Maria Inês António Roseiro

CARDIOVASCULAR RISK EVALUATION
FUSION OF CLINICAL EVIDENCE AND NEW RISK FACTORS

1 2 9 0 UNIVERSIDADE D COIMBRA

1 2 9 0

UNIVERSIDADE D
COIMBRA

Maria Inês António Roseiro

CARDIOVASCULAR RISK EVALUATION

FUSION OF CLINICAL EVIDENCE AND NEW RISK FACTORS

Dissertation in the context of the Master in Informatics Engineering, Specialization in Intelligent Systems, advised by Professor Jorge Henriques and Professor Simão Paredes and Dr. José Pedro Sousa and presented to Faculty of Sciences and Technology / Department of Informatics Engineering.

September 2021

This page is intentionally left blank.

Faculty of Sciences and Technology

Department of Informatics Engineering

Cardiovascular Risk Evaluation Fusion of Clinical Evidence and New Risk Factors

Maria Inês António Roseiro

Dissertation in the context of the Master in Informatics Engineering, Specialization in Intelligent Systems advised by Prof. Jorge Henriques and Prof. Simão Paredes and Dr. José Pedro Sousa and presented to the
Faculty of Sciences and Technology / Department of Informatics Engineering.

September 2021

UNIVERSIDADE D
COIMBRA

This page is intentionally left blank.

Acknowledgements

To my advisers, Jorge Henriques and Simão Paredes, for their guidance and support throughout the past year. I could not have accomplished this without you. To Dr. José Pedro Sousa, for the constant clinical advice, crucial in this thesis's development.

To my parents and sister, for constant encouragement and support through my life. And to my uncles for always having and being a home for me.

To all my friends, to the ones I made when I first entered DEI, and to the ones who were already there, thank you for making this path so much better. And to Gonçalo, who is always there when I need him.

Moreover, this work was supported by the lookAfterRisk Project (POCI-01-0145-FEDER-030290).

This page is intentionally left blank.

Agradecimentos

Aos meus orientadores, Jorge Henriques e Simão Paredes, pelo acompanhamento e ajuda no último ano, e sem os quais não teria conseguido desenvolver este trabalho. Ao Dr. José Pedro Sousa, pelo constante acompanhamento a nível clínico, crucial para o desenvolvimento desta dissertação.

Aos meus pais e à minha irmã, por todo o apoio incondicional e suporte ao longo da vida. E aos meus tios, por terem e serem sempre casa para mim.

A todos os meus amigos, os que fiz no DEI, e os que já antes faziam parte da minha vida, obrigada por fazerem este caminho tão melhor e mais leve. E a ti Gonçalo, por estares sempre comigo.

Este trabalho foi realizado no âmbito do projeto lookAfterRisk (POCI-01-0145-FEDER-030290).

This page is intentionally left blank.

Abstract

According to the World Health Organization, cardiovascular diseases are the world's leading cause of death. Although most of them can be prevented by addressing behavioural risk factors (like an unhealthy diet and obesity, smoking habits, physical inactivity, or the harmful use of alcohol), the number of occurrences tends to increase globally. In the cardiovascular diseases field, risk models are very useful tools to deal with patients who are sick, helping to predict the occurrence of some events (e.g. re-hospitalisation or death). The GRACE model is one of the most accepted clinical models, developed from a registry with a population across all spectrum of Acute Coronary Syndrome with the premise of predicting events in a short-term period. Although the GRACE risk model is frequently adopted (in Portugal, it is the most used risk stratification method to assess patients with Acute Coronary Syndromes), it uses a limited range of variables to preserve simplicity. However, it is recognised by the clinical community that the use of more information, particularly from clinical records, may complement the performance of this risk score. Moreover, some recent research has suggested that an inflammatory process may be associated with a worse prognosis on adverse cardiac events.

The first purpose of this work is to analyse if inflammation biomarkers can be applied to improve the GRACE score performance and, considering the obtained results, advance with our primary objective, which is the development of a new risk stratification tool. By using computational intelligence methodologies, capable of combining data from the already accepted GRACE model with additional knowledge on patients inflammation biomarkers, we aim to improve GRACE discriminate power. To accomplish that, our work proposes an innovative machine learning-based approach by implementing a system that combines knowledge (clinical evidence reflected on GRACE model) and data-driven techniques. This approach addresses simultaneously two of the central requirements of Explainable Artificial Intelligence: interpretability and personalisation, without impairing the model's performance. Briefly, our system is based on the creation of several rules, and uses a Machine Learning model to verify their correctness for each patient, creating a subset of useful rules, that aims to simulate clinical reasoning. The created subset is then combined to compute the final risk score. Since our model is determining the most valuable rules for each patient, we assure personalisation. Interpretability is guaranteed through the rule creating process and was validated by our clinical partner, giving their importance.

Moreover, this study was performed in collaboration with the Centro Hospitalar e Universitário de Coimbra (CHUC), which provided the clinical dataset used for our strategy validation.

Lastly, the evaluation of the developed approach enables us to conclude that an inflammatory process can affect acute coronary syndrome outcomes and the proposed methodology gains 5% on the AUC performance when compared with GRACE Risk Score, which is considered relevant in clinical domains based on decision-support systems, while guaranteeing the interpretability and personalisation of the model.

Keywords

Computational Intelligence, Cardiovascular Diseases, Acute Coronary Syndrome, GRACE Risk Score, Interpretability, Personalisation

This page is intentionally left blank.

Resumo

De acordo com os registos da Organização Mundial de Saúde, as doenças cardiovasculares são responsáveis pelo maior número de mortes em todo o mundo. Ainda que a sua prevenção ocupe um papel fundamental na redução de casos, nomeadamente recorrendo a hábitos de vida saudáveis, a verdade é que o número de ocorrências tem vindo a aumentar.

No campo das doenças cardiovasculares, a análise de riscos e a construção de modelos de risco são ferramentas úteis e habitualmente utilizadas em doentes com manifestações agudas, de maneira a auxiliar na previsão de uma série de fatores. O modelo GRACE, atualmente o mais utilizado em Portugal, é um dos modelos cientificamente aceites. Este modelo foi desenvolvido a partir de um registo internacional de doentes, baseado em dados de todos os espectros das doenças coronárias agudas. Ainda que muito usado, este modelo apresenta algumas limitações, nomeadamente devido à utilização de um número limitado de variáveis. O uso de informação mais detalhada sobre cada indivíduo poderia de alguma forma melhorar a performance deste modelo, nomeadamente recorrendo a informação sobre os marcadores de inflamação de cada doente. Pesquisas recentes sugerem uma associação entre a presença de marcadores de inflamação com a existência de um prognóstico mais delicado em manifestações agudas de doenças cardíacas.

O trabalho desenvolvido tinha como objectivo inicial analisar se a existência de marcadores de inflamação teria algum valor preditivo, de maneira a melhorar o modelo GRACE, correntemente aceite. Considerando os resultados desta análise, o principal objectivo desta dissertação é desenvolver um novo modelo de risco utilizando métodos de inteligência computacional que combinem o conhecimento já existente a nível do modelo GRACE com novas informações de cada paciente no âmbito dos marcadores de inflamação, procurando melhorar o poder discriminativo do modelo já existente. De maneira a alcançar o objectivo mencionado, o nosso trabalho propõe uma abordagem inovadora, baseada em técnicas de *machine learning*, ao implementar um sistema que combine técnicas baseadas em conhecimento com técnicas baseadas em dados. Esta abordagem tem em conta dois dos principais pilares da área de Explainable AI: a interpretabilidade e a personalização do modelo, sem que a sua performance seja prejudicada. De uma maneira geral, o nosso sistema baseia-se num conjunto de regras simples, e utiliza um modelo de Machine Learning para verificar quais as regras aplicáveis a cada paciente, de maneira a simular o raciocínio clinico, e guarda um subconjunto regras úteis, garantindo assim a personalização do modelo, que será a base para o cálculo de risco final. A interpretabilidade é garantida através do uso de regras simples. Este último passo foi validado pelo nosso parceiro clínico, dado a sua importância. Este estudo foi realizado em parceria com o Centro Hospitalar e Universitário de Coimbra (CHUC), que nos forneceu o dataset utilizado para validar a estratégia desenvolvida.

Por fim, a avaliação das estratégias desenvolvidas permite-nos concluir que um processo inflamatório tem de facto influência no desenvolvimento de um síndrome coronário agudo, e o modelo proposto ganha cerca de 5% ao nível da AUC, quando comparado com o modelo GRACE, o que se considera relevante em domínios clínicos baseados em tomadas de decisão.

Palavras-Chave

Inteligência Computacional, Doenças Cardiovasculares, Síndrome Coronária Aguda, Modelo de Risco GRACE, Interpretabilidade, Personalização

This page is intentionally left blank.

Contents

Acronyms

ACC American College of Cardiology. 22

ACS Acute Coronary Syndrome. 3, 4, 7, 8, 10, 16, 17, 20–23, 27, 29–31, 33, 39, 40, 43, 46, 48, 50, 51, 61

AI Artificial Intelligence. 4, 14, 20, 21, 27, 51

ANN Artificial Neural Network. 15, 21, 22

AUC Area Under the Curve. xv, 10, 22, 26, 31, 32, 43, 45, 46, 48–50

AUROC Area Under ROC Curve. 26, 29

CHUC Centro Hospitalar e Universitário de Coimbra. 4

CRP C-reactive Protein. 13, 14, 21, 29, 31, 43, 44, 50, 59

CVD Cardiovascular Diseases. 7, 8

DNN Deep Neural Network. 15

DT Decision Trees. 15–18, 46

ECG Electrocardiogram. 3, 8, 9, 14, 22

FN False Negative. 23, 29

FP False Positive. 23, 29

GBRT Gradient Boosting Regression Trees. 15, 18, 22, 23, 47

GM Geometric Mean. 29, 31, 32, 34, 35, 43, 45, 46, 48–50

GRACE Global Registry of Acute Coronary Events. i, xv, 3, 4, 10, 22, 23, 29–32, 39, 43–46, 49–52, 59

KNN K-nearest Neighbours. 30, 40

LDA Linear Discriminant Analysis. 20, 22, 35, 47, 48

MI Myocardial Infarction. 8, 9, 11, 21, 22

ML Machine Learning. i, xv, 4, 7, 14, 15, 19–23, 27, 29–32, 36–39, 43–51

NICE National Institute for Health and Clinical Excellence. 30

NSTEMI Non-ST elevated Myocardial Infarction. 8, 10, 11, 17, 30, 39, 41

PURSUIT Platelet glycoprotein IIb/IIIa in Unstable agina: Receptor Suppression Using Integrilin. 3, 10, 12

RF Random Forest. 15, 17, 18, 22, 31, 44, 47, 49, 50

ROC Receiver Operating Characteristic. xv, 24, 26, 43, 44, 46

SMOTE Synthetic Minority Oversampling Technique. 30

STEMI ST-elevated Myocardial Infarction. 8, 9, 17, 30, 39, 41

SVM Support-vector Machines. 15, 16, 21, 22, 47

TIMI Thrombolysis in Myocardial Infarction. 3, 10, 11, 22

TN True Negative. 23, 29

TP True Positive. 23, 29

UA Unstable Angina. 8, 17, 21, 30, 39, 41

WBC White Blood Cell. xv, 13, 14, 29, 31, 40, 43, 44, 59, 60

WHO World Health Organization. 8

XAI Explainable Artificial Intelligence. 4, 27

List of Figures

This page is intentionally left blank.

List of Tables

This page is intentionally left blank.

Chapter 1

Introduction

In Europe, cardiovascular diseases are the leading cause of death, responsible for 3.9 million deaths annually, approximately 37% of all deaths. Several European countries have implemented measures to decrease mortality, such as treatment improvements, disease prevention campaigns, and better disease management at unit cares. However, despite the decreasing mortality rates from cardiovascular diseases, the disease prevalence for non-fatal events of Acute Coronary Syndrome (ACS) remains high, and the economic costs are immense [9].

As part of this effort, Portugal has established several strategies to indirectly fight cardiovascular disease mortality, including a smoking ban in 2008 and salt reduction regulations in 2010. Moreover, the program "Coronary Fast-Track System" was implemented in Portuguese hospitals to produce early reliable diagnoses at the beginning of 2000. It is of utmost importance to perform an early diagnosis and determine the risk stratification on every patient. These methods are based on the past clinical history, Electrocardiogram (ECG) results, and laboratory tests on cardiac biomarkers. Not only it facilitates a more accurate prognostic of the illness but also determines the triage of patients with ACS. Therefore, several different risk scores have been proposed over the last decades to help risk stratification. For short-term prognosis, the most commonly known are the GRACE risk score, clinically developed on international databases and Thrombolysis in Myocardial Infarction (TIMI) and Platelet glycoprotein IIb/IIIa in Unstable agina: Receptor Suppression Using Integrilin (PURSUIT) risk scores, derived from clinical trials analysis.

1.1 Motivation

Despite GRACE, TIMI and PURSUIT Risk Scores having been validated, they present some limitations. As they are population-based models with fixed variables, they typically manifest a poor performance when applied to the risk profile of an individual. In addition, they have fundamental similarities to linear models, limiting their ability to explore higher-order interactions like incorporating new knowledge and new risk factors, ending with sub-optimal performances.

Furthermore, some reports on haematological indices of inflammation have been associated with poorer prognosis or major adverse cardiac events after an ACS event, which may suggest that an inflammation process has a primary role in the progression of coronary artery lesions. However, the knowledge on their influence is a recent research field.

With the growing availability of data, Artificial Intelligence (AI) has achieved notable accomplishments, and its research influence in healthcare issues is unquestionable. Despite all the success, the development of AI strategies in this practice is somehow limited. The most understandable barriers to successfully implementing these methods are the lack of transparency, interpretability and personalisation. These elements are crucial when assuring confidence to the clinical community, but, most importantly, the patient's safety. According to International Business Machines Corporation (IBM), Explainable Artificial Intelligence (XAI) aims to achieve a solution that allows human users to understand and trust the results and outputs derived from ML algorithms and efficiently supports clinicians in their decisions [41].

The considerations mentioned above constitute the main challenge that motivates this work. This study will focus on creating a model that ensures interpretability and personalisation to the medical community without deteriorating the ML model's prediction metrics, as well as a study on inflammation biomarkers as new risk factors.

1.2 Objectives

The objectives of this work can be divided into two major categories. In the clinical field, we established a clinical research question that aims to understand if the introduction of new risk factors, in our case, the inflammation biomarkers, can improve GRACE performance. To perform the mentioned analysis, clinical evidence (GRACE risk factors) are combined with new knowledge (inflammation biomarkers) and our hypothesis is tested on 6-months mortality after an ACS event.

In the innovative field, our main objective is to develop a ML-based model that merges GRACE risk factors with the previously obtained knowledge on inflammation biomarkers. Considering that, our research work proposes a hybrid approach by developing a more explainable ML-based decision support system. Even though the final model needs to be predictive, physicians also have to find them interpretative. The need for interpretability on the development of the classification model is vital to increase professionals confidence and, accordingly, the model's acceptance. Moreover, our ML-based system will also approach personalisation since it is helpful to explore the most significant risk factors for each patient and can improve sub-optimal performances from the standard risk score models.

Our strategy will involve three major phases. The first one is the development of a rule-set based on clinical knowledge, overcoming the limitations of conventional ML (black-boxes) models. Secondly, we will implement a personalisation system to select the most appropriate rules for each patient. The third and final phase will connect the previous two stages and compute a prediction score for an ACS outcome.

Finally, our strategy will be validated through a real Portuguese dataset provided by the Centro Hospitalar e Universitário de Coimbra (CHUC) hospital. The dataset contains 1544 patients and presents a mortality rate of 9% (in 6-months) after experiencing an ACS episode. Comparisons with the traditional GRACE and a standard ML-based approach will be performed to perform a direct evaluation.

1.3 Structure

This document is structured in five chapters. Chapter 2 of this dissertation introduces pertinent background concepts in clinical and scientific fields, followed by a second section, which is intended to enumerate related healthcare and cardiology work. In the third section, the evaluation metrics used in our dissertation are detailed.

The proposed methodology is presented in Chapter 3, alongside additional motivation behind the importance of this research topic and the scientific choices for developing our approach. In Chapter 4, we discuss all the experiments performed, as well as their results and evaluation. Finally, Chapter 5 draws our conclusions and enumerates several ideas for future work.

This page is intentionally left blank.

Chapter 2

Background and Related Work

> It is not the strongest of the species that survives, nor the most intelligent, but the one most responsive to change.
>
> *Charles Darwin*

Nowadays, the volume of data produced by clinical devices can be overwhelming, and, if not appropriately well exploited, it is of little value. Therefore, methods that automatically explore this data and establish patterns and associations while discovering valuable evidence in clinical diagnoses are vital for decision-making processes in the healthcare area.

This chapter is dedicated to the background knowledge for the subject of this dissertation, and it also presents related work and the used metrics for evaluation, which is why it is divided into three sections.

The first section introduces fundamental concepts to understand our research question and the model proposed in our thesis.

The second section presents our thesis state of the art, comprising two different themes: studies on the inflammation biomarkers as Acute Coronary Syndrome (ACS) risk factors, and Machine Learning (Machine Learning (ML)) applications in cardiology.

The third section presents a description of the used metrics for evaluating all the developed models.

2.1 Background

In this section we will explain clinical concepts related to Acute Coronary Syndromes as diagnoses and Risk Scores and inflammation biomarkers, our study variables, as well as scientific concepts on ML domains.

2.1.1 Acute Coronary Syndrome

By definition, Cardiovascular Diseases (CVD) are a group of heart disorders, including vein thrombosis, pulmonary embolisms, and injuries on the blood vessels that supply several

organs and members of our body.

