UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR INGENIERÍA INDUSTRIAL VALENCIA

BIOMEDICAL ENGINEERING MASTER THESIS

APPLICATION OF HEART RATE VARIABILITY IN THE PREDICTION OF VASCULAR EVENTS IN HYPERTENSIVE PATIENTS

AUTHOR: ROBERTO TORNERO COSTA

SUPERVISOR: JOSÉ JOAQUÍN RIETA IBAÑEZ

Academic year: 2020-21

To my family, for all their support during this long year.

To José Joaquín Rieta, for giving me the opportunity to work with him to achieve this thesis.

To all the people with whom I have shared this wonderful master's degree and who have helped to make it a little easier.

Abstract

Vascular events can be cardiovascular or cerebrovascular, that is, myocardial infarction or cerebrovascular accident, which are the main cause of premature death and disability in developed countries. Due to this, there is great interest in the development of computational tools for their prognosis and diagnosis. A very relevant variable in its evaluation is the heart rate, which is a dynamic signal that fluctuates over time. Heart rate variability (HRV) is defined as the beat-to-beat variation in heart rate and may be indicative of the possible presence of a pathological condition. On the other hand, high blood pressure is an important risk factor for many cardiovascular diseases. Thus, high blood pressure is associated with increased risks of stroke, coronary heart disease, chronic kidney disease, heart failure, and death in general. In fact, small reductions in blood pressure are known to markedly reduce cardiovascular morbidity and mortality in the population.

The objective of this work is to evaluate the predictive value of short-term HRV by developing models based on data mining algorithms to provide an automatic tool for stratifying the risk of vascular accident for hypertensive patients. For this specific framework, a dataset (SHAREE) from the University Hospital of Naples in 2015 has been used. The methodology of the original research was tested and a new methodology was also used to validate the hypothesis. 5 minutes, 30 minutes and 1 hour of HRV analysis during sleep stage - 00 a.m to 06 a.m - were compared. Classification models showed a similar performance for the different times, but those trained with short-term HRV analysis showed a higher sensitivity for 5 minutes analysis and a slightly higher F1 score metric. Therefore, this project concludes that it is feasible to use short-term HRV analysis to predict the risk of vascular accident for hypertensive patients.

Keywords: Heart rate variability; HRV; short-term; blood pressure; hypertensive; machine learning; classification models; vascular accident; heart attack

Resumen

Los eventos vasculares pueden ser de tipo cardiovascular o cerebrovascular, es decir, infarto de miocardio o accidente cerebrovascular, que son la principal causa de muerte prematura y discapacidad en los países desarrollados. Debido a ello, existe un gran interés en el desarrollo de herramientas computacionales para el pronóstico y diagnóstico de los mismos. Una variable muy relevante en su evaluación es la frecuencia cardíaca, que es una señal dinámica que fluctúa a lo largo del tiempo. La variabilidad de la frecuencia cardíaca (VFC) se define como la variación, latido a latido, de dicha frecuencia y puede ser indicativa de la posible presencia de una afección patológica. Por otro lado, la hipertensión arterial es un importante factor de riesgo de muchas enfermedades cardiovasculares. Así, una presión arterial elevada se asocia con mayores riesgos de accidente cerebrovascular, enfermedad coronaria, enfermedad renal crónica, insuficiencia cardíaca y muerte en general. De hecho, se sabe que pequeñas reducciones de la presión arterial reducen notablemente la morbilidad y mortalidad cardiovascular de la población.

El objetivo de este trabajo es evaluar el valor predictivo de la VFC a corto plazo (5 minutos) desarrollando modelos basados en algoritmos de minería de datos para proporcionar una herramienta automática de estratificación del riesgo de accidente vascular para pacientes hipertensos. Para este proyecto, se usó un dataset del proyecto SHAREE del Hospital Universitario de Nápoles en 2015. Se puso a prueba la metodología original y se usó una nueva para validar la hipótesis. Se compararon los distintos modelos entrenados con el análisis de VFC de 5 minutos, 30 y 1 hora extraídos durante el sueño nocturno (de 00 a.m a 06 a.m). Los modelos mostraron rendimientos similares para los distintos tiempos. Aquellos entrenados con solo 5 minutos de análisis obtuvieron una superior puntuación en sensibilidad y levemente superior en F1 score. Por lo tanto, la conclusión de este proyecto es que sí es posible utilizar 5 minutos de análisis VFC para discriminar el riesgo de eventos cardiovasculares en pacientes hipertensos.

Palabras clave: Variabilidad de la frecuencia cardiaca; HRV; presión arterial; aprendizaje automático; modelos de clasificación; accidente vascular; infarto.

