLOS ONE* 14.5 (May 2019). Ed. by K. Aalto-Setala, e0213653. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0213653.

[14] R. Nakanishi, P. J. Slomka, R. Rios, J. Betancur, et al. "Machine Learning Adds to Clinical and CAC Assessments in Predicting 10-Year CHD and CVD Deaths". In: *JACC: Cardiovascular Imaging* 14.3 (Mar. 2021), 615–625. ISSN: 1936-878X. DOI: 10.1016/j.jcmg.2020.08.024.

[15] J. You, Y. Guo, J.-J. Kang, H.-F. Wang, M. Yang, J.-F. Feng, J.-T. Yu, and W. Cheng. "Development of machine learning-based models to predict 10-year risk of cardiovascular disease: a prospective cohort study". In: *Stroke and Vascular Neurology* 8.6 (Apr. 2023), 475–485. ISSN: 2059-8696. DOI: 10.1136/svn-2023-002332.

[16] B. N. Krupa, K. Bharathi, M. Gaonkar, S. Karun, S. Nath, and M. A. M. Ali. "Multiclass Classification of APG Signals using ELM for CVD Risk Identification: A Real-Time Application". In: *The 16th International Conference on Biomedical Engineering*. Springer Singapore, 2017, 32–37. ISBN: 9789811042201. DOI: 10.1007/978-981-10-4220-1_7.

[17] T. Jeon, B. Kim, M. Jeon, and B.-G. Lee. "Implementation of a portable device for real-time ECG signal analysis". In: *BioMedical Engineering OnLine* 13.1 (Dec. 2014). ISSN: 1475-925X. DOI: 10.1186/1475-925x-13-160.

[18] H. Ozkan, O. Ozhan, Y. Karadana, M. Gulcu, S. Macit, and F. Husain. "A Portable Wearable Tele-ECG Monitoring System". In: *IEEE Transactions on Instrumentation and Measurement* 69.1 (Jan. 2020), 173–182. ISSN: 1557-9662. DOI: 10.1109/tim.2019.2895484.

[19] T. R. Dawber, G. F. Meadors, and F. E. Moore. "Epidemiological Approaches to Heart Disease: The Framingham Study". In: *American Journal of Public Health and the Nations Health* 41.3 (Mar. 1951), 279–286. ISSN: 0002-9572. DOI: 10.2105/ajph.41.3.279.

[20] T. R. Dawber and W. B. Kannel. "The Framingham Study An Epidemiological Approach to Coronary Heart Disease". In: *Circulation* 34.4 (Oct. 1966), 553–555. ISSN: 1524-4539. DOI: 10.1161/01.cir.34.4.553.

[21] W. B. Kannel, T. R. Dawber, A. Kagan, N. Revotskie, and J. Stokes. "Factors of Risk in the Development of Coronary Heart Disease—Six-Year Follow-up Experience: The Framingham Study". In: *Annals of Internal Medicine* 55.1 (July 1961), 33–50. ISSN: 1539-3704. DOI: 10.7326/0003-4819-55-1-33.

[22] M. Majidi, V. Eslami, P. Ghorbani, and M. Foroughi. "Are women more susceptible to ischemic heart disease compared to men? A literature overview". In: *Journal of geriatric cardiology: JGC* 18.4 (2021), p. 289.

[23] V. Saini, L. Guada, and D. R. Yavagal. "Global Epidemiology of Stroke and Access to Acute Ischemic Stroke Interventions". In: *Neurology* 97.20$_{supplement_2}$ (Nov. 2021). ISSN: 1526-632X. DOI: 10.1212/wnl.0000000000012781.

[24] M. A. Ikram, R. G. Wieberdink, and P. J. Koudstaal. "International Epidemiology of Intracerebral Hemorrhage". In: *Current Atherosclerosis Reports* 14.4 (Apr. 2012), 300–306. ISSN: 1534-6242. DOI: 10.1007/s11883-012-0252-1.

[25] J. A. Diamond and R. A. Phillips. "Hypertensive Heart Disease". In: *Hypertension Research* 28.3 (2005), 191–202. ISSN: 1348-4214. DOI: 10.1291/hypres.28.191.