Coronary Heart Diseases are one group of CVD in which atherosclerosis occurs in the vessels that supply oxygen and blood to the heart muscle. Atherosclerosis in coronary arteries blocks the blood flow to the heart, increasing the patient's risk for critical conditions. Acute Coronary Syndrome is the clinical manifestation of these severe conditions that follow the disruption of a coronary arterial plaque and are immediate causes for medical emergency care and hospitalisation.

Although physicians should immediately consider the hypothesis of ACS in all patients, a formal medical decision is only complete with a review of several clinical features: an electrocardiogram (ECG), the biomarkers of myocardial necrosis and a thorough assessment of the patient's history and findings on physical examinations. Clinical signs and symptoms typically include chest pain, dyspnea, nausea, fatigue and abdominal pain. Accurate diagnosis and early risk stratification are essential for guiding treatment and saving lives.

2.1.2 Diagnoses

Acute Coronary Syndrome (ACS) includes 2 different myocardial ischemic states, designated by **ST-elevated Myocardial Infarction (STEMI)** and non–ST elevation ACS, which includes **Unstable Angina (UA)** and **Non-ST elevated Myocardial Infarction (NSTEMI)** because the two are often indistinguishable at presentation [47].

STEMI and NSTEMI diagnoses are confirmed with positive results on myocardial necrosis in the context of myocardial ischaemia, which is the inadequate blood and oxygen supply to the heart provoked by thrombotic occlusions of a coronary vessel. This event happens mainly due to atherosclerosis, which is the accumulation of cholesterol plaques in artery walls.

ST-elevated myocardial infarction (STEMI) is determined with symptoms characteristic of cardiac ischaemia with persistent findings of ST complex elevations visible on Electrocardiogram (ECG). Nevertheless, several conditions may result in myocardial injuries and cause these elevations. Consequently, it can not be a single indicator to perform a correct diagnosis of Myocardial Infarction (MI).

At the same time, non-ST elevated myocardial infarction (NSTEMI) diagnoses are congruent with high Cardiac Biomarkers levels but not with cardiac ischaemia commonly characterized by ST-segment elevations.

Unstable angina (UA) provokes symptoms suggestive of cardiac ischaemia without elevated Cardiac Biomarkers, and any ECG changes.

Cardiac Biomarkers

According to World Health Organization (WHO), a biomarker is an accurate indicator that can be measured in the body and can have influence or predict the incidence of outcome or disease [49].

Cardiac Biomarkers, also known as **Troponins** are a group of proteins from the cardiac muscle fibres that regulate muscular contractions. Troponin T and Troponin I are highly sensitive markers of myocardial injury that may indicate necrosis in myocardial cells. Troponin T is measured using a simple analysis, with a cutoff value of 0.1 mg/litre showing myocardial damage [47]. On the other hand, The European Society of Cardiology and the

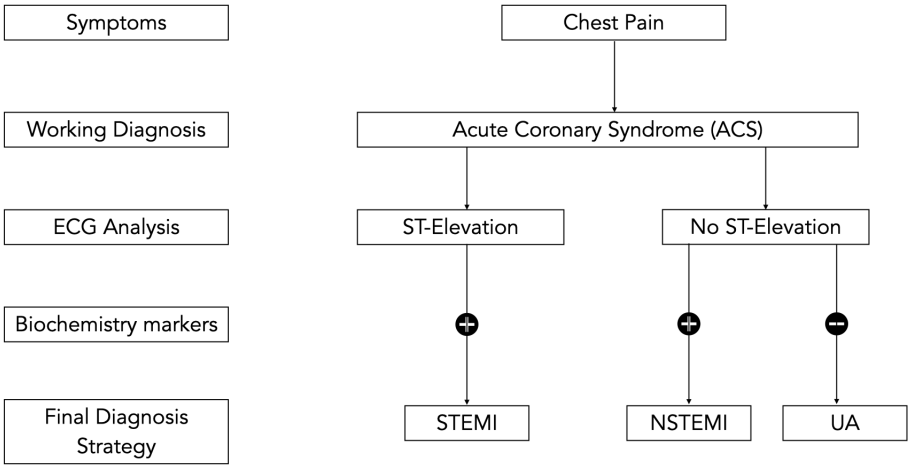


Figure 2.1: Definition of different diagnoses of Acute Coronary Syndrome. Adapted from: [9]

American College of Cardiology recommend that each laboratory determine its Troponin I cutoffs, defining myocardial damage at the value that exceeds the 99th percentile of a standard reference population [45]. With these criteria, the values indicative of myocardial damage range from 0.1 to 2 mg/litre.

Electrocardiogram Changes

An electrocardiogram (ECG) is a compilation of an individual cardiac electrical activity over a short period of continuous-time. An atrial depolarisation (P wave) is followed by a ventricular depolarisation (QRS complex) and with another ventricular depolarisation succeeding (T wave). This **P–QRS–T** sequence is successively recorded for each cardiac series. Depending on heart rate and rhythm, the interval between waves and cycles can be variable. [9].

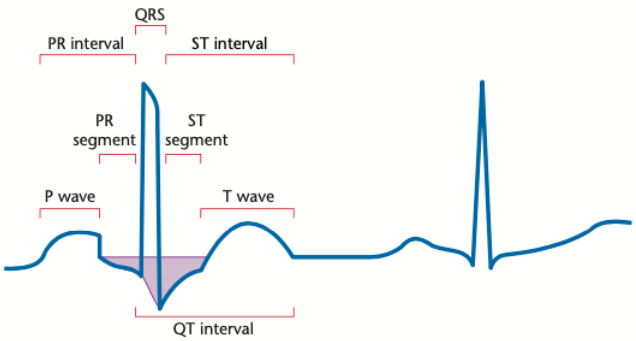


Figure 2.2: ECG morphology showing the different waves and intervals. Source: [9]

The *cutoff* that defines a STEMI diagnose it is an ST-segment elevation of $2mm$ in men and $1.5mm$ in women on at least two ECGs contiguous leads [21]. Similar to cardiac biomarkers, the ECG alone is insufficient to diagnose an acute MI [47]. ECG changes such as ST deviation may be present in other conditions, such as left ventricular hypertrophy.

If no elevation is visible, only the values from the biomarkers of myocardial cell injury (troponins) will determine whether a patient is having an NSTEMI with no standard ST-segment elevations [9].

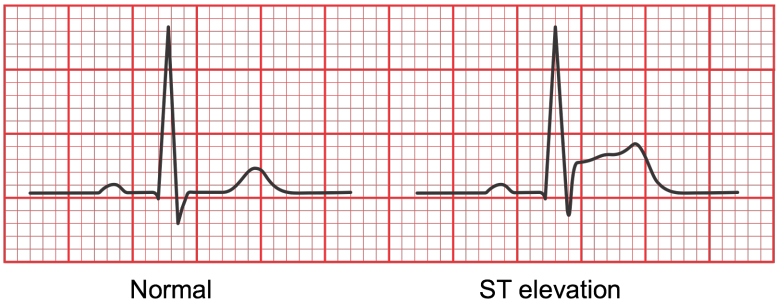


Figure 2.3: Example of the ST segment elevation (between S and T waves). This segment is clearly elevated from the baseline.

2.1.3 Risk scores in Acute Coronary Syndrome

Early risk stratification plays a central role in any field of medicine, mainly when dealing with acute diseases like Acute Coronary Syndrome (ACS). Risk stratification enables the physician to predict the illness and decide the mode and urgency of therapy. Nevertheless, the optimal management of underlying risk factors can be complex. There is a need for a simple tool accepted by the medical community to help with the triage process of these patients and evaluate the potential of different therapeutic interventions [1].

Platelet glycoprotein IIb/IIIa in Unstable agina: Receptor Suppression Using Integrilin (PURSUIT), **Thrombolysis in Myocardial Infarction (TIMI)** and **Global Registry of Acute Coronary Events (GRACE)** Risk Scores are 3 of the most acknowledged models that intend to estimate a patients mortality risk. Furthermore, they were all developed for a short-term prognosis ACS, to determine the patient's mortality risk, within 6 months for the GRACE, 14 days for TIMI, and 30 days for the PURSUIT model [12].

In an interesting Portuguese study, published in the European Heart Journal, *Gonçalves et al.* [12], the three mentioned risk scores, PURSUIT, TIMI and GRACE, were applied to the same group of 460 patients and analysed short-term as well as long-term mortality. They have demonstrated that GRACE score was the best in predicting the risk of death one year after admission and enabled the identification of subsets of high-risk patients who will benefit most from myocardial revascularisation therapeutics if early performed.

Global Registry of Acute Coronary Events (GRACE)

The **GRACE** scoring system is the latest and has originated from an international registry of a population across the entire spectrum of ACS [12]. Moreover, it is the most highly developed score system, with robust features like heart rate, blood pressure, survival from cardiac arrest, serum Creatinine, and Killip class [10]. Therefore, nowadays, it is the most used risk model in Portugal.

The applied risk factors contain more than 90% of the predictive capacity of the entire model, with the Area Under the Curve (AUC) for in-hospital death at 0.84 and further

myocardial infarction events at 0.71 [9]. The final score (ranging between 2 and 383) is the sum of all variables scores. The complete scoring system is given below in 2.1.

Variable	**Spectrum**	**Score**
Age (years)	<40	0
	40 – 49	18
	50 – 59	36
	60 – 69	55
	70 – 79	73
	>= 80	91
Heart rate (bpm)	<70	0
	70 – 89	7
	90 – 109	13
	110 – 149	23
	150 – 199	36
	>200	46
Systolic BP (mmHg)	<80	63
	80 – 99	58
	100 – 119	47
	120 – 139	37
	140 – 159	26
	160 – 199	11
	>200	0
Creatinine (mg/dL)	0 – 0.39	2
	0.4 – 0.79	5
	0.8 – 1.19	8
	1.2 – 1.59	11
	1.6 – 1.99	14
	2 – 3.99	23
	>4	31
Killip class	I	0
	II	21
	III	43
	IV	64
Cardiac arrest at admission	No	0
	Yes	43
Elevated cardiac markers	No	0
	Yes	15
ST-segment deviation	No	0
	Yes	30

Table 2.1: GRACE risk score specification variables and correspondent scores. Source: [12]

Thrombolysis in Myocardial Infarction (TIMI)

TIMI risk score was developed with the databases from large clinical trials of NSTEMI diagnoses [12]. The scoring system considers seven different risk factors, which are presented in the following table 2.2, and gives one mark to each of them if present. In addition, for the *known coronary artery stenosis (>= 50%)* risk factor was also attributed one mark to a history of myocardial infarction (MI) and coronary revascularization. This methodology is closer to real-world practice, and it has been validated by the authors of the TIMI

risk score [12]. The final score (0-7) returns the sum of every mark from the seven factors.

Variable	Spectrum	Score
Age (years) >= 65	No	0
	Yes	1
>= 3 risk factors for CHD	No	0
	Yes	1
Use of aspirin within the past 7 days	No	0
	Yes	1
Known coronary artery stenosis (>= 50%)	No	0
	Yes	1
1 episode of angina in <24 hours	No	0
	Yes	1
ST-segment deviation	No	0
	Yes	1
Elevated Cardiac Biomarkers	No	0
	Yes	1

Table 2.2: TIMI risk score variables and scores. Source: [12]

Platelet glycoprotein IIb/IIIa in Unstable angina: Receptor Suppression Using Integrilin (PURSUIT)

The **PURSUIT** scoring system was developed on extensive clinical trials databases. It predicts 30-day-risk by fusing information from early vital signs. The comprehended risk factors are age, gender, worst Canadian Cardiovascular Society (CCS) angina classification in the previous six weeks (table 2.4), heart rate, systolic blood pressure, signs of heart failure, and ST depression, detailed in table 2.3, with the respective score. The final score (0-20) is the sum of all five risk factors individual score. However, this system has not gained much popularity, and it is practically not applied.

Variable	Spectrum	Unstable Angina Score	Myocardial Infarction Score
Age (years)	<50	0	0
	50-59	8	11
	60-69	9	12
	70-79	11	13
	>80	12	14
Sex	Male	1	1
	Female	0	0
Worst CCS-class in previous 6 weeks	No angina or CCS I/II	0	0
	CCS III/IV	2	2
Signs of heart failure	No	0	0
	Yes	2	2
ST-depression on presenting ECG	No	0	0
	Yes	1	1

Table 2.3: PURSUIT risk score specification variables and correspondent scores. Source: [12]

Class	Name	Description of Angina Severity
0	Asymptomatic Angina	Light myocardial ischaemia with no symptoms.
I	Angina only with strenuous exertion	Ordinary physical activity (walking or climbing stairs) does not cause angina. Angina occurs with strenuous, rapid, or prolonged exercise.
II	Angina with moderate exertion	Slight limitations of ordinary activity. Angina occurs on walking or climbing stairs rapidly, under emotional stress, or only during the few hours after awakening. Angina occurs on walking more than 2 blocks on the level and climbing more than one flight of ordinary stairs at a normal pace and in normal condition.
III	Angina with mild exertion	Marked limitations of ordinary physical activity. Angina occurs on walking one to two blocks on the level and climbing one flight of stairs in normal conditions and at a normal pace.
IV	Angina at rest	Inability to do any activity without discomfort. Rest pain.

Table 2.4: Canadian Cardiovascular Society (CSS) Classification of Angina. Source: [53]

2.1.4 Biomarkers of inflammation

Over the last years, a collection of evidence has demonstrated that inflammation plays a crucial role in atherosclerosis occurrences and its complications. Several reports expressed that differences in the levels of inflammatory markers predict an unfavourable cardiovascular outcome in patients with acute coronary syndrome. Therefore, improved knowledge on the mechanisms of inflammation might improve the prognostic stratification process [6].

Inflammation is an automatic defence against infections. The inflammatory response is a defence mechanism that intends to kill pathogens and initiate tissue repair processes, helping restore homeostasis at infected and damaged spots. The human body response during an inflammatory process always assumes four phases. The first one is the increased blood supply to the inflammation site, followed by an increased capillary permeability caused by the retraction of endothelial cells, permitting larger molecules to deliver some soluble mediators to the inflammation site. Next, the leukocytes migrate from the capillaries to the surrounding tissues. Once in the tissue, the leukocytes move to the inflammation area. Finally, leukocytes release some mediators at the site of inflammation [8].

What starts as an acute response, if not adequately treated, can turn chronic. As a result, inflammation biomarkers are always present and can contribute to a range of diseases. As mentioned in 2.1.2, biomarkers are reliable markers of the human body and enable physicians to understand a patient's condition. The concentrations of **C-reactive Protein (CRP)**, **Albumin**, and **White Blood Cell (WBC)** are some of the strongest biomarkers that may indicate an inflammatory process. These substances can fluctuate widely in the blood flow during an acute response to an infection. Plasma concentrations of CRP can become 10000 times higher, and the leukocyte count can triple. The concentration of serum albumin can fall by about 20 per cent [11].

White Blood Cell count (WBC)

The definition of WBC (also known as leukocytes) stands for a type of blood cell made in the bone marrow found in the blood and lymph tissue. Leukocytes are part of the body's immune system and can help the body fight infection and other diseases [22].

An average value for leukocytes count has a standard range of 4.500/L to 10.500/L in adults. Above 11.000/L is considered to be high and indicates leukocytosis as a response to an infection process, inflammatory disorders, or an abnormal production as leukaemia [48].

Albumin Serum

Albumin is the main protein in blood plasma. Low serum albumin levels occur in people with malnutrition, inflammation, and severe liver and kidney disease [39].

The referenced average values are between 35 and 55 g/L. Albumin levels decrease when an inflammation process occurs. A hypoalbuminemia process occurs to albumin values of less than 25 g/L, and a mild one to albumin values of 2.5 to 3.5 g/L [32].

C-reactive Protein (CPR)

The last biomarker of inflammation that we will take into account in this research is C-reactive Protein (CRP). Produced by the human liver, it is discharged into the bloodstream in response to inflammation [55].

The standard value is less than 10.0 mg/L and indicates low risk of cardiac injuries. On the other hand, 10.0 to 30.0 mg/L suggests an average risk, and more than 30.0 mg/L designates a high risk of cardiac problems. Very high CRP levels (more than 10 mg/L) can also indicate a weakened immune response or an inflammatory disease [33].

2.1.5 Machine Learning in Acute Coronary Syndrome

ML is a sub-field of Artificial Intelligence (Artificial Intelligence (AI)) that studies computational algorithms ability to improve the performance of a specific task with data through experience, without direct human intervention. With the growing availability of clinical data, ML can undoubtedly contribute to a promising expansion of traditional risk scoring methods, with metrics predictions equivalent or superior to well-validated risk scores.

In healthcare, inputs in ML algorithms can include patient information such as age, gender, and medical history [25] or even disease-specific data like ECG results, clinical symptoms, medication, and several biomarkers. Outcomes usually comprise patient data such as disease indicators, patient survival (or the number of survivals to an episode), re-hospitalisation or even death registries.

ML can be divided into several subcategories: the major ones are Supervised Machine Learning and Unsupervised Machine Learning. Semi-supervised learning is accepted as a hybrid approach, and it is suitable for scenarios in which some outcome data are missing.

Supervised learning

Supervised learning uses labelled data-sets to train algorithms to classify data or predict outcomes. There are two groups of problems inside supervised learning techniques: classification and regression [26].

Classification

Classification problems use an algorithm to accurately assign test data into specific categories, such as separating death from survival outcomes. Linear classifiers, Support-vector Machines (SVM), Decision Trees (DT), Random Forest (RF) and Gradient Boosting Regression Trees (GBRT) are all common classification algorithms [26].