Resum

Els esdeveniments vasculars poden ser de tipus cardiovascular o cerebrovascular, és a dir, infart de miocardi o accident cerebrovascular, que són la principal causa de mort prematura i discapacitat als països desenvolupats. A causa d'això, existeix un gran interés en el desenvolupament d'eines computacionals per al pronòstic i diagnòstic d'aquests. Una variable molt rellevant en la seua avaluació és la freqüència cardíaca, que és un senyal dinàmic que fluctua al llarg del temps. La variabilitat de la freqüència cardíaca (VFC) es defineix com la variació, bategat a batec, d'aquesta freqüència i pot ser indicativa de la possible presència d'una afecció patològica. D'altra banda, la hipertensió arterial és un important factor de risc de moltes malalties cardiovasculars. Així, una pressió arterial elevada s'associa amb majors riscos d'accident cerebrovascular, malaltia coronària, malaltia renal crònica, insuficiència cardíaca i mort en general. De fet, se sap que xicotetes reduccions de la pressió arterial redueixen notablement la morbiditat i mortalitat cardiovascular de la població.

L'objectiu d'aquest treball és avaluar el valor predictiu de la VFC a curt termini (5 minuts) desenvolupant models basats en algorismes de mineria de dades per a proporcionar una eina automàtica d'estratificació del risc d'accident vascular per a pacients hipertensos. Per a aquest projecte, es va usar un dataset del projecte SHAREE de l'Hospital Universitari de Nàpols en 2015. Es va posar a prova la metodologia original i s'use una nova per a validar la hipòtesi. Es van comparar els diferents models entrenats amb l'anàlisi de VFC de 5 minuts, 30 i 1 hora extrets durant el somni nocturn (de 00 a.m a 06 a.m). Els models van mostrar rendiments similars per als diferents temps. Aquells entrenats amb només 5 minuts d'anàlisis van obtindre una superior puntuació en sensibilitat i lleument superior en F1 score. Per tant, la conclusió d'aquest projecte és que sí que és possible utilitzar 5 minuts d'anàlisis VFC per a discriminar el risc d'esdeveniments cardiovasculars en pacients hipertensos.

Paraules clau: Variabilitat de la freqüència cardíaca; HRV; pressió arterial; aprenentatge automàtic; models de classificació; accident vascular; infart

Contents

Chapter 1

Introduction

1.1 Motivation

At the turn of the last century, several studies began to analyse the connection between an abnormally high blood pressure (HBP) and the risk of undergoing a cardiovascular event or disease (CVD). Moreover, at the beginning of this century, the impact of high and normal blood pressure on risk of CVD was published, showing the risk of developing cardiovascular events due to hypertension [1]. Furthermore, it is suggested that HBP may result in a reduced adaptation of the autonomic nervous system to haemodynamic changes in some patients. A low or depressed heart rate variability (HRV), where the sympathetic system dominates, may reduce this adaptation and manifest precedence to a disease or a CVD. Combined with other reasons, such as the immune system being affected by alterations in the autonomic nervous system, this may lead to an increased likelihood of CVDs [2] [3]. Therefore, there is a predisposition to develop cardiac disorders due to HBP and alterations of the autonomic nervous system, measurable by HRV.

A report from the World Health Organization (WHO) classifies CVDs as the main causes of mortalities in the last century, from the years 2000 to 2019, – Leading causes of death and disability: A visual summary of global and regional trends 2000-2019 [4]. In this report, ischemic heart diseases and strokes are revealed as the two main factors of death around all the world – uninterruptedly since 2000. Ischemic heart diseases led to almost 9 million of deaths only in 2019 and the 16% of deceases in this century. Furthermore, when looking at morbidity and life expectancy, it is clear that cardiovascular diseases have resulted in the loss of more than 320 million of years of life expectancy (DALYs) only in 2019.

High blood pressure – also known as arterial hypertension – is one of the prominent risk factors of developing a new CVD or undergoing a heart event. Also, it has been investigated how the supervision of blood pressure helps to significantly decrease the incidence of either damage in target tissue – vascular tissue – or morbidity and mortality due to diseases in other related tissues – such as heart tissue [5]. However, HBP shows one of the worst control rates in regard to other risk cardiac factors among people, patients or not of a cardiovascular disease [6] [7]. Having an impact in 20-30% of adult population, HBP rates are increasingly apparent amongst the elderly population. Nonetheless, HBP cannot be considered as a true disease rather than a risk factor for a wide group of diseases. Nowadays, there are multiple sort of diagnostic tests to distinguish HBP and its relation with different CVDs [8].