[26] B. Liu, X.-y. Huang, Z.-y. Hao, J. Wang, Y.-t. Fan, Q. Shao, P.-z. Miao, R.-g. Li, B. He, and L. Jiang. "Burden of Atrial Fibrillation and Flutter and its Attributable Risk Factors in 204 Countries and Territories, 1990 -2021: Results from the Global Burden of Disease Study 2021". In: (2024). DOI: 10.2139/ssrn.4939225.

[27] M. J. Antunes. "The Global Burden of Rheumatic Heart Disease: Population-Related Differences (It is Not All the Same!)" In: *Brazilian Journal of Cardiovascular Surgery* 35.6 (2020). ISSN: 1678-9741. DOI: 10.21470/1678-9741-2020-0514.

[28] P. Song, Y. He, D. Adeloye, Y. Zhu, X. Ye, Q. Yi, K. Rahimi, and I. Rudan. "The Global and Regional Prevalence of Abdominal Aortic Aneurysms: A Systematic Review and Modeling Analysis". In: *Annals of Surgery* 277.6 (Sept. 2022), 912–919. ISSN: 0003-4932. DOI: 10.1097/sla.0000000000005716.

[29] J. B. Herrick. "CLINICAL FEATURES OF SUDDEN OBSTRUCTION OF THE CORONARY ARTERIES". In: *Journal of the American Medical Association* LIX.23 (Dec. 1912), p. 2015. ISSN: 0002-9955. DOI: 10.1001/jama.1912.04270120001001.

[30] M. A. DeWood, J. Spores, R. Notske, L. T. Mouser, R. Burroughs, M. S. Golden, and H. T. Lang. "Prevalence of Total Coronary Occlusion during the Early Hours of Transmural Myocardial Infarction". In: *New England Journal of Medicine* 303.16 (Oct. 1980), 897–902. ISSN: 1533-4406. DOI: 10.1056/nejm198010163031601.

[31] M. S. Brown and J. L. Goldstein. "Receptor-mediated endocytosis: insights from the lipoprotein receptor system." In: *Proceedings of the National Academy of Sciences* 76.7 (July 1979), 3330–3337. ISSN: 1091-6490. DOI: 10.1073/pnas.76.7.3330.

[32] F. C. Novello and J. M. Sprague. "BENZOTHIADIAZINE DIOXIDES AS NOVEL DIURETICS". In: *Journal of the American Chemical Society* 79.8 (Apr. 1957), 2028–2029. ISSN: 1520-5126. DOI: 10.1021/ja01565a079.

[33] R. Ross and J. A. Glomset. "The Pathogenesis of Atherosclerosis: (First of Two Parts)". In: *New England Journal of Medicine* 295.7 (Aug. 1976), 369–377. ISSN: 1533-4406. DOI: 10.1056/nejm197608122950707.

[34] F. H. Epstein, D. Steinberg, S. Parthasarathy, T. E. Carew, J. C. Khoo, and J. L. Witztum. "Beyond Cholesterol". In: *New England Journal of Medicine* 320.14 (Apr. 1989), 915–924. ISSN: 1533-4406. DOI: 10.1056/nejm198904063201407.

[35] J. L. Goldstein, Y. K. Ho, S. K. Basu, and M. S. Brown. "Binding site on macrophages that mediates uptake and degradation of acetylated low density lipoprotein, producing massive cholesterol deposition". In: *Proceedings of the National Academy of Sciences* 76.1 (Jan. 1979), 333–337. ISSN: 1091-6490. DOI: 10.1073/pnas.76.1.333.

[36] S. Chien. "Mechanotransduction and endothelial cell homeostasis: the wisdom of the cell". In: *American Journal of Physiology-Heart and Circulatory Physiology* 292.3 (Mar. 2007), H1209–H1224. ISSN: 1522-1539. DOI: 10.1152/ajpheart.01047.2006.