Our research will explore four different classification algorithms: Artificial Neural Network (ANN) and Support Vector Machines (SVM) since they are well-known and performing classifiers. Decision Trees (DT) based algorithms as Random Forest (RF) and Gradient Boosting (GBRT) will also be considered due to their similarities with medical decision procedures, where descriptive attributes resemble clinical variables.

Artificial Neural Network

Artificial Neural Networks (ANN) are mathematical models inspired by the structure and functions of biological neural networks. Figure 2.4 presents a typical Neural Network (ANN), which is divided into several layers: input, hidden layers, and an output layer, and can have various architectures.

Their connections are outputs of neurons, which became an input to the next neuron of the network. Each connection has an associated weight that represents its relative importance in the neural network. Every neuron can possess many to many relationships with multiple inputs and output connections. In every hidden layer node, an activation function produces a signal output resulting from their weighted sum. Next, the weights are optimised through a learning rule. The purpose is to estimate the weights through input and outcome data so that the average error between the outcome and their predictions is minimised [25].

Deep Neural Network (DNN) are a subset of ANN which have multiple hidden layers between the input and output. It adds complexity to the network's topology, allowing them to solve more complex problems. DNNs can be further specialised for particular tasks by choosing the operations used in each layer's nodes.

Support Vector Machine

The objective of Support Vector Machines (SVM) algorithms is to find a hyperplane in an N-dimensional space, where N is the number of features, that distinctly classifies the data points [15] . The formulation of SVM learning is based on minimizing structural risk. Instead of minimizing an objective function based on the training samples, it attempts to mitigate the bound on the generalization error. Generalization error is the error made by the ML algorithm on test data [13].SVM can separate two classes of data points with various hyperplanes. The main goal is to find the optimal hyperplane, which is the plane that maximizes the distance between data points of both classes [15]. Maximizing the margin distance provides some support so that the model can classify future data points with more confidence.

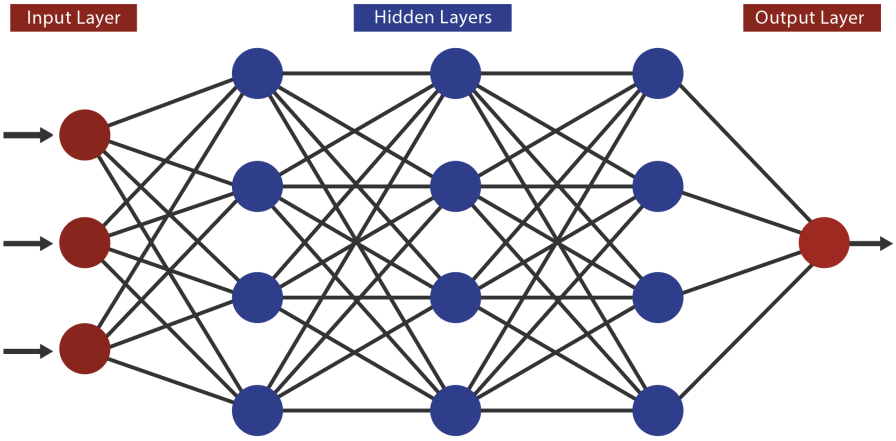


Figure 2.4: Example of a neural network architecture.

Hyper-planes are decision boundaries that support data point classification. Data points on distinct sides of the hyper-plane are attributed to different classes (see Fig. 2.5). Furthermore, the hyper-plane dimension depends on the number of features. Assuming that the number of input features is 2, the hyper-plane is a line.

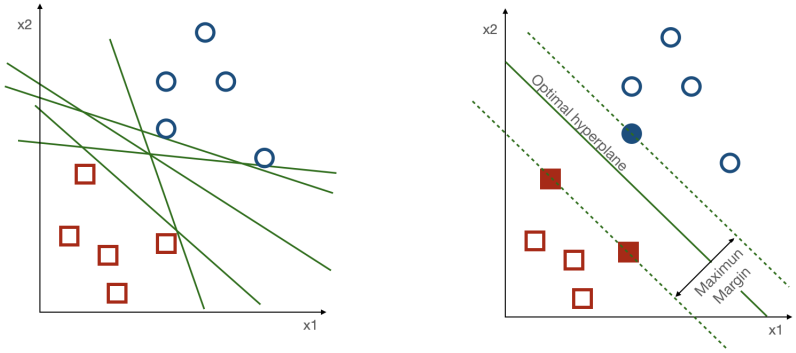


Figure 2.5: SVM hyperplanes in 2D illustration. Source:

Support vectors are data points closer to the hyper-plane and influence its position and orientation. Employing these, we maximise the margin of the classifier. If the support vectors are removed, the line equation would be altered since these are the points that construct the SVM. However, there are cases such that no line can classify the data set into two classes.

Decision Trees

Decision trees (DT) are models based on a tree-like graph architecture composed of decision nodes (which represent attributes) linked to two or more sub-trees and leaf nodes (which represent outcomes), with a decision caused by one or more conditions. The condition in each node tries to divide the data to maximise the discrimination power among the different classes. An instance is classified by beginning at the root node of the tree and moving along decision nodes based on the convergence of attributes [40]. When a leaf is encountered, its label presents the predicted class of the instance.

An example of a simple DT related to the diagnose of ACS can be seen in Figure 2.6. The

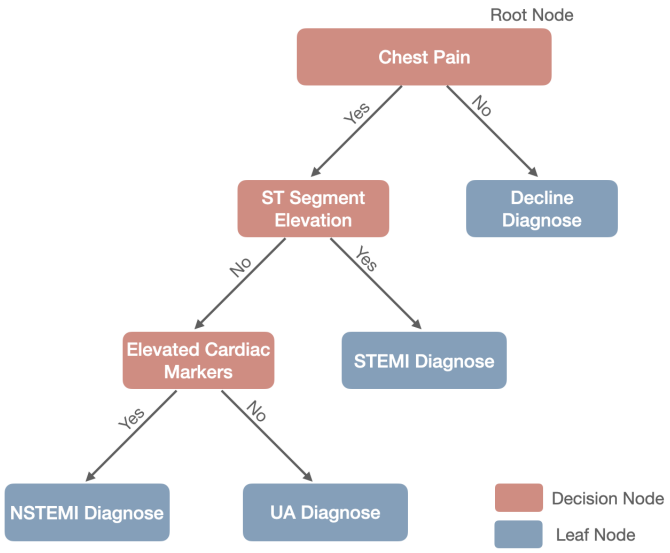


Figure 2.6: Decision Tree Example. Based on: [40]

interpretation of DT is very simple. From this tree the following rules can be deduced:

1. If a patient has no chest pain, then the ACS diagnose is declined.
2. If a patient has chest pain, and elevations of ST segment, then he has a STEMI diagnose.
3. If a patient has chest pain, no elevations of ST segment, and elevated cardiac markers, then he has a NSTEMI diagnose.
4. If a patient has chest pain, no elevations of ST segment, and no elevated cardiac markers, then he has UA.

Random Forest

Random Forest (RF) is an ensemble method that consists of many Decision Trees (DT). An ensemble is a collection of predictors combined to predict complex problems. Ensembling techniques can be classified into Bagging and Boosting. A Random Forest (RF) algorithm uses Bagging techniques [43].

Bagging is a simple ensembling method that creates several models independently and then combines them with model averaging techniques.

RF overcomes the limitations of a DT algorithm, which can learn highly irregular patterns leading to over-fitting scenarios. Taking this into account, RF increases precision by taking the average of the output from the multiple trees and establishes an outcome based on their predictions.

In summary, RF model averages a set of M Decision Trees (DT),

$$F(x) = \frac{\sum_{m=1}^{M} h_m(x)}{M} \tag{2.1}$$

where $h_m(x)$ is a modified DT that chooses a random subset of the features. [23].

RF models specify several hyper-parameters that can either be used to improve their predictive power or to make them faster. In the improvement field, there is the n estimators parameter, which is the number of Decision Trees (DT) that the algorithm builds before taking the maximum vote. Generally, a more significant number of trees increases the performance and makes the predictions more stable, but it also slows down the computation. Another two relevant hyperparameters are minimum samples split and minimum samples leaf. *Minimum samples split* specifies the minimum samples required to split an internal node, while *minimum samples leaf* specifies the minimum number of samples needed at a leaf node. We will not further explore hyper-parameter tuning to increase a model's speed since it is not relevant for our analysis.

Gradient Boosting

Gradient boosting (GBRT) is a Boosting ensemble technique in which the final model form is an ensemble of weak prediction models, normally Decision Trees (DT)[17].

Boosting is an ensemble technique in which the predictors are not made independently but sequentially [31]. In particular, we will mention additive predictive models of the form:

$$F(x) = \sum_{m=1}^{M} \lambda_m h_m(x) \tag{2.2}$$

where $h_m(x)$ are prediction models. GBRT will then expand the additive model in a greedy model:

$$F_m(x) = F_{m_1}(x) + \lambda_m h_m(x) \tag{2.3}$$

The initial model $F0$ is the mean of target values when the loss function is least squares regression or a quantile when using other loss functions.

GBRT techniques are well suited to a wide range of problems. Being a DT based algorithm makes GBRT a versatile method that can be used for features with different units or features that can't be easily normalised (features with minimal and massive values) [23].

Regression

Regression is a supervised learning method that uses an algorithm to explain a relationship between dependent and independent variables. Regression models can help predict numerical values based on different data points, such as risk variables projections for a given patient. Some popular regression algorithms are linear regression and logistic regression [26].

Considering that unsupervised learning is commonly used for feature extraction techniques and supervised learning to establish associations between the input (as patient features) and the output (as outcomes), supervised learning can contribute more to obtain clinically relevant results. In consequence, computational intelligence applications in healthcare often use supervised learning techniques [25].

Since our approach aims to explore a risk score that predicts an individual's mortality risk, the core of our system will make use of **supervised learning** methodologies.

Unsupervised Learning

Unsupervised Learning uses ML algorithms to interpret and cluster unlabelled data sets. These algorithms detect hidden patterns in data without human intervention. Unsupervised learning models can be used for two main tasks: clustering and association rules [26].

Clustering

Clustering is a Data Mining technique that groups unlabelled data based on their similarities, differences or relations among samples [29]. There are two significantly different types of clustering methods, hierarchical and partitional.

Hierarchical clustering method organises instances successively, merging the most similar two groups of objects based on the pair of distances between two groups until the process terminates, creating a hierarchical representation.

Partitioning methods create n partitions of the data. Similar instances are parts of the same group, and different instances are data points from other groups. Accordingly, the dissimilarity metric is the calculated distance between the groups. For instance, K-means clustering algorithms assign similar data subjects into groups and represent each cluster as the mean value of all elements in one group. The K value designates the number of groupings. This technique will be helpful to separate distinct risk patients (high-risk and low-risk) [29].

Association Association techniques use different rules to find relationships, frequent patterns, or structures between variables in a given dataset. A rule can be seen as an inference composed of two items: the antecedent and the consequent [16]. For example, equation 2.4 illustrates a simple rule that converts an individual risk factor into a binary outcome.

$$if AGE \geq 80 \quad then \quad m_i = 1 \tag{2.4}$$

With m_i being the mortality risk of a patient ($m_i = 0, 1$).

Dimensionality reduction

Dimensionality reduction reduces the number of input features in a dataset without neglecting data integrity.

This technique is often applied in the preprocessing data stage and for data visualisation. Nevertheless, dimensionality reduction can be used in ML models to simplify a classification or regression in order to fit better a predictive model.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), also known as Fisher's LDA, is a method used in statistics, pattern recognition and ML to find a linear combination that separates two or more classes of features. There are two significant objectives in the separating process: the description of group separation and the prediction and allocation of observations to groups. The resulting combination may be used as a linear classifier or, more commonly, for dimensionality reduction before classification [2].

LDA space transformation is obtained by maximising the ratio of between-class variance to within-class variance to reduce data variation in the same class and increase the separation between classes.

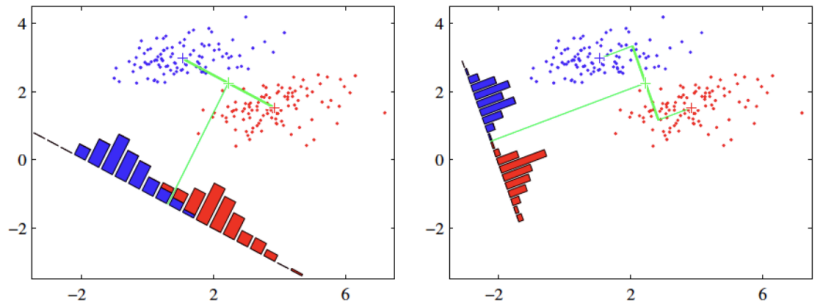


Figure 2.7: LDA Example. Source: [18]

The Figure 2.7 left plot shows samples from two classes, and the histograms that results from the projection onto the line joining the class means [18]. The Figure 2.7 right plot shows the corresponding forecast based on the Fisher's LDA with a considerably improved class separation.

2.2 Related Work

Artificial Intelligence (AI) can have extensive social impacts in the healthcare domain. In a growing industry of smartwatches, fit bands, and devices that constantly gather a collection of health data, the prevalence of Machine Learning (ML) to analyse this information is stronger than ever. Furthermore, ML can be the solution to a series of medical problems such as analysing structured data such as imaging, genetic and laboratory data. Generally, in medical applications, the ML procedures attempt to cluster patient traits or infer the probability from a disease outcome.

Several pieces of research have proven the potential of this expanding area, mainly around three groups of diseases: cancer, nervous system and cardiovascular diseases. For instance, Somashekhar *et al.* [24] demonstrated that IBM Watson (a platform that helps businesses to predict future market outcomes) for oncology would be a reliable AI system for assisting cancer diagnosis through a double-blinded validation study. Esteva *et al.* [24] analysed clinical images to identify skin cancer subtypes. In the neurology field, Bouton *et al.* [24] developed an AI system to restore movement control in patients with quadriplegia.

This section will review some relevant medical studies on the impact of inflammation biomarkers on ACS outcomes. Empirical studies and projects about AI in Cardiology fields will also be reviewed. Note that definitions and theoretical concepts presented in

Chapter 2 are essential to understand related concepts found in this chapter.

2.2.1 Inflammation in Acute Coronary Syndrome

Inflammation is a well-established risk factor for the development of Acute Coronary Syndromes (ACS) since it is a common component of atherosclerotic plaques. Although it also appears to be a component responsible for the sudden origin of coronary instability, less is known about its influence on the outcome of ACS.

Moreover, high values of C-Reactive Protein (CRP) are observed in approximately 70% of patients with severe Unstable Angina (UA) and are correlated with risk for developing an acute infarction in the short term.

Odeberg J et al. [37], researched on the *influence of pre-existing inflammation on the outcome of acute coronary syndrome*. This study intended to learn if blood biomarkers of inflammation were explicitly associated with myocardial infarction (MI) or Unstable Angina (UA) in patients with ACS. The concentrations of plasma biomarkers of inflammation, known as CRP and fibrinogen serum analysed at admission were found to be associated with MI over UA, in ACS occurrences. A stronger association was found for blood cell markers of inflammation, that is, counts of neutrophils, monocytes and thrombocytes were established to be significantly lower in patients with MI compared to those with UA.

In addition, *Maseri et al.* [30] conducted a study to determine whether the inflammatory process detected by elevated CRP levels originates in the coronary arteries or somewhere else in the body. However, ACS are rare events, and the knowledge in such a complex pathogenic scenario is scarce. So, although high CRP values are effectively correlated with the risk of developing an acute infarction in the short term, this study sustained no further conclusions on the origin of such an event.

Both medical studies analysed the influence of inflammation biomarkers on ACS outcome. Despite the established correlation on C-reactive Protein (CRP) high values with the risk for developing an acute infarction in the short term (approximately 70% of patients), the study on fibrinogen serum is an expensive laboratory analysis. Therefore, only a few Hospitals can manage to perform them. Following this, our analysis will be centralised on three inflammation biomarkers, leukocyte count, CRP, and Albumin serum, which are inexpensive and straightforward, so our approach can be an important contribution for a global usage of this biomarkers.

2.2.2 Machine Learning Applications in Cardiology

Machine Learning (ML) might help the diagnostic work-up of cardiovascular diseases. For example, neural networks classifiers could facilitate the detection of patterns of congestive heart failure on chest x-rays.

Some ML studies in the cardiology domain have revolutionised the way that AI is perceived in the medical community. *Mininini et al.* [25] proposed a wearable device for collecting data about normal or pathological gaits for stroke predictions. The data was collected and modelled with hidden Markov models and Support Vector Machines (SVM). The final classifier correctly predicted 90.5% of the subject. Asadi *et al.* [25] compiled a clinical database of 107 patients with acute anterior or posterior circulation stroke who underwent intra-arterial therapy and analysed the data with ANN and SVM classifiers, obtaining a

final accuracy above 70%. Additionally, Zhang *et al.* [25] to better support the clinical decision-making process, proposed a model for predicting 3-month treatment outcome by analysing physiological parameters during 48hours after stroke, using Logistic Regression.

Weng et al. [44] demonstrated that applying ML algorithms to routine medical data significantly improved the accuracy in predicting the first cardiovascular event over ten years when compared to established algorithms with standard American College of Cardiology (ACC) guidelines. Random Forest (RF), Linear Regression, Gradient-boosting Machines (GBRT), and Neural Networks (ANN) were the classifiers employed in this project.

Rezaianzadeh *et al.* [42] predicted the length of stay in the coronary care unit for patients with ACS, using SVM, ANN, and logistic regression classifiers. Afsar *et al.* [5] applied an ANN for the detection of ST-changes, using wavelet transformed ECG signals as input and achieved a positive predictive value of 89.2% on the European Long-Term ST-T Database.

Finnaly, Frizzell *et al.* [42] predicted the 30-day readmission of patients with heart failure using tree-augmented naive Bayesian network and RF algorithm.

Machine Learning Studies on Acute Coronary Syndrome

An empirical research by *Al-Zaiti et al.* [4] studied the possibility of using 12 electrocardiogram (ECG) features for predicting underlying acute myocardial ischaemia (MI) in patients with chest pain. By applying Logistic Regression, Gradient Boosting (GBRT) and Neural Networks (ANN) classifiers, their final fusion model achieved a 52% gain in sensitivity and a 37% gain when compared to experienced physicians. ECG-based clinical decision support tools like this one can help perform early diagnoses by improving clinical outcomes and reducing unnecessary costs in patients with chest pain.

Moreover, *Gibson et al.* [34] conducted a study to evaluate the performance of ML models when compared to traditional risk stratification methods in patients with Acute Coronary Syndrome (ACS) treated with antithrombotic therapy. Data on 24.178 ACS patients were merged from four trials. A super learner ensemble algorithm selected weights for 23 ML models. This algorithm was compared to the traditional TIMI risk score. The super learner algorithm was able to produce a AUC of 0.734, higher than the TIMI risk score (0.489). It was profoundly calibrated on both efficiency and safety produced the highest AUC for prediction of ACS when compared to traditional risk stratification methods. In conclusion, this analysis demonstrated a contemporary application of ML that can guide patient-level antithrombotic therapy treatment decisions.

Noh et al. [36] studied the possibility of using ML algorithms to predict major ACS events that require revascularization on patients presenting early-stage angina like symptoms. They analysed over 20 features relevant to ACS, by applying standard algorithms, SVM and Linear Discriminant Analysis (LDA), and achieved a relevant AUC of 0.860 for the ACS predictions that requires revascularization.

Pieszko et al. [38] developed a study on Predicting Long-Term Mortality after Acute Coronary Syndrome Using Machine Learning Techniques and Haematological Markers. With electronic medical records from 5053 patients with ACS, a the authors trained a ML classifier to predict death during hospitalisation and within 180 and 365 days from admission and compared their results with the Global Registry of Acute Coronary Events (GRACE). They studied haematological markers of inflammation (the ones presented in 2) to understand the usefulness of ML techniques in predicting mortality after ACS events since such features in ML domain have not been studied before. As already validated,

this study defends that haematological markers of inflammation show a strong correlation with ACS outcomes and that they can be successfully incorporated into numerical models designed to support clinical decisions. The Gradient Boosting (GBRT) classifier with haematological markers of inflammation predicted long-term mortality better than the standard GRACE risk score. However, this approach requires further validation.

The potential of solutions like the previously mentioned and the one we intend to develop lie in taking advantage of the easily available inflammation biomarkers in a simple blood test. Eliminating the necessity to enter the results of clinical examinations or past medical history into the model can reduce the patients waiting time and improving the success of prescribed therapeutics. Nevertheless, ML models can never replace physicians reasoning and validation. That's why our approach aims to implement interpretability and personalisation concepts so that both areas can benefit from each others potential.

2.3 Models Evaluation Metrics

This section will present the different measures and concepts that will be considered to evaluate and compare the developed approaches.

2.3.1 Performance measures

Performance measures intend to evaluate the efficiency of a model in the classification process and are calculated considering the comparisons between predicted and actual classes.

These metrics have as basis Figure 2.8, and are further explained below.

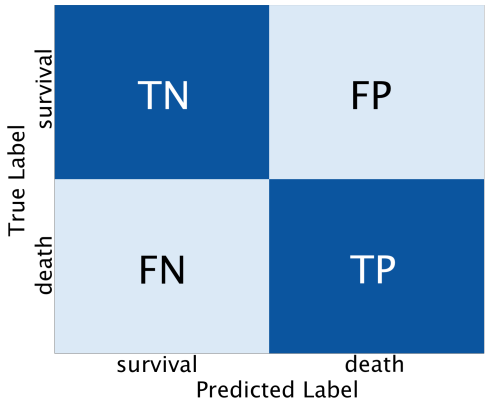


Figure 2.8: Confusion Matrix

1. True Positive (TP) : The predicted class is positive and the real class is also positive;
2. True Negative (TN) : The predicted class is negative and the real class is also negative;
3. False Positive (FP) : The predicted class is positive and the real class is negative;
4. False Negative (FN) : The predicted class is negative and the real class is positive;
5. Negative (N) : Number of instances with negative label;
6. Positive (P) : Number of instances with positive label;

In order to quantify the performance of a classifier, some of the adopted metrics are presented on Table 2.5.

Measure	Formula	Context
Accuracy (ACC)	$\frac{TP+TN}{P+N}$	Percentage of instances correctly classified
Sensitivity	$\frac{TP}{TP+FN}$	Percentage of positive instances classified as positive
Specificity	$\frac{TN}{TN+FP}$	Percentage of negative instances classified as negative
Precision	$\frac{TP}{TP+FP}$	Percentage of instances classified as positive that are really positive
Geometric Mean	$\sqrt{sensitivity * specificity}$	Average set of values rate of return
F-Score	$2 * \frac{precision * sensitivity}{precision + sensitivity}$	Weighted average of the precision and sensitivity

Table 2.5: Performance measures

2.3.2 Comparing Classifiers

For a suitable choice of a classifier, this study will consider two strategies to compare classifiers: using statistical tests of significance and analysing the Receiver Operating Characteristic (ROC) curve [19].

Statistical Tests of Significance

Statistical tests of significance are crucial to analyse how two different classifiers behave when using the same test data. When evaluating their performance, two kinds of errors have to be taken into account: **type I** and **type II** errors, which are defined through the null hypothesis [19] (see Table 2.6).

Significance Test	Null hypothesis true	Null hypothesis false
Reject null hypothesis	Type I error	No error
Do not reject null hypothesis	No error	Type I error

Table 2.6: Types of error

Where,

Null hypothesis is defined by the absence of a relationship between two study variables.

Alternative hypothesis is defined by the existence of a relationship between two study variables.

Type I error means rejecting the null hypothesis when it's true.

Type II error means failing to reject the null hypothesis when it's false.

Several statistical tests are available to compare the performance of different approaches.

Comparison Comparison tests look for differences among group means and can test the effect of a categorical variable on the mean value of some other characteristic [19]. T-tests and ANOVA are parametric tests performed to compare the means of precisely two groups, or more, respectively [52]. However, if the data does not follow a normal distribution, a non-parametric alternative should be applied [50]. The *Shapiro-Wilk* test for normality is one of three general normality tests designed to detect all departures from normality. If the *p-value* from the Shapiro-Wilk test is greater than 0.05, the data is considered to be normal. If not, it significantly deviates from a normal distribution.

Test	**Number of Groups**	**Parametric?**	**Outcome Variable**
Paired T-test	2 Groups	Parametric	Quantitative (groups from the same population)
Wilcoxon Signed-rank Test	2 Groups	Non-parametric	Quantitative (groups from the same population)
ANOVA	2 or more Groups	Parametric	Quantitative 1 outcome
Pearson's χ^2	2 or more Groups	Non-parametric	Quantitative
Mann-Whitney U test	2 Groups	Non-parametric	Quantitative

Table 2.7: Comparison Statistical Tests examples. Based on: [50]

Correlation Correlation tests check if there is a relation between two variables without assuming any cause-and-effect relationship [19]. In statistical terms we use correlation to denote an association between two variables.

	Predictor variable	**Parametric?**	**Outcome variable**
Pearson's r	Continuous	Parametric	Continuous
Spearman's r	Quantitative	Non-Parametric	Quantitative

Table 2.8: Correlation Statistical Tests examples. Based on: [7]

Parametric tests are based on assumptions about the parameters of the population distribution from which the sample is drawn [20].

Non-parametric tests are not based on assumptions. That is, the data can be collected from a sample that does not follow a specific distribution, such as normal distribution [20].

In statistical contexts, the **p-value** is the crucial evaluation parameter. It reflects the significance level of a hypothesis test and whenever a null hypothesis is accepted or rejected when formulated before the study performance. For example, a p-value of 0.05 means a 5% probability that the results are due to random chance [46]. In consequence, under the Null hypothesis, **the smaller the p-value, the more significant the evidence that you should reject the null hypothesis.**

Moreover, confidence intervals definition also needs to be introduced in order to understand the p-value. A confidence interval gives information on the interval in which the experience value is considered relevant if the experiment is replicated with the same initial conditions and parameters.

Values of p	Inference
p >0.10	No evidence against the null hypothesis
0.05 <p <0.10	Weak evidence against the null hypothesis
0.01 <p <0.05	Moderate evidence against the null hypothesis
0.05 <p <0.01	Good evidence against the null hypothesis
0.001 <p <0.001	Strong evidence against the null hypothesis
p <0.001	Very strong evidence against the null hypothesis

Table 2.9: Interpretation of p-value. Based on: [46]

Receiver Operating Characteristic Curve

In clinical studies, the ability to correctly predict the output is often determined through Area Under ROC Curve (AUROC). AUROC is a performance measurement for a classification problem at various threshold settings.

As presented in Figure 2.9, ROC is the probability curve, and the AUC represents the likelihood of a model to distinguish observations from two classes [35]. In sum, it is a plot of 1− Specificity (in the x-axis) versus the Sensitivity (in the y-axis) for several threshold values (see Section 2.3.

The higher the AUC, the better the model predicts true positives classes and true negatives in a binary problem. An AUC of 0.5 represents a test with no discriminating ability, while an AUC of 1.0 represents perfect discrimination [56]. Consequently, a greater area implies a more significant result.

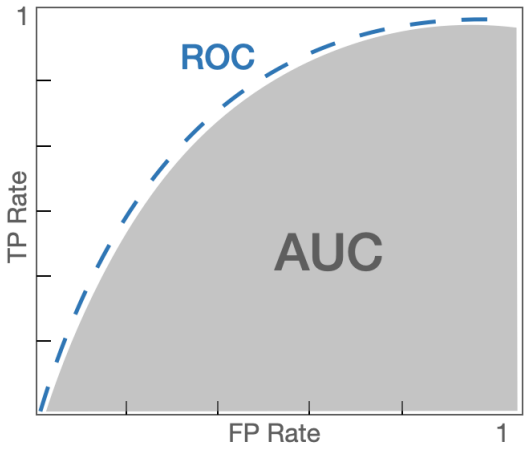


Figure 2.9: ROC curve plot explanation

2.3.3 Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) is a research field that aims to produce more explainable models and maintain high-performance values. Explanations in AI support understandability and transparency to enable greater trust toward AI-based solutions [41]. Therefore, Explainable Artificial Intelligence (XAI) is acknowledged as a crucial feature for satisfying the fundamental rights of AI users related to AI decision-making.

In the healthcare domain, some reports have revealed that the decisions and recommendations of such systems may be biased. Due to that, there is a need for interpretability to ensure that the developed methods are free from data bias in assessing different populations.

Interpretability

Interpretability is the degree of human understanding. By explaining the reasoning behind predictions, interpretable machine learning systems give users reasons to accept or reject predictions and recommendations. Higher interpretability means easier understanding and explanation of the predictions in the clinical field, guiding physicians for personalised decisions that can lead to a higher quality of service.

Interpretability concepts can be particularly challenging to implement in healthcare domains due to medical and ethical requirements, laws, and regulations [3]. Nonetheless, there is an absolute need for interpretability in ML clinical models. The lack of interpretability can have adverse or even life-threatening consequences, for example, a scenario where the insights from a black box model are operated without the recognition that the predictive model is not prescribed.

While the application of ML methods to healthcare problems is inevitable given the complexity of analysing vast amounts of data, it is essential to standardise the expectation for interpretable ML in this domain [3].

Personalisation

Machine learning-based personalisation provides a more scalable and accurate way to achieve unique experiences for individual users. Nowadays, this concept is visible in almost every online platform, in product or content recommendations. When employing ML personalisation, our priority is not only to maintain but to optimise the user experience [27]. Platforms like Amazon and Netflix are well-known examples of personalisation in the recommendation systems domain that deliver these one-to-one experiences.

In the healthcare domain, personalisation intends to estimate the acceptance of the baseline rules when applied to an individual (patient), avoiding the principle that one fits all. When looking at ACS diagnoses, physicians typically select the most relevant rules for each patient, considering their experience, which is precisely the ML based personalisation goal for their models.

This page is intentionally left blank.

Chapter 3

Methodology

In order to answer to our research question on the influence of inflammation biomarkers on Acute Coronary Syndromes (ACS) outcomes, we implemented three different models to perform a direct comparison: an implementation of GRACE Risk Score (See 2.1.3), a ML classifier, using as features the GRACE risk factors and a ML classifier considering GRACE risk factors and the selected inflammation biomarkers (C-Reactive Protein (CRP), Leukocyte Count (WBC) and Albumin serum) and statistically analysed their performances.

Considering the previous results, and to evaluate our methodology that relies on an innovative ML-based method that keeps interpretability and personalization, we considered three different methods to perform a direct comparison: the standard GRACE approach, a standard ML-based classifier, and our developed system. Moreover, the implemented ML systems make use of several classifiers in order to understand if several ML structures can be used without loss of generality.

The performance of our algorithms can be quantified with metrics from Table 2.5. However, accuracy is not a suitable performance measure when a class is eminently bigger than the other. With unbalanced classes, metrics like AUROC and Geometric Mean (GM) should be computed instead, since they contemplate both sensitivity and specificity measures on their calculations. Confusion matrix can also contribute to understanding which and how well classes are being correctly predicted (see Figure 2.8). True negative (TN) and false positive (FP) represent the number of survival cases ($m_i = 0$) that have been, respectively correctly ($\hat{m_i} = 0$) and incorrectly identified ($\hat{m_i} = 1$). True positive (TP) and false negative (FN) represent the number of death cases ($m_i = 1$) that have been, respectively correctly ($\hat{m_i} = 1$) and incorrectly identified ($\hat{m_i} = 0$). Moreover, to evaluate our results consistently, our dataset was divided into 70% for the training process and 30% for testing.

We develop our code in Python language by recurring to Interactive Python. It has a robust interactive Python shell and a Jupyter kernel to work in Jupyter notebooks and includes ample support for ML, data analysis and manipulation through several Python libraries.

3.1 Impact of Inflammation Biomarkers on Acute Coronary Syndrome Outcome

As previously specified, one of the primary objectives of this work is to understand if the introduction of inflammation biomarkers as risk factors in the currently accepted models can somehow accomplish better results in the mortality prediction domain. To address this objective, we applied the well-established GRACE risk score performance as the benchmark to establish a comparison from the ML-based algorithms, since it has been proved that the GRACE model has a good ability to assess a patient's risk.

In order to overcome our imbalanced data issue (9% mortality targets), and considering that we also intend to expose the excellent performances that typical ML algorithms can obtain, we used an oversampling technique denominated Synthetic Minority Oversampling Technique (SMOTE). There are two techniques to deal with this subject: under-sampling and oversampling. However, in most cases, oversampling is preferred over under-sampling methods. Under-sampling can remove instances with vital information and corrupt our model [54].

SMOTE generates synthetic samples from the minority class (i.e. death class) by randomly selecting one or more of the K-nearest Neighbours (KNN) for each instance in the minority class (we generated 557 mortality samples plus 1544 existent from our dataset). After the oversampling process, the data is reconstructed, and classification models can be applied to the processed data [54].

3.1.1 Global Registry of Acute Coronary Events Risk Score

GRACE risk score was developed in the late 1990s and early 2000s, and in 2010, the United Kingdom National Institute for Health and Clinical Excellence (NICE) recommended its international adoption.

As previously mentioned in Chapter 2, this scoring system was created on a prospective observational registry enrolling patients with the full spectrum of ACS and was conceived to study an unbiased population from multiple geographic locations. It includes components as age, heart rate, systolic blood pressure, Killip class, cardiac arrest, ST-segment deviation, serum creatinine, and initial cardiac troponins, which are all well-established risk factors in ACS.

The algorithm registers every patient's risk factor values and deduces a final risk score (0-383), calculating the sum from every variables.

Finally, study patients were divided into three groups according to their final risk score [14]:

1. **Low-risk** group: GRACE score ≤ 108 for NSTEMI and UA diagnoses and GRACE score ≤ 125 for STEMI diagnoses.

2. **Intermediate-risk** group: GRACE score $109 - 140$ for NSTEMI and UA diagnoses and GRACE score $126 - 154$ for STEMI diagnoses.

3. **High-risk** group: GRACE score ≥ 141 for NSTEMI and UA diagnoses and GRACE score ≥ 155 for STEMI diagnoses.

3.1.2 Standard Machine Learning Based Approach with GRACE Risk Factors

To evaluate the impact of the inflammation biomarkers in ACS outcomes, we first implemented a ML classifier, aiming to extract accurate predictions about cardiovascular mortality. This model will exclusively contain information relative to the GRACE established risk factors.

In order to determine the most powerful classifier, we first conducted a grid search. A grid search is a process that performs hyper-parameter tuning to determine the optimal values for our model [28]. The finest model was a Random Forest (RF) classifier with 200 estimators, 2 minimum sample leaves and 2 minimum samples split.

As presented in Figure 3.1, the developed model is trained to employ a data-driven supervised method, based on the available training dataset, where the inputs are the patient's GRACE risk factors (X). The output is the actual mortality outcomes (t_i) with N being the number of instances (patients), and the target t the respective mortality class $t = survival, death$.