[37] W. A. Pryor and K. Stone. "Oxidants in Cigarette Smoke Radicals, Hydrogen Peroxide, Peroxynitrate, and Peroxynitritea". In: *Annals of the New York Academy of Sciences* 686.1 (May 1993), 12–27. ISSN: 1749-6632. DOI: 10.1111/j.1749-6632.1993.tb39148.x.

[38] V. J. Dzau and R Re. "Tissue angiotensin system in cardiovascular medicine. A paradigm shift?" In: *Circulation* 89.1 (Jan. 1994), 493–498. ISSN: 1524-4539. DOI: 10.1161/01.cir.89.1.493.

[39] T. C. T. S. Group. "Effects of Enalapril on Mortality in Severe Congestive Heart Failure". In: *New England Journal of Medicine* 316.23 (June 1987), 1429–1435. ISSN: 1533-4406. DOI: 10.1056/nejm198706043162301.

[40] J. J. McMurray, M. Packer, A. S. Desai, J. Gong, et al. "Angiotensin–Neprilysin Inhibition versus Enalapril in Heart Failure". In: *New England Journal of Medicine* 371.11 (Sept. 2014), 993–1004. ISSN: 1533-4406. DOI: 10.1056/nejmoa1409077.

[41] P. M. Ridker, B. M. Everett, T. Thuren, J. G. MacFadyen, et al. "Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease". In: *New England Journal of Medicine* 377.12 (Sept. 2017), 1119–1131. ISSN: 1533-4406. DOI: 10.1056/nejmoa1707914.

[42] S. D. Solomon, J. J. McMurray, B. Claggett, R. A. de Boer, et al. "Dapagliflozin in Heart Failure with Mildly Reduced or Preserved Ejection Fraction". In: *New England Journal of Medicine* 387.12 (Sept. 2022), 1089–1098. ISSN: 1533-4406. DOI: 10.1056/nejmoa2206286.

[43] J. Hippisley-Cox, C. Coupland, and P. Brindle. "Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study". In: *BMJ* (May 2017), j2099. ISSN: 1756-1833. DOI: 10.1136/bmj.j2099.

[44] R Conroy. "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project". In: *European Heart Journal* 24.11 (June 2003), 987–1003. ISSN: 0195-668X. DOI: 10.1016/s0195-668x(03)00114-3.

[45] D. C. Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady, et al. "2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines". In: *Circulation* 129.25$_{s}uppl_{2}$ (June 2014). ISSN: 1524-4539. DOI: 10.1161/01.cir.0000437741.48606.98.

[46] W. H. Organization et al. "Prevention of cardiovascular disease: guidelines for assessment and management of total cardiovascular risk". In: *Prevention of cardiovascular disease: guidelines for assessment and management of total cardiovascular risk*. 2007.

[47] G. Assmann, P. Cullen, and H. Schulte. "Simple Scoring Scheme for Calculating the Risk of Acute Coronary Events Based on the 10-Year Follow-Up of the Prospective Cardiovascular Munster (PROCAM) Study". In: *Circulation* 105.3 (Jan. 2002), 310–315. ISSN: 1524-4539. DOI: 10.1161/hc0302.102575.

[48] R. B. D'Agostino, R. S. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro, and W. B. Kannel. "General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study". In: *Circulation* 117.6 (Feb. 2008), 743–753. ISSN: 1524-4539. DOI: 10.1161/circulationaha.107.699579. URL: http://dx.doi.org/10.1161/CIRCULATIONAHA.107.699579.

[49] D. R. Cox. "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 34.2 (Jan. 1972), 187–202. ISSN: 1467-9868. DOI: 10.1111/j.2517-6161.1972.tb00899.x.

[50] F. E. Harrell. "Evaluating the Yield of Medical Tests". In: *JAMA: The Journal of the American Medical Association* 247.18 (May 1982), p. 2543. ISSN: 0098-7484. DOI: 10.1001/jama.1982.03320430047030.

[51] P. J. Heagerty, T. Lumley, and M. S. Pepe. "Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker". In: *Biometrics* 56.2 (June 2000), 337–344. ISSN: 1541-0420. DOI: 10.1111/j.0006-341x.2000.00337.x.