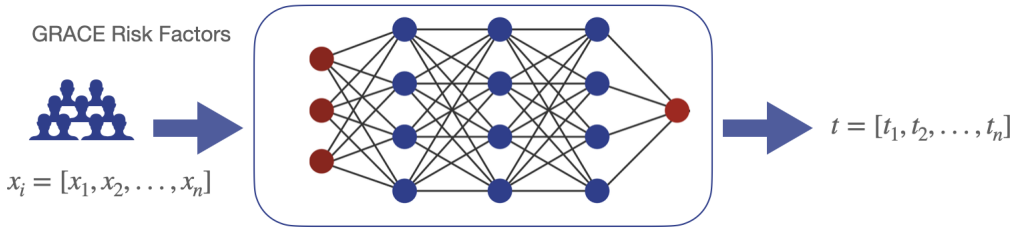


Figure 3.1: ML-based Scheme Using GRACE Risk Factors

3.1.3 Standard Machine Learning Based Approach with GRACE Risk Factors and Inflammation Biomarkers

To perform a direct comparison, we implemented the same classifier, i.e., a Random Forest (RF) with 200 estimators, 2 minimum sample leaves and 2 minimum samples split, but also including the three referenced biomarkers of inflammation (CRP, WBC and albumin serum), as given in Figure 3.2.

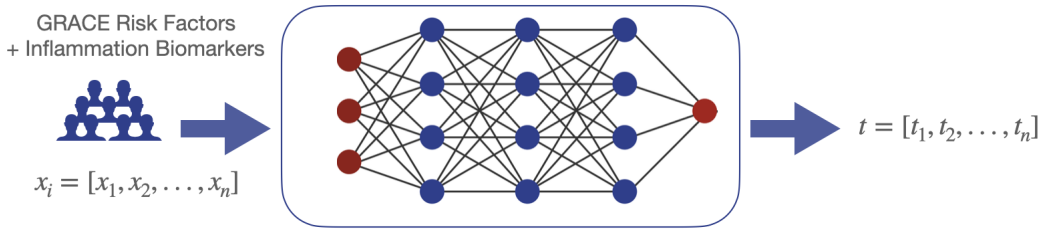


Figure 3.2: ML-based Scheme Using GRACE Risk Factors and Inflammation Biomarkers

Although typical ML structures can reach remarkable performances (GM and AUC metrics), they usually fail on keeping interpretability, which is an essential requirement for the acceptance of decision support systems. In fact, the mentioned ML methods are black-box models; therefore, they are not directly explainable or interpretable. In addition, they follow the policy "one fits all" (i.e. the model is applied in identical ways to all subjects) without be the ability to understand which of the used features can be of utmost use.

Considering all that, in daily clinical practices, these classification models can not be helpful. Consequently, these vulnerabilities forced the development of different ML-based approaches.

3.2 Proposed Machine Learning Based Approach

In order to develop a new risk assessment tool that physicians can benefit from, we started by looking at the GRACE model and its risk factors. As previously mentioned, although ML models can provide notable results in terms of performance, i.e., GM and AUC metrics, they neglect some essential requirements for the acceptance of medical decision support systems: interpretability and personalization. It is impracticable to develop a black-box ML approach for clinical use, even if remarkably good.

Briefly, our approach is composed of three main phases. The first one is constructing several interpretable rules considering clinical knowledge. The second stage is to introduce a ML-based system to identify the more appropriated subset of rules for each patient. Finally, the third phase combines the previous subset of relevant rules to estimate each patient cardiovascular mortality risk. The fundamental point of our approach is the paradigm of how a ML model is applied. Basically, it intends to simulate a physicians reasoning. With his clinical knowledge (set of rules), facing specific characteristics of a patient (personalised choice of rules), the clinician will select and apply only the most suitable rules to achieve a diagnosis.

3.2.1 Rules Structure and Definition

As aforementioned, the initial point of our approach consisted of specifying a set of rules that can perform a risk prediction or a diagnose.

A rule can be a simple binary association based on a risk factor, as demonstrated in 3.1. If a patients Killip value is 4, then the patient is expected to die.

$$if\ Killip = 4\ then\ m_r = 1 \tag{3.1}$$

Despite that, more complex rules that combine two or more risk factors will can also be considered.

The creation of rules can be based on two different strategies: knowledge and data-driven. **Knowledge-driven** methods use the experience of clinical professionals and literature to obtain the rules. In **data-driven** scenarios, the available data is used. Nevertheless, all the developed rules need to reflect clinical regulations to be directly interpretable and accepted. Consequently, our set of baseline rules were externally validated by a physician.

3.2.2 Rules based on Virtual Patients

One of the applied data-driven methods of our approach was based on the idea of virtual patients collected through a clustering strategy.

More specifically, considering that patients with similar symptoms and characteristics can be grouped, we determined two centroids based on each established risk factor (GRACE risk factors plus inflammation biomarkers). Each centroid designates one class: patients

who survived and patients who died, and they represent the average characteristics from the two virtual patients groups.

With both centroids calculated, the distance of a new patient to them can be applied to measure his classification, i.e., a patient closer to the *death centroid* is expected to die. Otherwise, if it's closer to the survival centroid, the patient is expected to survive. Actually, this approach in some way also resembles a physicians reasoning. The clinician uses their past experiences, recalling similar cases and standard guidelines and applies them to the subject to diagnose.

Figure 3.3 demonstrates the core of this clustering methodology with the combination of two risk factors. In this specific case, since the patient's distance to the survival centroid (d_0) is more considerable than their distance to the death centroid (d_1), the patient's mortality risk will be defined as death.

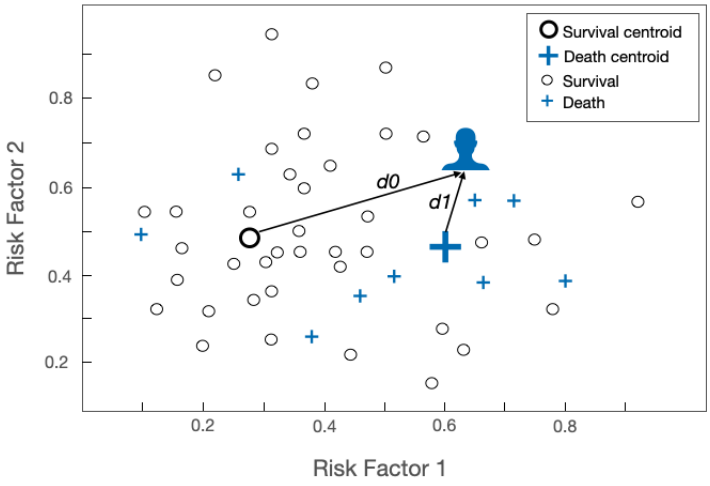


Figure 3.3: Virtual Patient Clustering example with the combination of 2 Risk Factors.

Considering that ACS risk factors possess different scaling systems (for example, Killip scale is from 1 to 4 and Age is a continuous variable), the distance from a patient to the centroids is determined through a normalised distance d_n, presented in 3.2.

$$d_n = \frac{d_0}{d_1 + d_0} \tag{3.2}$$

Where d_1 represents the Euclidean distance of a patient to the death centroid and d_0 represents the patient's distance to the survival centroid.

This distance varies between $d_n = 0$ (the case corresponds to the virtual patient representing the survival group) and $d_n = 1$ (the case corresponds to the virtual patient representing the death group).

Therefore, d_n can be transformed into a rule for the classification task, as presented in 3.3.

$$if\ d_n \geq L\ then\ m_r = 1 \tag{3.3}$$

Where L represents a threshold that places the individual into the death or survival groups. The standard case stipulates L as the mean distance between the two centroids, therefore

$L = 0.5$. However, the threshold for each risk factor can be optimised, provided that medical guidelines are not violated.

Finally, we implemented two types of rules: individual rules, where centroids are computed considering only one risk factor (3.4) , and rules that combine the knowledge of two risk factors (3.5), and consequently, our centroids analyse two variables.

$$if\ d_n(RF_x) \geq L\ then\ m_r = 1 \tag{3.4}$$

$$if\ d_n(RF_x \wedge RF_y) \geq L\ then\ m_r = 1 \tag{3.5}$$

Threshold optimisation

Although a threshold value of 0.5 ($L = 0.5$) can be acceptable, the maximum accuracy for each rule is not always assured. Therefore, the computation of an optimal L for each risk factor (or for the combination of 2) was performed, and Geometric Mean (GM) was the selected measure to maximise our model performance. Moreover, since we are dealing with a binary classification task (survival and death), formulas for specificity and sensitivity can be rewritten as 3.6 and 3.7.

$$SE = \frac{1}{T_1}(t' \cdot \hat{t}) \tag{3.6}$$

$$SP = \frac{1}{T_0}((1 - t') \cdot (1 - \hat{t})) \tag{3.7}$$

In sensitivity formulation (3.6), t' represents the transposed target-risk and $\hat{t}$, our calculated-risk vector with the rules classification process. $T0$ and $T1$ represent, respectively, the number of survivals and deaths from the target vector t. Basically, the dot product between the vectors t' and $\hat{t}$, quantify the number of simultaneously equal deaths in both vectors.

To quantify specificity, a similar operation is performed, by computing the values that are simultaneously equal to survival occurrences.

Considering GM formula referred in 3.8, general formulation for our optimal GM can be understood as the maximum value from the GM vector computations.

$$GM = \sqrt{SE \cdot SP} \tag{3.8}$$

In conclusion, the maximisation of each rule can be described as 3.9.

$$if\ dn_i \geq L_i\ then\ \hat{t_{ij}} = 1 \quad i = 1, ..., M \tag{3.9}$$

Where dn_i is the i^{th} risk factor ($i = 1, ..., M$) normalised distance and L_i the respective threshold. The outcome t_{ij} presents the estimated mortality risk for the patient i, by using the j rule.

Fisher's Linear Discriminant Analysis

Fisher's Linear Discriminant Analysis (LDA) was another implemented methodology for the rule development process that combines knowledge from two risk factors. As mentioned in Chapter 2, LDA is a well-known classification technique whose primary purpose is to separate samples of distinct groups by transforming the data to a particular optimal space for distinguishing between the classes. So, instead of calculating the optimal threshold for the two risk factors, with *Fisher's* LDA we determined the optimal axis that separated both classes and estimated a new d_n referent to the designed axis.

As presented in figure 3.4, LDA process considers our patient's combination of 2 risk factors and projects the data points in order to maximise the separability of the known categories of our target ($0 = survival;\ 1 = death$).

The process of finding the optimal axis is performed in two different steps: maximising the difference between the means of our mortality target (m_s *and* m_d) and minimising the variance (s^2). In our approach *Fisher's* LDA produced a baseline set of optimal rules based on a new axis that maximises class separation. Moreover, the rule evaluation method was similar to the one presented in the threshold optimisation process, based on calculating GM metrics for every established rule.

This method of finding the optimal axis that separates our target class produced an optimal baseline set of rules that combined two risk factors. Furthermore, the rule evaluation process was similar to the one presented in the threshold optimisation process, based on calculating GM metrics for every produced rule.

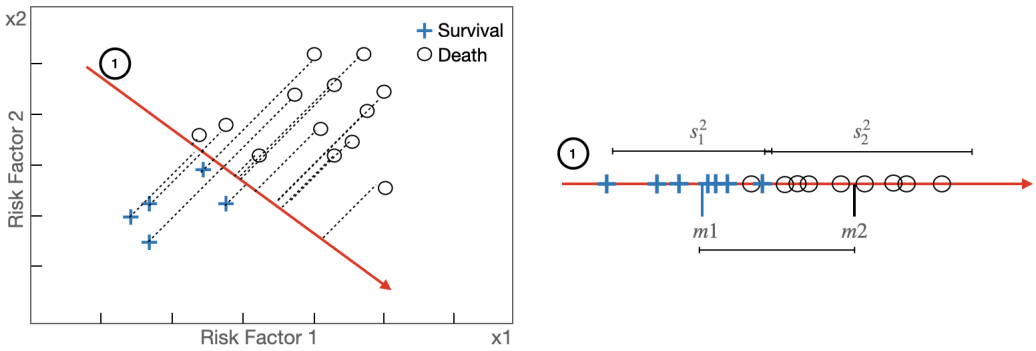


Figure 3.4: Fisher's LDA optimal class separation

3.2.3 Rule Personalisation

As aforementioned, this particular step is the critical point of our strategy. In some way, this method simulates the logic of a professional by not selecting the same rules for every patient. Instead, he/she determines the subset of rules to perform a diagnosis, considering each patient's specific characteristics. Only the rules which are likely to be suitable for a patient are going to be selected. In sum, the interpretability of the model is preserved.

Prediction of Correct Rules

With the set of rules already defined, the second step from our approach was to train a ML classifier. Instead of using the mortality data as our target (similar to Chapter 3), our ML model applied the computed rule correctness from every patient. The computation of each rule correctness is simple and can be expressed as a yes or no response ($correct = 0$; $incorrect = 1$) since the patient's outcome, and the output of each rule are already known.

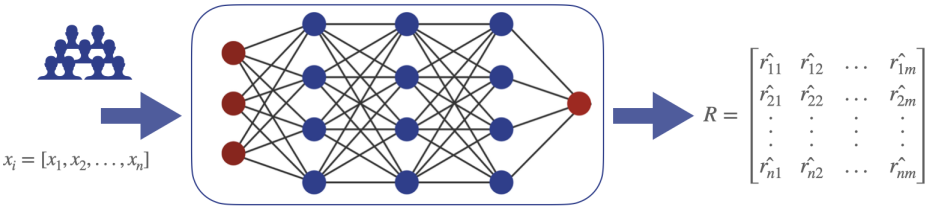


Figure 3.5: Training our ML model to estimate the correctness of the rule-set.

For example, if we consider a patient with an albumin serum of 30 that dies and the previously established rule defines that $if\ albumin \leq 25\ then\ \hat{t} = 1$, we conclude that this rule is not correct for this patient and should not be applied.

Therefore, our ML training used the risk factors as the input, and the target output is the correctness of the rules. In other words, the ML models do not learn how to classify the risk of a given patient but are trained to determine if a rule is corrected or not when applied to a specific patient.

New Patient Rules Acceptance Process

Next to the training process and given a new patient (p_i), the ML output ($\hat{r_{ij}}$) defines the correctness from each of the base rules (j) by establishing their acceptance as a binary output. If ($\hat{r_{ij}}$) $= 1$, it means that the rule was accepted for the patient (i) and should be applied in the final mortality risk assessment. On the other way, a value of ($\hat{r_{ij}}$) $= 0$ means that the correspondent rule (j) is considered not correct and, therefore, not useful for our final risk calculation.

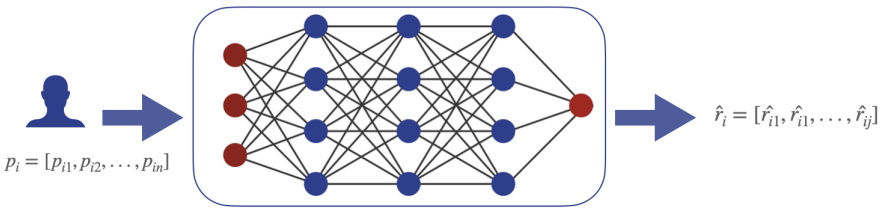


Figure 3.6: Estimating each rule correctness for a specific patient.

Finally, the final mortality score was estimated, taking into account only the subset of previously accepted rules. For a better understanding, let's consider the existence of four different rules ($r_j = r_1, r_2, r_3, r_4$) presented in 3.10, and a patient who survived ($t_i = 0$), characterised by a normalised distance of $d_n = 0.4$ to the *Killip* Rule, a $d_n = 0.4$ to the *Age* rule, a $d_n = 0.5$ for the *Albumin* rule and a $d_n = 0.3$ for the *Creatinine* rule. Table

3.1 present an example from the rules output and the respective ideal acceptance for this patient.

$$\begin{aligned} r1 : if\ d_n(Killip) \geq 0.5\ then\ \hat{t_{i1}} = 1 \\ r2 : if\ d_n(Age) \geq 0.7\ then\ \hat{t_{i2}} = 1 \\ r3 : if\ d_n(Albumin) \geq 0.4\ then\ \hat{t_{i3}} = 1 \\ r4 : if\ d_n(Creatinine) \geq 0.4\ then\ \hat{t_{i4}} = 1 \end{aligned} \tag{3.10}$$

	Rule 1	Rule 2	Rule 3	Rule 4
Target (t_i)	0 (survival)			
Rule Output ($\hat{t_{ij}}$)	1	0	1	0
Rule Acceptance i.e. ML Output ($\hat{r_{ij}}$)	1	1	0	1

Table 3.1: Example of the rule estimation and their acceptance process. An ideal scenario.

Each output rule ($\hat{t_{ij}}$) is directly evaluated by using the patient established variables and the statement of the 4 rules.

The rule acceptance ($\hat{r_{ij}}$), which was provided from our ML model, defined which rules were helpful in the risk assessment of our patient. Regarding our example, our ML model considered r_1, r_2 and r_4 as useful rules. Therefore, these are going to be the used rules to estimate the patient mortality risk.

Finally, in order to estimate the patient mortality risk, we used a majority voting score, which is obtained from the most common output of the accepted rules. Consequently, these patient risk will be survival ($\hat{t_i} = 0$), obtained from $\hat{t_{ij}} = [1, 0, 0]$.

3.2.4 Mortality Prediction

The third phase combines the accepted rules in the previous step ($\hat{r_{ij}}$) and infer an estimation for the mortality patients risk. Our proposed mortality score is obtained for each patient (i) according to Equation 3.11, using the subset of T applicable rules, previously established as the condition $\hat{r_{ij}} = 1$.

In sum, a patient mortality score is calculated by the ratio between the number of accepted rules that suggest his/her mortality ($\hat{r_{ij}} = 1 \wedge \hat{t_{ij}} = 1$), and the number of all accepted rules (T).

$$m_r = \frac{1}{T} \cdot \sum_{j=1}^{T} \hat{r_{ij}} \cdot \hat{t_{ij}} \tag{3.11}$$

The final prediction of ($0 = survival\ ; 1 = mortality$) is defined in 3.12.

$$m_p = \begin{cases} death & m_r \geq 0.5 \\ survival & m_r < 0.5 \end{cases} \tag{3.12}$$

Figure 3.7 presents a global scheme of our ML approach. From the set of interpretable rules derived in Phase 1, our ML model (Phase 2) proposes to estimate their acceptance, avoiding the principle that *one fits all*, and, therefore, assures personalisation. In addition, the interpretability of the rules is fully preserved since our approach does not modify the rules, only elects a subset of them. The last phase (Phase 3) determines each patient prediction.

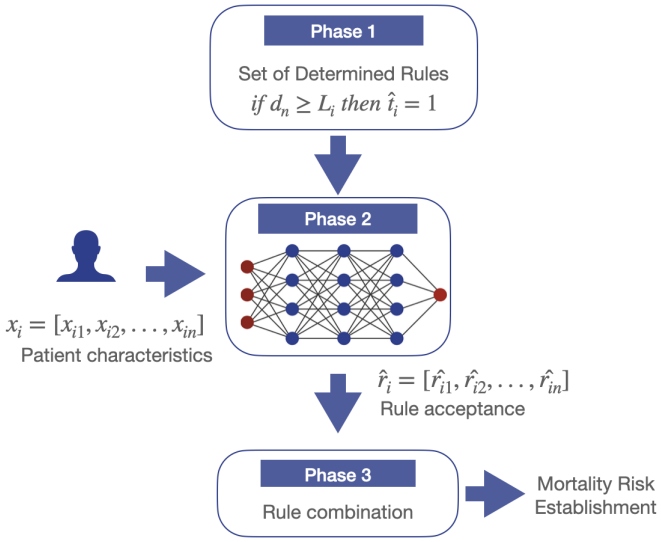


Figure 3.7: Overview from our approach

Chapter 4

Experimental Analysis

This chapter describes the used dataset and the relevant experiments performed throughout this dissertation. Two main groups of experiments were performed, explained in further detail in the following sections, including the preprocessing involved.

The first experiment intends to prove the usefulness of ML methods on decision support systems and at the same time analyse if the inflammation biomarkers can help to improve ACS prediction outcomes. It consists on comparing the traditional GRACE model with a ML-based model that only uses GRACE risk factors and a ML-based model that uses GRACE risk factors and inflammation biomarkers as features.

The other experiment has the purpose of validating our proposed model. To accomplish that, our experiment will rely on the comparison of three different models, with no generated synthetic samples for the minority class. The traditional GRACE model, ML-based model that applies GRACE risk factors and inflammation biomarkers and our proposed ML-based approach that introduces interpretability and personalisation.

4.1 Dataset

This section describes in more detail the patient's characteristics from the provided database. The dataset includes information such as baseline characteristics, medication and the presence of comorbidities on 1544 patients admitted between 2009 and 2016 at the *Hospital dos Covões Cardiology ICU*, containing patients from all spectrum of ACS diagnoses (STEMI, NSTEMI and UA). Prior approval was obtained to conduct this study, and the patient data were anonymized.

4.1.1 Pre-processing

Data on all-cause death or survival and the exact date of death was also provided. Patients who had incomplete records on death information were excluded from the study (75 patients). The percentage of all missing values can be viewed in detail in Table 4.1.

Idade	Missing Values Percentage (%)
Systolic Pressure Admission (PAS)	1.43%
Cardiac Frequency Admission (FCA)	1.56%
Leukocyte count Admission	0.6%
Albumin	58.2%
Troponin Admission	0.54%
C Reactive Protein Admission (CRP)	0.75%
Maximum Creatinine	1.50%
Death	5.17%

Table 4.1: Number of missing values from the CHUC dataset

Missing data is inherent in critical analysis that needs to be immediate, particularly in the medical field. Moreover, the high percentage of missing values on albumin serum can be related to albumin discriminative power on ACS being a recent finding, and most laboratory reports do not include this study variable. While the most natural way to handle missing data is to discard the subjects, it is not suitable for our work due to the number of missing values that could create biased estimations.

However, compared with the deletion of missing data, imputation is a more optimal technique. Imputation techniques make the consequence of statistical analyses and data processing more accurate and reliable. Considering all that, to overcome our limitations, we employed KNN imputation algorithm. KNN finds the K nearest neighbours (in our approach, we used five neighbours) with missing data from complete samples and then fill in the missing data based on the mean.

Outliers existence will cause more significant deviations when estimating missing values. Therefore, the only existent outlier was removed from our study (present in WBC) to reduce their influence before imputation process.

4.1.2 Exploratory Data Analysis

This subsection describes our data characteristics and some relevant analyses on influential risk factors and inflammation biomarkers in detail.

Dataset Baseline Characteristics

The mortality rate within six months from admission was 10.22% (n = 150). The cardiovascular risk death analysis is typically a highly imbalanced scenario, and our results follow the same reality. The baseline clinical features and laboratory test results according to survival or death status are presented in Table 4.2.

For computing elevated cardiac markers, the troponin value was used. A 34 g/L or higher value has been defined as a positive occurrence (our clinical partner established the threshold). For the cardiac arrest at admission variable, it was considered Killip class at 4.

Statistical analyses were performed using the RStudio Software. The *Shapiro-Wilk* test was used to test the distribution of the variables for normality. However, none of the analysed variables followed a normal distribution.

The chosen measures of central tendency were mean and interquartile ranges (IQR). Mann-Whitney U test [51] and Pearson χ^2 tests were performed to compare survival-death continuous and categorical (STEMI and Killip) data, respectively [14].

All risk factors suggests that there is a difference between mortality and survival classes (p-value ≤ 0.05).

Risk Factor	**Survival Mean (n = 1319)**	**Survival IQR Range**	**Death Mean (n = 150)**	**Death IQR Range**	**p value**
Age	66.87	(57-77)	77.49	(74.0-83.0)	<0.001
Systolic Pressure Admission (PAS)	135.05	(119.5-150)	123.07	(104.25-140.75)	<0.001
Cardiac Frequency Admission (FCA)	75.73	(64-85)	84.02	(70.0-97.75)	<0.001
Troponin Admission	19.47	(0.11-9.61)	29.58	(0.80-24.77)	<0.001
Maximum Creatinine	125.71	7180	241.34	(112.35-310.35)	<0.001
STEMI	0.36	(0.0-1.0)	0.46	(0.0-1.0)	0.009
Maximum Killip	1.35	(1.0-1.0)	2.53	(1.0-4.0)	<0.001
C-Reactive Protein Admission (CRP)	2.28	(0.5-1.6)	5.60	(0.70-6.77)	<0.001
Leukocyte count Admission (WBC)	10186.71	(7180.0-11450.0)	11576.80	(8100.0-13250.0)	<0.001
Albumin	36.11	(34.0- 38.2)	33.36	(31.0-35.95)	<0.001

Table 4.2: Dataset baseline characteristics.

In addition, the descriptions from the patient's medical records were analysed in order to identify patients who had an ST-segment elevation (n = 543). From the 1469 qualified patients, 36.97% (n=543) had a STEMI, 32.74% (n=481) a NSTEMI and 30.29% (n=445) an UA diagnose (see Figure 4.1). The associated mortality to STEMI diagnoses are 4.77% (n=70), 3.33% for NSTEMI (n=49) and 2.11% (n=31) for UA.

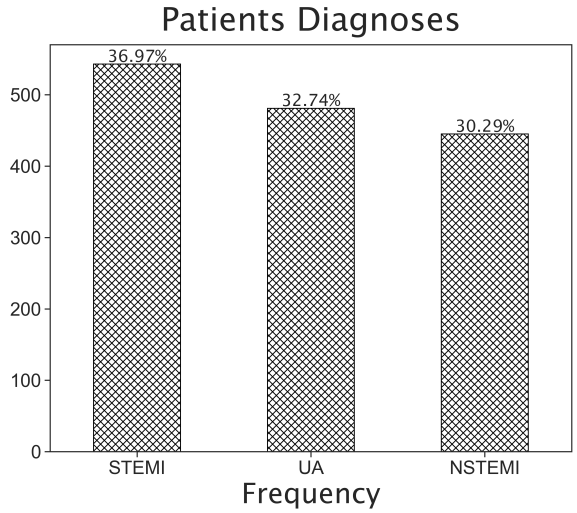


Figure 4.1: Dataset ACS Diagnoses Frequency

Moreover, the dataset contains 36.49% (n=536) high-risk patients, 34.65% (n=509) intermediate-risk patients and 28.86% (n=424) low-risk patients.

As presented in plot 4.2, from the total mortality rate (10.22%), 8.31% are from high-risk patients, 1.43% from intermediate-risk patients and only 0.48% are from low-risk patients. From this plot analysis, we can understand that risk stratification on a patient can give both physicians and patients some guidelines and trust to assume a prognosis.

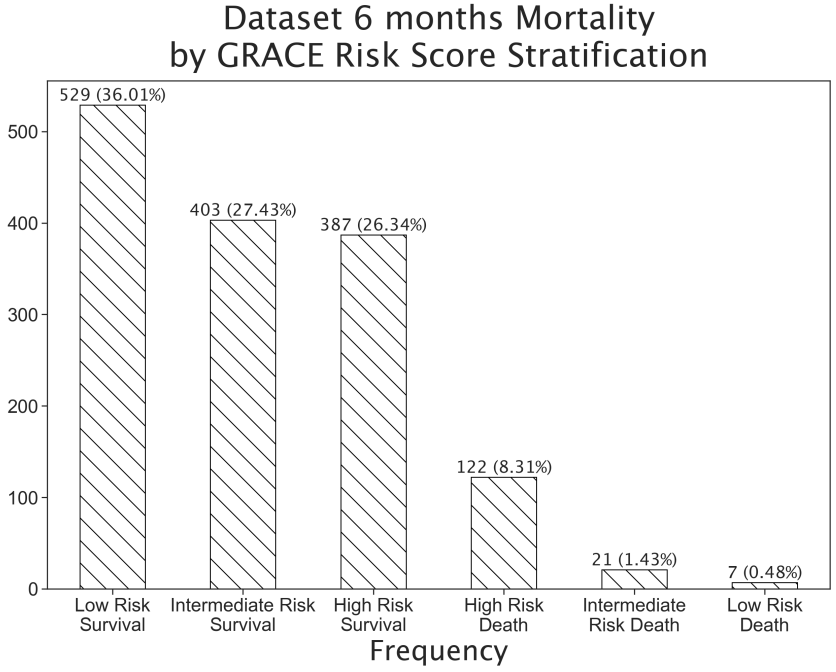


Figure 4.2: 6 months mortality distribution by Grace risk stratification

Inflammation Biomarkers

To understand the occurrence of an inflammation process in our patients, we analysed each biomarker individually. Concerning the 1469 patients, 1336 (90.9%) presented at least one inflammation biomarker and 130 death records, and 68(4.62%) recorded the three markers, with 8 death records. From the 150 deaths, 88 (58.67%) died with hypoalbuminemia (low albumin), 83 (55.30%) with high CRP levels and 72 (48.00%) with large WBC counts.

The GRACE risk stratification was also analysed in common with inflammation biomarkers. The respective information plots can be seen in Appendix 5.

4.1.3 Selected Risk Factors and Validation Strategy

In conformity with our principal research question on the influence of inflammation biomarkers on ACS outcomes, the selected risk factors were the GRACE established risk factors, plus the inflammation biomarkers (C-Reactive Protein (CRP), Leukocyte count (WBC) and Albumin Serum).

4.1.4 Validation Strategy

The methodology presented in this work was validated through the analysis of 500 runs of each experiment, wherein each run, 70% data was used to train the model a 30% to test, except for the traditional GRACE method. To perform a coherent analysis, the model was only applied to 30% of the population.

In order to assess the performance of the proposed methodology, the Geometric Mean GM and Area Under the Receiver Operating Characteristic curve AUC were considered.

4.2 Impact of Inflammation Biomarkers on Acute Coronary Syndrome Outcome Results

This section will specify the obtained results on our research question that intends to determine if inflammation biomarkers have discriminative power when predicting an ACS event outcome.

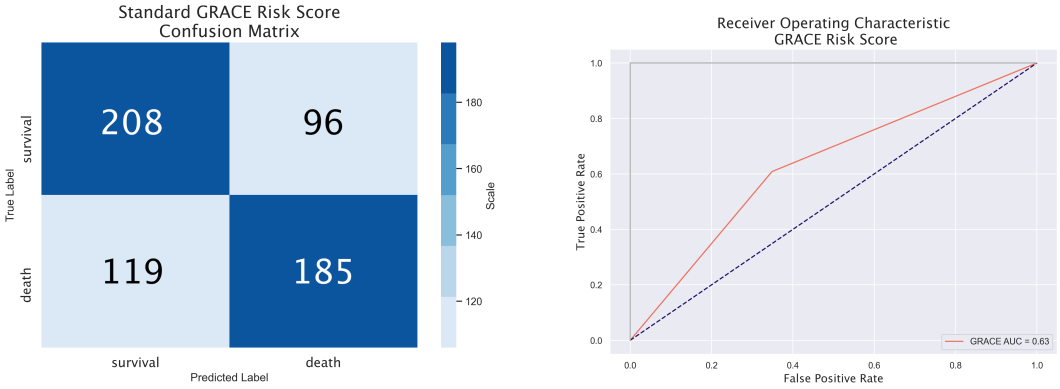
4.2.1 GRACE Traditional Algorithm

The figure 4.3 shows the confusion matrix, and the ROC plot respectively, from the conventional GRACE approach, with no ML. Table 4.3 presents in more detail the aforementioned chosen performance metrics.

Considering our AUC value (0.63%), we can conclude that our results are in conformity with regular GRACE AUC values (normally are values from 60% to 70%). Therefore, we can assume that our data is representative.

Sensitivity	Specificity	Geometric Mean	AUROC
0.61	0.64	0.62	0.63

Table 4.3: Obtained results using the conventional GRACE method



(a) Confusion Matrix (b) ROC curve plot

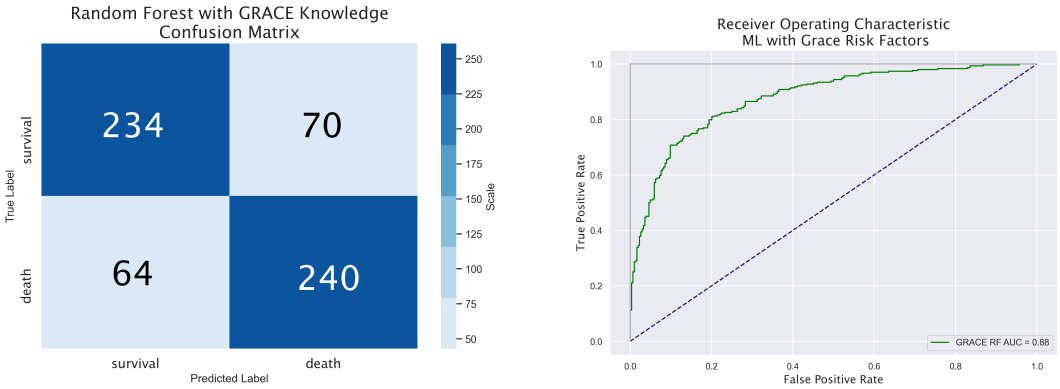
Figure 4.3: Metrics from Conventional GRACE Risk Score

4.2.2 Machine Learning Algorithm with GRACE Risk Factors

Figure 4.4 presents the respective Confusion Matrix, the ROC curve and the detailed performance of this classifier is detailed on Table 4.4.

Sensitivity	Specificity	Geometric Mean	AUROC
0.75	0.85	0.80	0.88

Table 4.4: Obtained results using the conventional GRACE method



(a) Confusion Matrix (b) ROC curve plot

Figure 4.4: Metrics from ML-based Approach with GRACE Risk Factors

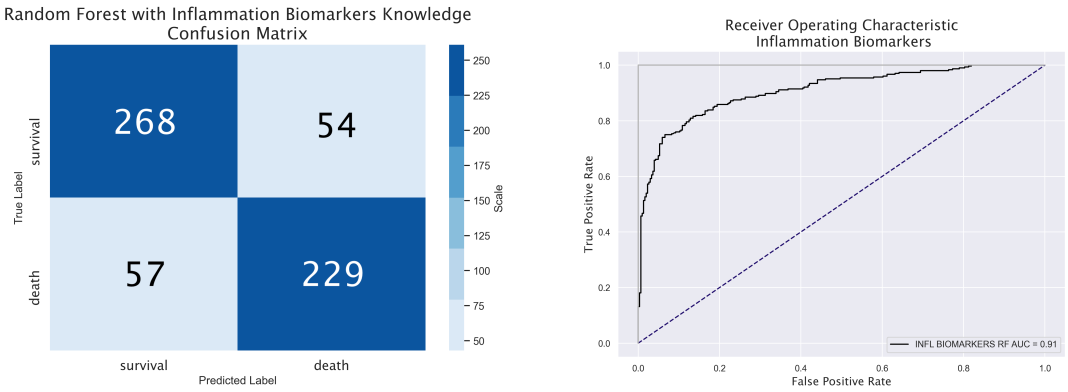
4.2.3 Machine Learning Algorithm with GRACE Risk Factors and Inflammation Biomarkers

Finally, we implemented the same classifier, i.e., a Random Forest (RF) with 200 estimators, 2 minimum sample leaves and 2 minimum samples split.

Figure 4.5 contains the Confusion Matrix and the ROC plot for the ML classifier with information on GRACE risk factors and on inflammation biomarkers (WBC, CRP and albumin serum). The final performance metrics detailed in Table 4.5 will be used as a comparison against our preceding approach.