[52] W. Einthoven. "The string galvanometer and the measurement of the action currents of the heart". In: *Nobel Lecture, December* 11 (1925), pp. 10–6.

[53] W. Einthoven. "THE DIFFERENT FORMS OF THE HUMAN ELECTROCARDIOGRAM AND THEIR SIGNIFICATION." In: *The Lancet* 179.4622 (Mar. 1912), 853–861. ISSN: 0140-6736. DOI: 10.1016/s0140-6736(00)50560-1.

[54] M. Sokolow and T. P. Lyon. "The ventricular complex in left ventricular hypertrophy as obtained by unipolar precordial and limb leads". In: *American Heart Journal* 37.2 (Feb. 1949), 161–186. ISSN: 0002-8703. DOI: 10.1016/0002-8703(49)90562-1.

[55] L. Sörnmo and P. Laguna. "Electrocardiogram (ECG) Signal Processing". In: Wiley, Apr. 2006. ISBN: 9780471740360. DOI: 10.1002/9780471740360.ebs1482.

[56] S. Butterworth et al. "On the theory of filter amplifiers". In: *Wireless Engineer* 7.6 (1930), pp. 536–541.

[57] D. Ricciardi, I. Cavallari, A. Creta, G. Di Giovanni, et al. "Impact of the high-frequency cutoff of bandpass filtering on ECG quality and clinical interpretation: A comparison between 40Hz and 150Hz cutoff in a surgical preoperative adult outpatient population". In: *Journal of Electrocardiology* 49.5 (Sept. 2016), 691–695. ISSN: 0022-0736. DOI: 10.1016/j.jelectrocard.2016.07.002.

[58] C. Bingham, M. Godfrey, and J. Tukey. "Modern techniques of power spectrum estimation". In: *IEEE Transactions on Audio and Electroacoustics* 15.2 (June 1967), 56–66. ISSN: 0018-9278. DOI: 10.1109/tau.1967.1161895.

[59] A. Akan and L. F. Chaparro. *Signals and systems using MATLAB®*. Elsevier, 2024.

[60] B. Mandelbrot. "How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension". In: *Science* 156.3775 (May 1967), 636–638. ISSN: 1095-9203. DOI: 10.1126/science.156.3775.636.

[61] C. E. Shannon. "A mathematical theory of communication". In: *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1 (Jan. 2001), 3–55. ISSN: 1931-1222. DOI: 10.1145/584091.584093.

[62] A. N. Kolmogorov. "Three approaches to the quantitative definition of information*". In: *International Journal of Computer Mathematics* 2.1–4 (Jan. 1968), 157–168. ISSN: 1029-0265. DOI: 10.1080/00207166808803030.

[63] P. Welch. "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms". In: *IEEE Transactions on Audio and Electroacoustics* 15.2 (June 1967), 70–73. ISSN: 0018-9278. DOI: 10.1109/tau.1967.1161901.

[64] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle. "Recurrence Plots of Dynamical Systems". In: *Turbulence, Strange Attractors and Chaos*. WORLD SCIENTIFIC, Sept. 1995, 441–445. DOI: 10.1142/9789812833709_0030.

[65] F. Takens. "Detecting strange attractors in turbulence". In: *Dynamical Systems and Turbulence, Warwick 1980*. Springer Berlin Heidelberg, 1981, 366–381. ISBN: 9783540389453. DOI: 10.1007/bfb0091924.

[66] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen. "NeuroKit2: A Python toolbox for neurophysiological signal processing". In: *Behavior Research Methods* 53.4 (Feb. 2021), 1689–1696. ISSN: 1554-3528. DOI: 10.3758/s13428-020-01516-y.

[67] J. Martinez, R. Almeida, S. Olmos, A. Rocha, and P. Laguna. "A Wavelet-Based ECG Delineator: Evaluation on Standard Databases". In: *IEEE Transactions on Biomedical Engineering* 51.4 (Apr. 2004), 570–581. ISSN: 0018-9294. DOI: 10.1109/tbme.2003.821031.