Sensitivity	Specificity	Geometric Mean	AUROC
0.83	0.84	0.83	0.91

Table 4.5: Obtained results from the ML model with GRACE risk factors and inflammation biomarkers

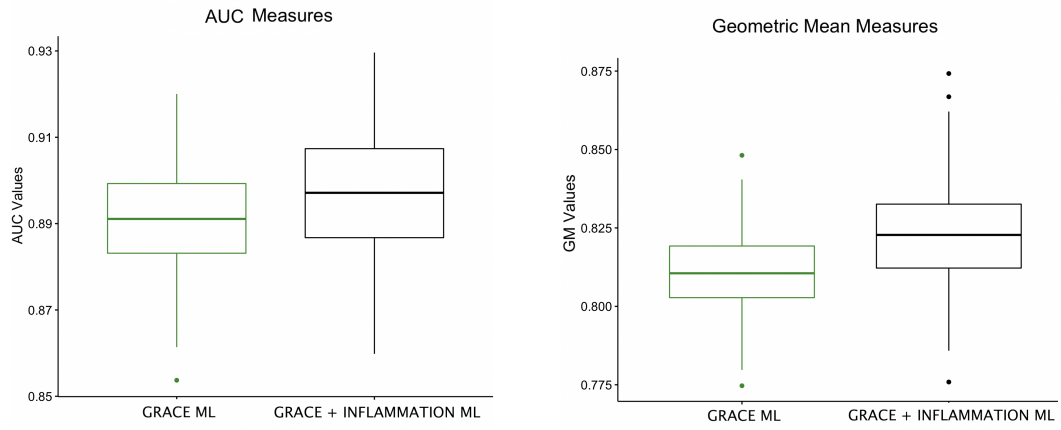


(a) Confusion Matrix (b) ROC curve plot

Figure 4.5: Metrics from ML GRACE and Inflammation Biomarkers Approach

4.2.4 Statistical Analysis

In order to statistically test the differences between the ML approaches, we ran our ML algorithms 500 times and collected their AUC and GM metrics to test if their means were statistically different. Since we were dealing with paired samples (the same group of patients), we performed a paired t-test and a Shapiro Wilk to check data normality.



(a) AUC Data Distribution - Boxplot (b) GM Data Distribution - Boxplot

Figure 4.6: Metrics from ML GRACE and Inflammation Biomarkers Approach

The *p-value* for the Shapiro Wilk tests for AUC metrics, are respectively, 0.86 for GRACE ML and 0.48 for the GRACE plus inflammation biomarkers approach. From these values, the p-value is greater than the significance level of 0.05, implying that the distribution of the differences is not significantly different from a normal distribution. Therefore we can assume normality.

The normality for the GM analysis was also assumed, with a *p-value* of 0.95 for GRACE ML and 0.26 for the GRACE plus inflammation biomarkers approach.

The *p-value* from the *t-test* related to GM data is $1.678e^{-}08$, less than the significance level of 0.05, so we reject the null hypothesis and conclude that the average GM performance with and without inflammation biomarkers is significantly different. Finally, the *p-value* for the AUC comparisons is 0.001. Therefore, we can also reject the null hypothesis and conclude that the average GM performances are significantly different.

4.2.5 Results Discussion

As expected, our ML models demonstrated a tremendous potential in predicting the mortality risk in ACS events, and even more, when compared to traditional ones. GM improved at least 14% and AUC at least 25%. Nevertheless, we can not ignore their limitations. Despite their similarities with the physicians reasoning (DT based-model), these models are undoubtedly not easy to interpret and, consequently, not adequate.

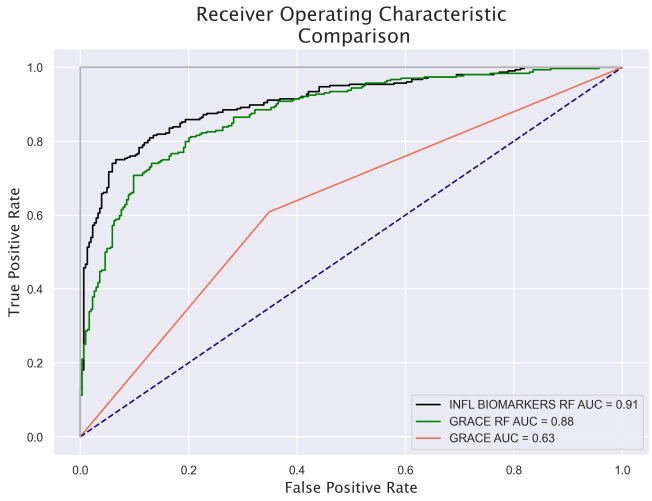


Figure 4.7: ROC and AUC metrics from the 3 developed models

Regarding our research question that intends to understand if the introduction of inflammation biomarkers can improve the GRACE Risk Factors discriminative power, Figure 4.7 shows the comparison between the ROC and auc metrics. From this figure and considering the statistical analysis results, we can conclude that the existence of an inflammatory process is relevant to a patient ACS outcome and, therefore, can be an important addition to the already existing risk stratification models.

4.3 Proposed Machine Learning Based Approach Results

In order to perform a comparison between our interpretable approach and the already established models, we used four methods: the Standard GRACE Risk Score, which is the risk assessment model of typical use, a Classic ML model, where the classifiers are considered black-boxes, and our proposed approach, where despite the usage of a ML model, the paradigm of how the model is employed is entirely different with and without inflammation biomarkers, for performing a final comparison.

Since our interpretable ML approach needed a deeper exploration, different classifiers were implemented to maximise our performance. We compared four classifiers: a ML Perceptron, Support Vector Machines (SVM), a Gradient Boosting (GBRT) and a Random Forest (RF).

4.3.1 Rule Performance Metrics

The performance from the discrimination of individual risk factor rules are presented in Table 4.6.

Risk Factor	$d_n > threshold$	**SE**	**SP**	**AUC**	**GM**
Maximum Killip	0.5	0.740	0.751	0.746	0.746
Maximum Creatinine	0.4	0.507	0.871	0.665	0.689
Age	0.7	0.507	0.871	0.613	0.624
Albumin	0.4	0.693	0.740	0.610	0.615
CRP Admission	0.4	0.413	0.537	0.600	0.642
Cardiac Frequency Admission	0.4	0.727	0.449	0.572	0.588
Troponin Admission	0.6	0.733	0.431	0.562	0.582
STEMI	0.5	0.467	0.641	0.547	0.554
WBC Admission	0.4	0.700	0.397	0.527	0.549
Systolic Pressure Admission	0.4	0.760	0.332	0.502	0.546

Table 4.6: Individual Risk Factors

Table 4.7 presents performance results from the combination of 2 Risk Factors, with Fisher's LDA implementation.

Risk Factor	$d_n > threshold$	**SE**	**SP**	**AUC**	**GM**
Maximum Killip; Age	0.5	0.713	0.680	0.697	0.697
Age; WBC Admission	0.5	0.723	0.667	0.695	0.695
Age; STEMI	0.7	0.716	0.650	0.683	0.683
WBC Admission; Maximum Creatinine	0.5	0.713	0.652	0.683	0.682
Albumin; Maximum Creatinine	0.5	0.713	0.636	0.675	0.673
Age; CRP Admission	0.5	0.703	0.639	0.671	0.670
Systolic Pressure Admission; Maximum Killip	0.6	0.681	0.656	0.669	0.669
Age; Cardiac Frequency Admission	0.5	0.717	0.619	0.668	0.666
Age; Albumin	0.5	0.692	0.639	0.665	0.665
Systolic Pressure Admission; Cardiac Frequency Admission	0.5	0.687	0.641	0.664	0.664

Table 4.7: Combination of 2 Risk Factors using Fisher's LDA

Rule Performance Discussion

Our strategy was developed in a straightforward way, where each rule definition is clearly explained, guaranteeing the interpretability of the final model. Our clinical partner externally validated this step (including the involved optimal threshold).

As presented in Tables 4.6 and 4.7, it is possible to understand that the most discriminant rule is the Killip Value, with a very acceptable value. Killip Value is a well-established Risk Factor in ACS, and it is a combination of several difficult conditions.

Regarding the performance metrics of combinations of 2 Risk Factors with Fisher's LDA, we can perceive that all rules benefit from the fusion of 2 risk factor inflammation, except for the Killip Value. Killip is one of the most decisive risk factors in ACS outcomes since it comprises several comorbidities. A Killip of 4 can lead to multiple organ dysfunction, which is a severe clinical condition. Moreover, Fisher's LDA combination of risk factors can offer interesting explanations for the created associations, which can be further explored. For example, the Age variable is extremely powerful, without any clinical surprise, since it can improve the performance from the most risk factors, when combined. However, and considering the sub-optimal performance from the mentioned rule and the importance of Killip as a Risk Factor, our developed approach will only consider individual rules.

4.3.2 Performance Evaluation

The results of the 6-months mortality risk prediction considering all the previously mentioned methodologies were analysed in terms of discrimination. Furthermore, the AUC and GM metrics for the rule definition process will also be presented.

Proposed Approach with GRACE and Inflammation Knowledge

Table 4.8 discriminates performance results from ML using simple rules with individual risk factor in the **training** phase of the interpretable approach, and Table 4.9 discriminates performance results in the **testing** phase.

ML Classifier	Specificity	Sensitivity	AUROC	Geometric Mean
Random Forest	0.808	0.752	**0.78**	**0.78**
ML Perceptron	0.808	0.752	**0.78**	**0.78**
Gradient Boosting	0.808	0.752	**0.78**	**0.78**
Support Vector Machine	0.811	0.743	0.777	0.777

Table 4.8: Individual Rule Performance on Mortality Prediction - Trainning

ML Classifier	Specificity	Sensitivity	AUROC	Geometric Mean
Random Forest	0.763	0.778	**0.77**	**0.77**
ML Perceptron	0.808	0.711	**0.76**	**0.758**
Gradient Boosting	0.669	0.867	**0.762**	**0.768**
Support Vector Machine	0.811	0.711	**0.76**	**0.759**

Table 4.9: Individual Rule Performance on Mortality Prediction - Testing

Proposed Approach with GRACE Risk Factors

Table 4.10 details the performance metrics for our interpretable approach, considering a RF classifier (similar to 4.2.3), but only using GRACE risk factors, i.e., with no knowledge on inflammation biomarkers.

	Specificity	**Sensitivity**	**AUROC**	**Geometric Mean**
Train Set	0.633	0.81	0.721	0.716
Test Set	0.846	0.644	0.738	0.745

Table 4.10: Interpretable Model with GRACE Risk Factors Evaluation Metrics

Machine Learning Approach with GRACE and Inflammation Knowledge

Table 4.11 states the performance metrics from a black-box RF (similar to 4.2.3), to further compare them with our interpretable developed approach.

	Specificity	**Sensitivity**	**AUROC**	**Geometric Mean**
Train Set	0.948	0.999	0.974	0.973
Test Set	0.863	0.604	0.733	0.722

Table 4.11: ML Model with GRACE Risk Factors and Inflammation Biomarkers Evaluation Metrics

GRACE Model

Finally, Table 4.12 presents the performance metrics for the medical GRACE model. This table will also be used to perform comparisons with our developed approach.

	Specificity	**Sensitivity**	**AUROC**	**Geometric Mean**
Train Set	0.607	0.875	0.723	0.713
Test Set	0.599	0.848	0.729	0.72

Table 4.12: GRACE Model Evaluation Metrics

4.3.3 Statistical Analysis

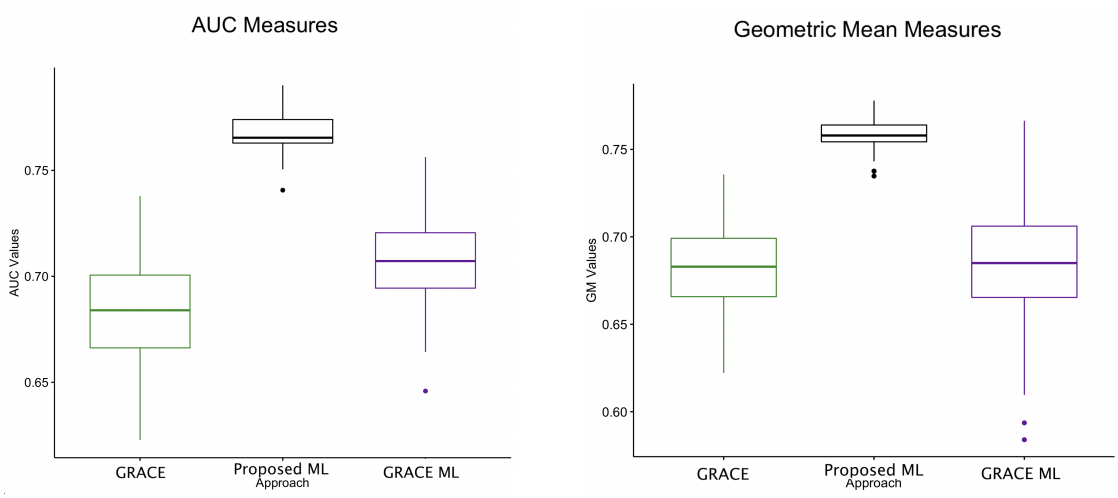
In order to statistically test the comparison between our proposed ML-based model with the standard GRACE and the "black-box model", we ran the three algorithms 500 times and collected their AUC and GM performance measures to test if their means were statistically different.

We were dealing with paired samples (the same group of patients). However, normality assumptions were not preserved. Therefore, we used the non-parametric Wilcoxon signed-rank test. The *p-value* from the *Wilcoxon signed-rank test* related all data comparisons is $< 2.2e-16$, strongly less the significance level of 0.05, so we reject the null hypothesis and conclude that the average performances measures (GM and AUC) when comparing

our proposed approach with the GRACE model and the standard ML model, are both significantly different.

4.3.4 Results Discussion

Considering our testing data and the statistical performed analysis, the GRACE risk score metrics were 0.729 and 0.72, respectively for AUC and GM. Our proposed approach obtained a AUC and GM of 0.77, which means a gain of 5%. In clinical domains, it is a considerable increase of performance.



(a) AUC Data Distribution - Boxplot

(b) GM Data Distribution - Boxplot

Figure 4.8: Comparison from Performance Metrics to validate the proposed approach

Moreover, as observed in Figure 4.8, the proposed methodology achieves slightly better performances than the standard RF classifier. Perhaps the computation of the optimal threshold can have contributed to the model's performance. Moreover, since no under or oversampling techniques have been applied to the data, the standard ML approach may lose some performance power.

Finally, considering our model with and without inflammation biomarkers influence, we can recognise that these variables can be influential additions to the already known risk scores. In fact, in our data, albumin serum is the $4th$ most discriminant variable, followed by CRP. Therefore, the same assumption as in Chapter 2 can be established: The existence of an inflammatory process is likely to have discriminative power to predict an ACS outcome.

Summing up, we consider our systems a generic and robust approach, able to provide a mechanism to support physicians in their decision-making process and potentially to increase their support on the use of ML procedures.

Chapter 5

Conclusions

The influence of inflammation biomarkers on ACS outcomes is an emerging research topic since the search for new risk factors that may include discriminant power to one of the most significant causes of death is inevitable. Appropriate answers for those concerns may be peremptory as a base for support critical methods.

As known, in clinical decisions support systems, misleading evaluations may result in a critical human wellness outcome. Therefore, there is no confidence in the medical community to apply models that can not be easily understood. Our objective was to develop a strategy that not only implements a hybrid scheme capable of combining ML and medical domains, but also analyses the influence of inflammation biomarkers on ACS outcome.

To achieve the main objectives of this dissertation, we started by studying the critical background and the existent state-of-the-art techniques. Building from this knowledge, we conducted a study into black-boxed ML models to examine whether inflammation biomarkers had any predictive power in our data. It was possible to establish that the introduction of inflammation biomarkers produced more accurate predictions.

The second part of our thesis was focused on the development and evaluation of a new prediction method that benefits from ML capabilities from a different perspective. To develop our strategy, we considered the necessity for interpretability and personalisation without decreasing the regular performances. Therefore, we created a virtual patient-based mechanism that makes simple rules to establish a patient's outcome prediction. According to our analysis, our interpretable solution keeps interpretability and medical acceptance, with even more significant performances than a standard ML and GRACE models.

Finally, we can assume that both of the main objectives (clinical and scientific) of this dissertation were achieved. We analysed our proposed clinical research question and developed a methodology that can be useful for clinicians since our procedure is understandable and generic. Moreover, this strategy can be applied in other contexts, related or not.

All in all, the present thesis supports that inflammation biomarkers as Albumin, C-Reactive Protein and the Leukocyte Count influence an Acute Coronary Syndrome outcome and that well-established risk factors as GRACE Risk Score could be improved using this knowledge. Moreover, we consider that our approach is a powerful tool in the explainable AI field since it can provide a mechanism to assist the physicians in their decision-making process.

This study can be considered an initial step, being the considerations validated in predicting a patient's 6-months mortality risk after an Acute Coronary Syndrome event. For future research, supplementary studies using other data-sets will be carried out to sub-

stantiate the proposed strategy's effectiveness further, and eventually, to incorporate the knowledge on inflammation biomarkers into GRACE Risk Score. Another direction may be the creation of several rules with different thresholds, almost as the GRACE risk score, but following the developed methodology. In addition, Fisher's LDA separation class process or other methodologies can be further explored to create more accurate rules based on the combination of different risk factors.

References

[1] Emad Abu-Assi, José M. García-Acuña, Carlos Peña-Gil, and José R. González-Juanatey. Validation of the grace risk score for predicting death within 6 months of follow-up in a contemporary cohort of patients with acute coronary syndrome. *Revista Española de Cardiología*, 2010.