[68] M. Vetterli and J. Kovacevic. *Wavelets and subband coding*. Vol. 87. Prentice Hall PTR Englewood Cliffs, NJ, 1995.

[69] H. Malvar. "Lapped transforms for efficient transform/subband coding". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 38.6 (June 1990), 969–978. ISSN: 0096-3518. DOI: 10.1109/29.56057.

[70] M. R. B. Clarke, R. O. Duda, and P. E. Hart. "Pattern Classification and Scene Analysis." In: *Journal of the Royal Statistical Society. Series A (General)* 137.3 (1974), p. 442. ISSN: 0035-9238. DOI: 10.2307/2344977.

[71] C. Spearman. "The Proof and Measurement of Association Between Two Things." In: *Studies in individual differences: The search for intelligence.* Appleton-Century-Crofts, 1961, 45–58. DOI: 10.1037/11491-005.

[72] S. S. Shapiro and M. B. Wilk. "An analysis of variance test for normality (complete samples)". In: *Biometrika* 52.3–4 (Dec. 1965), 591–611. ISSN: 1464-3510. DOI: 10.1093/biomet/52.3-4.591.

[73] W. H. Kruskal and W. A. Wallis. "Use of ranks in one-criterion variance analysis". In: *Journal of the American statistical Association* 47.260 (1952), pp. 583–621.

[74] G Rupert Jr et al. "Simultaneous statistical inference". In: Springer Science & Business Media, 2012.

[75] R. A. Fisher. "Statistical Methods for Research Workers". In: *Breakthroughs in Statistics*. Springer New York, 1992, 66–70. ISBN: 9781461243809. DOI: 10.1007/978-1-4612-4380-9_6.

[76] R. Battiti. "Using mutual information for selecting features in supervised neural net learning". In: *IEEE Transactions on Neural Networks* 5.4 (July 1994), 537–550. ISSN: 1045-9227. DOI: 10.1109/72.298224.

[77] K. Pearson. "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (July 1900), 157–175. ISSN: 1941-5990. DOI: 10.1080/14786440009463897.

[78] Y. Benjamini and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 57.1 (Jan. 1995), 289–300. ISSN: 1467-9868. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[79] S. Holm. "A simple sequentially rejective multiple test procedure". In: *Scandinavian journal of statistics* (1979), pp. 65–70.

[80] D. M. Allen. "The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction". In: *Technometrics* 16.1 (Feb. 1974), 125–127. ISSN: 1537-2723. DOI: 10.1080/00401706.1974.10489157.

[81] M. Stone. "Cross-Validatory Choice and Assessment of Statistical Predictions". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 36.2 (Jan. 1974), 111–133. ISSN: 1467-9868. DOI: 10.1111/j.2517-6161.1974.tb00994.x.

[82] Y. Freund and R. E. Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55.1 (Aug. 1997), 119–139. ISSN: 0022-0000. DOI: 10.1006/jcss.1997.1504.

[83] L. Breiman. "Bagging predictors". In: *Machine Learning* 24.2 (Aug. 1996), 123–140. ISSN: 1573-0565. DOI: 10.1007/bf00058655.

[84] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification And Regression Trees*. Routledge, Oct. 2017. ISBN: 9781315139470. DOI: 10.1201/9781315139470.

[85] P. Geurts, D. Ernst, and L. Wehenkel. "Extremely randomized trees". In: *Machine Learning* 63.1 (Mar. 2006), 3–42. ISSN: 1573-0565. DOI: 10.1007/s10994-006-6226-1.

[86] C. E. Rasmussen. "Gaussian Processes in Machine Learning". In: *Advanced Lectures on Machine Learning*. Springer Berlin Heidelberg, 2004, 63–71. ISBN: 9783540286509. DOI: 10.1007/978-3-540-28650-9_4.

[87] J. H. Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[88] T. Cover and P. Hart. "Nearest neighbor pattern classification". In: *IEEE Transactions on Information Theory* 13.1 (Jan. 1967), 21–27. ISSN: 1557-9654. DOI: 10.1109/tit.1967.1053964.