[2] Dia AbuZeina and Fawaz S. Al-Anzi. Employing fisher discriminant analysis for arabic text classification. *Comput. Electr. Eng.*, 66:474–486, 2018.

[3] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In Amarda Shehu, Cathy H. Wu, Christina Boucher, Jing Li, Hongfang Liu, and Mihai Pop, editors, *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2018, Washington, DC, USA, August 29 - September 01, 2018*, pages 559–560. ACM, 2018.

[4] S. Al-Zaiti, Lucas Besomi, Zeineb Bouzid, Ziad Faramand, Stephanie O Frisch, C. Martin-Gill, R. Gregg, S. Saba, C. Callaway, and E. Sejdić. Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. *Nature Communications*, 11, 2020.

[5] Subhi J. Al'Aref, Khalil Anchouche, and Gurpreet Singh et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imagings. *European Heart Journal*, 2018.

[6] Dominick J. Angiolillo, Luigi M. Biasucci, Giovanna Liuzzo, and Filippo Crea. Inflammation in acute coronary syndromes: Mechanisms and clinical implications. *Revista Española de Cardiología*, 2004.

[7] Rebecca Bevans. Statistical tests: which one should you use?

[8] P.C. Calder, N. Ahluwalia, R. Albers, N. Bosco, R. Bourdet-Sicard, D. Haller, S.T. Holgate, L.S. Jonhsson, M.E. Latuli, A. Marco, J. Moreines, C. M'Rini, M. Muller, G. Pawelec, R.J.J. van Neerven, B. Watzl, and J. Zhao. A consideration of biomarkers to be used for evaluation of inflammation in human nutritional studies. *British Journal of Nutrition*, January 2013.

[9] A. John Camm, Thomas F. Lusher, and Patrick W. Serruys. *The ESC Textbook of Cardiovascular Medicine*. BlackWell Publishing, 2006.

[10] K. Sarat Chandra. Composite risk scores for acute coronary syndromes. *Indian Heart Journal*, page 270–272, 2012.

[11] John Danesh, Peter Whincup, Mary Walker, Lucy Lennon, Andrew Thomson, Paul Appleby, J Ruth Gallimore, and Mark B Pepys. Low grade inflammation and coronary heart disease: prospective study and updated meta-analyses. *BMJ - British Medicak Journal*, 2000.

[12] Pedro de Araújo Gonçalves, Jorge Ferreira, Carlos Aguiar, and Ricardo Seabra-Gomes. Timi, pursuit, and grace risk scores: sustained prognostic value and interaction with revascularization in nste-acs. *European Society of Cardiology*, page 865–872, 2005.

[13] I. El-Naqa, Yongyi Yang, M.N. Wernick, N.P. Galatsanos, and R.M. Nishikawa. A support vector machine approach for detection of microcalcifications. *IEEE Transactions on Medical Imaging*, 21(12):1552–1563, 2002.

[14] Basem Elbarouni, Shaun G. Goodman, Raymond T. Yan, Robert C. Welsh, Jan M. Kornder, J. Paul DeYoung, Graham C. Wong, Barry Rose, François R. Grondin, Richard Gallo, Mary Tan, Amparo Casanova, Kim A. Eagle, and Andrew T. Yan. Validation of the global registry of acute coronary event (grace) risk score for in-hospital mortality in patients with acute coronary syndrome in canada. *American nal*, 158(3):392–399, 2009.

[15] Rohith Gandhi. Support vector machine — introduction to machine learning algorithms.

[16] Dion H. Goh and Rebecca P. Ang. An introduction to association rule mining: An application in counseling and help-seeking behavior of adolescents. *Behavior Research Methods*, pages 259–266, 2007.

[17] Prince Grover. Gradient boosting from scratch.

[18] Vrutik Halani. Fischer's linear discriminant analysis in python from scratch.

[19] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques, 3rd edition*. Morgan Kaufmann, 2011.

[20] IBM. Statistics - parametric and nonparametric.

[21] Clerkship Directions in Emergency Medicine. St-elevation mi.

[22] National Cancer Institute. Nci dictionary of cancer terms.

[23] Gareth James, Witten Daniela, Hastie Trevor, and Tibshirani Robert. An introduction to statistical learning. *Springer*, 2015.

[24] F. Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Q. Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2:230 – 243, 2017.

[25] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Yilong Wang Sufeng Ma, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *CrossMark*, 2017.

[26] IBM Analytics Julianna Delua. Supervised vs. unsupervised learning: What's the difference?

[27] Anat Kalinski. Personalization using machine learning — from data science to user experience.

[28] Will Koehrsen. Hyperparameter tuning the random forest in python.

[29] T. Warren Liao. Clustering of time series data—a surveys. *Pattern Recognition*, page 1857–1874, 2005.

[30] A. Maseri and D. Cianflone. Inflammation in acute coronary syndromes. *European nal Supplements*, $4(\text{suppl}_B) : B8 - -B13, 032002$.

[31] Kjell Johnson Max Kuhn. *Applied Predictive Modeling*. Springer, 2016.

[32] Sridevi Devaraj Medscape. Albumin.

[33] Sridevi Devaraj Medscape. High-sensitivity c-reactive protein.

[34] Tarek Nafee, William Gibson, Ryan Travis, Megan Yee, Mathieu Kerneis, Magnus Ohman, and Michael Gibson. Machine learning versus traditional risk stratification methods in acute coronary syndrome: a pooled randomized clinical trial analysis. *Journal of the American College of Cardiology*, 73:258, 03 2019.

[35] Sarang Narkhede. Understanding auc - roc curve.

[36] Yung-Kyun Noh, Ji Young Park, Byoung Geol Choi, Kee-Eung Kim, and Seung-Woon Rha. A machine learning-based approach for the prediction of acute coronary syndrome requiring revascularization. *J. Medical Syst.*, 43(8):253:1–253:8, 2019.

[37] Jacob Odeberg, Michael Freitag, Henrik Forssell, Ivar Vaara, Marie-Louise Persson, Håkan Odeberg, Anders Halling, Lennart Råstam, and Ulf Lindblad. Influence of pre-existing inflammation on the outcome of acute coronary syndrome: a cross-sectional study. *BMJ Open*, 6(1), 2016.

[38] Konrad Pieszko, Jarosław Hiczkiewicz, Paweł Budzianowski, Jan Budzianowski, Janusz Rzeźniczak, Karolina Pieszko, and Pawel Burchardt. Predicting long-term mortality after acute coronary syndrome using machine learning techniques and hematological markers. *Disease Markers*, 2019:1–9, 01 2019.

[39] Medline Plus. Albumin blood test.

[40] Vili Podgorelec, Peter Kokol, Bruno Stiglicand, and Ivan Rozman. Decision trees: An overview and their use in medicine. *Journal of medical systems vol. 26,5 (2002): 445-63*, 2002.

[41] Anusha Mujumdar Rafia Inam, Ahmad Terra. Explainable ai – how humans can trust ai.

[42] Abbas Rezaianzadeh, Maryam Dastoorpoor, Majid Sanaei, Cirruse Salehnasab, Mohammad Javad Mohammadi, and Ali Mousavizadeh. Predictors of length of stay in the coronary care unit in patient with acute coronary syndrome based on data mining methods. *Clinical Epidemiology and Global Health*, 8(2):383–388, 2020.

[43] Section.io. Introduction to random forest in machine learning.

[44] Rahul Kumar Sevakula, Wan-Tai M. Au-Yeung, Jagmeet P. Singh, E. Kevin Heist, Eric M. Isselbacher, and Antonis A. Armoundas. State-of-the-art machine learning techniques aiming to improve patient outcomes pertaining to the cardiovascular system. *American Heart Association*, 2020.

[45] S Sharma, P G Jackson, and J Makan. Cardiac troponins. *Journal of Clinical Pathology*, 57(10):1025–1026, 2004.

[46] Padam Singh. P value, statistical significance and clinical significance. *J Clin Prev Cardiology*, 2013.

[47] Jennifer N. Smith, Jenna M. Negrelli, Megha B. Manek, Emily M. Hawes, and Anthony J. Viera. Diagnosis and management of acute coronary syndrome: An evidence-based update. *BJM Open*, 2014.

[48] PhD Sridevi Devaraj, Sep 2020.

[49] Kyle Strimbu and Jorge A. Tavel. What are biomarkers? *National Institute of Health*, 2010.

[50] Statistics How To. Comparison of means techniques.

[51] Statistics How To. Mann whitney u test: Definition.

[52] Statistical tools for high-throughput data analysis. Paired samples t-test in r.

[53] Michigan State University. Canadian cardiovascular society (css) classification of angina.

[54] Analytics Vidhya. Overcoming class imbalance using smote techniques.

[55] Merriam Webster. C-reactive protein.

[56] Dawn Teare Zhe Hui Hoo, Jane Candlish. What is a roc curve? *J Clin Prev Cardiology*, 2013.

Appendices

This page is intentionally left blank.

Appendix A

Inflammation Biomarkers Mortality Analysis considering Grace Risk Stratification

This appendix details the performed analysis on inflammation biomarkers mortality, based on the GRACE Risk Stratification. Histogram A.1 represents the distribution of deceased patients with hypoalbuminia (low levels of serum albumin). Histogram A.2 represents the distribution of deceased patients with high levels of C-Reactive Protein (CRP). Finally, histogram A.3 represents the distribution of deceased patients with high counts of White Blood Cells (WBC).

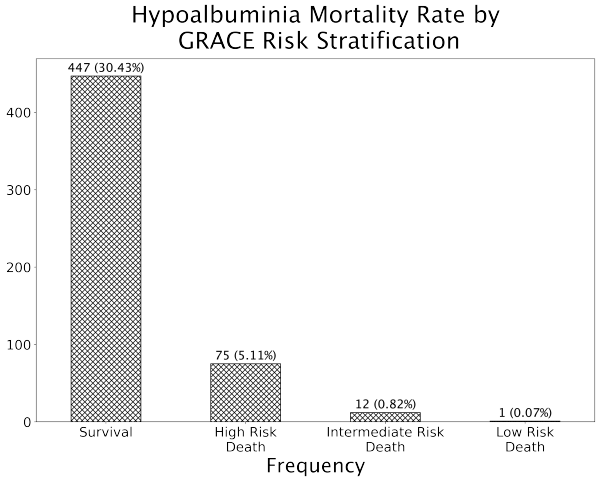


Figure A.1: Albumin Analysis Mortality Rate by Grace Risk Stratification

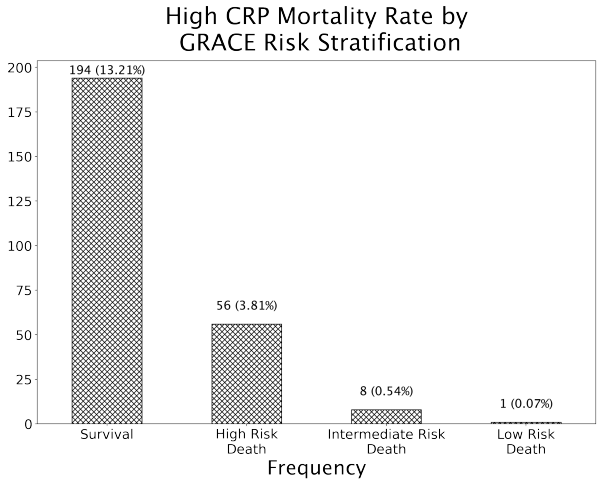


Figure A.2: High CRP Analysis Mortality Rate by Grace Risk Stratification

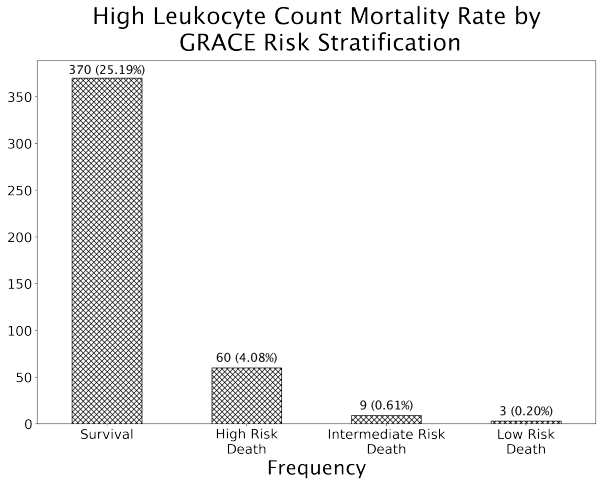


Figure A.3: High WBC Analysis Mortality Rate by Grace Risk Stratification

Appendix B

Work Plan

This appendix describes the work plan for the first and the second semester.

B.1 First Semester

Considering the planned work for developing our dissertation, the first semester was focused on studying clinical and scientific background concepts and related work. Besides this study, in the first semester, we intended to analyse our research question on the inflammation biomarkers influence.

Furthermore, our dataset was pre-processed for the validation of our strategy, and several experiments were conducted. Figure B.1 presents the tasks conducted during the first semester through a Gantt chart.

B.2 Second Semester

Apart from the continuation on studying the influence of inflammation biomarkers in ACS outcome, the work for the second semester can be divided into three major tasks.

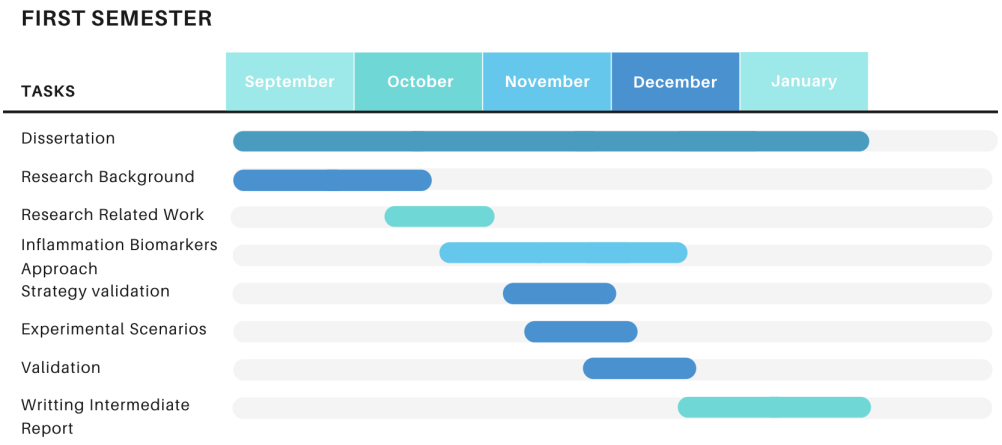


Figure B.1: Gantt Chart for the First Semester

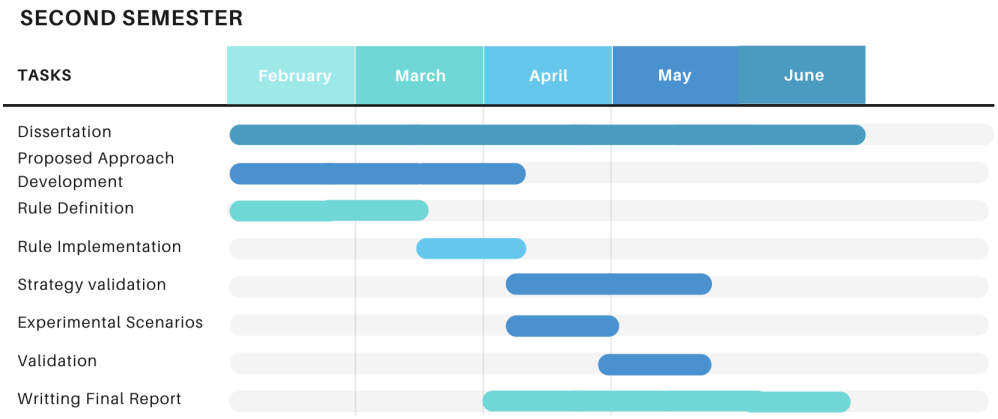


Figure B.2: Gantt chart for the second semester

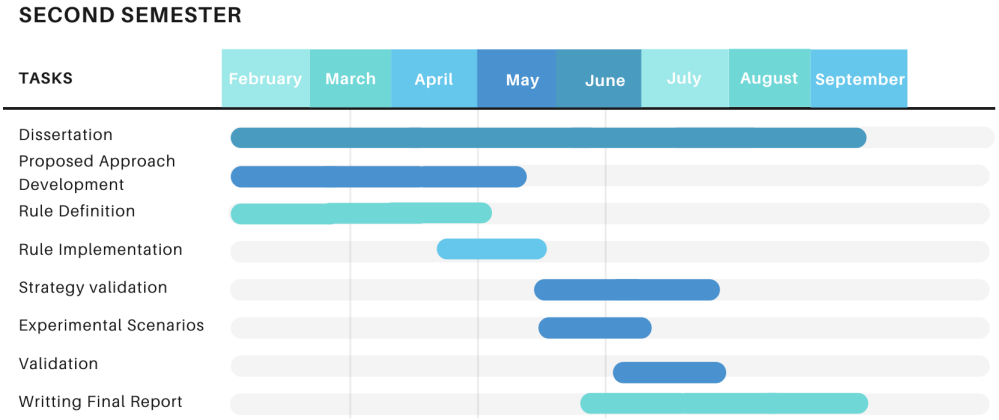


Figure B.3: Gantt chart for the second semester

The first task was the rule development process, calculating optimal thresholds in order to obtain predictive and at the same time clinically accepted rules. The second task consisted of the implementation of the previously established rules in our proposed methodology. And finally, the third task consisted validates our approach.

Figure B.2 presents the Predicted Gantt Chart, made in the first semester, for our second-semester task plan.

Figure B.3 shows the Gantt chart of the second semester with the revised scheduling. The main difference between the charts is the duration of the rule definition task. In fact, the task took more time than we anticipated due to some barriers between clinical and scientific knowledge.