[89] R. A. Fisher. "THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS". In: *Annals of Eugenics* 7.2 (Sept. 1936), 179–188. ISSN: 2050-1439. DOI: 10.1111/j.1469-1809.1936.tb02137.x.

[90] C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine Learning* 20.3 (Sept. 1995), 273–297. ISSN: 1573-0565. DOI: 10.1007/bf00994018.

[91] D. R. Cox. "The Regression Analysis of Binary Sequences". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 20.2 (July 1958), 215–232. ISSN: 1467-9868. DOI: 10.1111/j.2517-6161.1958.tb00292.x.

[92] F. Rosenblatt. "The perceptron: A probabilistic model for information storage and organization in the brain." In: *Psychological Review* 65.6 (1958), 386–408. ISSN: 0033-295X. DOI: 10.1037/h0042519.

[93] C. R. Rao. "The utilization of multiple measurements in problems of biological classification". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 10.2 (1948), pp. 159–203.

[94] L. Breiman. "Random Forests". In: *Machine Learning* 45.1 (Oct. 2001), 5–32. ISSN: 1573-0565. DOI: 10.1023/a:1010933404324.

[95] H. Robbins and S. Monro. "A stochastic approximation method". In: *The annals of mathematical statistics* (1951), pp. 400–407.

[96] J. Berkson. "Application of the Logistic Function to Bio-Assay". In: *Journal of the American Statistical Association* 39.227 (Sept. 1944), 357–365. ISSN: 1537-274X. DOI: 10.1080/01621459.1944.10500699.

[97] O. Mangasarian. "Arbitrary-norm separating plane". In: *Operations Research Letters* 24.1–2 (Feb. 1999), 15–23. ISSN: 0167-6377. DOI: 10.1016/s0167-6377(98)00049-2.

[98] T. Zhang. "Statistical behavior and consistency of classification methods based on convex risk minimization". In: *The Annals of Statistics* 32.1 (Feb. 2004). ISSN: 0090-5364. DOI: 10.1214/aos/1079120130.

[99] W. R. W. M. Nelder J. A. "Generalized Linear Models". In: *Journal of the Royal Statistical Society. Series A (General)* 135.3 (1972), p. 370. ISSN: 0035-9238. DOI: 10.2307/2344614. URL: http://dx.doi.org/10.2307/2344614.

[100] R. A. Fisher. "On the mathematical foundations of theoretical statistics". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222.594–604 (Jan. 1922), 309–368. ISSN: 2053-9258. DOI: 10.1098/rsta.1922.0009.

[101] R. C. Williams C. "Gaussian processes for regression". In: *Advances in neural information processing systems* 8 (1995).

[102] G. A. Seber and A. J. Lee. *Linear regression analysis*. John Wiley & Sons, 2003.

[103] N. Draper. *Applied regression analysis*. McGraw-Hill. Inc, 1998.

[104] A. E. Hoerl and R. W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 12.1 (Feb. 1970), 55–67. ISSN: 1537-2723. DOI: 10.1080/00401706.1970.10488634.

[105] R. Tibshirani. "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (Jan. 1996), 267–288. ISSN: 1467-9868. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[106] H. Zou and T. Hastie. "Regularization and Variable Selection Via the Elastic Net". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (Mar. 2005), 301–320. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2005.00503.x.

[107] J. Bergstra and Y. Bengio. "Random search for hyper-parameter optimization". In: *J. Mach. Learn. Res.* 13.null (Feb. 2012), 281–305. ISSN: 1532-4435.

[108] M. Buckland and F. Gey. "The relationship between Recall and Precision". In: *Journal of the American Society for Information Science* 45.1 (Jan. 1994), 12–19. ISSN: 1097-4571. DOI: 10.1002/(sici)1097-4571(199401)45:1<12::aid-asi2>3.0.co;2-l.

[109] R. Guns, C. Lioma, and B. Larsen. "The tipping point: F-score as a function of the number of retrieved items". In: *Information Processing & Management* 48.6 (Nov. 2012), 1171–1180. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2012.02.009.