A high percent of people with HBP are not aware of this state, hence, they are not given any treatment. Moreover, HBP is a risk which may remain silent for years, without showing any symptoms [7]. Sundry studies have already settled the Heart Rate Variability as a good analysis method with a great

performance for intelligent diagnosis and monitorisation of high blood pressure [9] [10] [11]. Heart rate and its variability can be measured in every patient using different diagnostic techniques, like ECG, PPG or ABP; techniques which may vary in simplicity of cost or form of use.

Due to the impact risk of high blood pressure in undergoing cardiovascular events, it is vital to monitor these patients in order to evaluate this risk and to act accordingly. For example, using predictive methods – with a great performance –, the patient's potential risk could be analysed easily and quickly during day-to-day checks-up. Even more, it may be possible to develop intelligent medical monitoring devices. In fact, always as a support decision tool for doctors. Although, a consensus about the best HRV techniques and methods has not been reached yet, further research is needed to conclude best mathematics, physic and medical approach to analyse HRV measures and to integrate it into intelligent predictive algorithms.

1.2 Hypothesis and Main Goal

Heart Rate Variability analysis has proven to be a helpful predictive tool for the autonomic nervous system, but sometimes it is only feasible to use in long-term evaluation. Reducing the ECG monitoring to short-term - 5 minutes or less - while ensuring good predictive results as long-term HRV does could largely decrease costs and greatly facilitate patient's collaboration. The Multidisciplinary Department of Medical, Surgical and Dental Sciences of Second University of Naples studied the potential of HRV analysis in 5 minutes ECG Holter in order to predict the risk of cardiovascular diseases [3]. This Master's dissertation aims to validate if this hypothesis could perform similar results as long-term HRV. Therefore, there are two main goals for this Master's dissertation:

- First, evaluate if proposal from the paper of predicting cardiovascular events through only 5 minutes short-term HRV is replicable. Hence, to validate if predicting CVDs in hypertensive people is feasible with only randomly chosen 5 minutes sections of ECG Holter with the methodology applied in. In this paper, he Multidisciplinary Department of Medical, Surgical and Dental Sciences of Second University of Naples suggests that 5 minutes short-term HRV analysis has a greater prediction ability than other clinic information in hypertensive people.

- The second goal is settle our own methodology to compare the predictive results of ML methods when using different randomly chosen time spans of the patients. In this way, the validity of short-term predictions will be compared with long-term predictions; the later much more widely acknowledged in research community. However, the dataset collected for the previous research and used in this project could be seen as having certain weaknesses. The collection of information for such health researching studies can be challenging, sometimes resulting in data that is difficult to work with. In this second part of the project, different aspects of data science will be applied to try to obtain the most accurate results that this dataset allows for 5 and 30 minutes as well as 1 hour of signals. Then, the outcomes will be compared to validated if 5 minutes of signals could be enough for this framework..

1.3 Structure of this document

This document is structured in the following chapters:

- **Chapter 2: General Background.** This chapter reviews some relevant concepts aimed at establishing the proper theoretical framework of this Master's Dissertation.
- **Chapter 3: Material and Methodology.** The material used and the methodology applied during this dissertation are explained alongside the required mathematical as well as engineering and data science framework. Methodology steps for preprocessing of time series, HRV analysis and machine learning classification as well as a brief explanation of its theoretical framework are introduced in this chapter.
- **Chapter 4: Results and Discussion** The evaluation procedure of the methodology and dataset provided in [3] as well as the evaluation of the main hypothesis introduced are described and the consequent results obtained are shown. Also, several aspects are discussed in this section, such as the assessment of the results obtained and the main strengths and limitations.
- **Chapter 5: Conclusion and Future Work.** The concluding remarks are exposed, as well as some future lines proposed to deepen in this research.
- **Appendix: Budget.** The costs associated with the development of this Master Dissertation.

Chapter 2

General Background

2.1 Blood pressure and its classification

Blood pressure is described as the force exerted by the circulation of the blood volume on the walls of arteries, veins and capillaries. Due to the intermittent contraction of the heart – so-called beats – blood pressure oscillates and that depends on the blood vessel and its characteristics. In the case of this project, we will focus on the arterial blood pressure and its subtypes.

First, it is necessary to introduce a brief explanation about arteries and its behaviour. This type of blood vessels not only acts as conduits for blood circulation. They have also a cushion function which helps to maintain blood pressure steadily along the first half of blood system. This function flattens the oscillation of blood pressure due to heart beats, highly reducing heart work. The increase on stiffness of blood vessels decrease this cushion function which is related with the developing of many cardiovascular diseases [12].

Blood pressure is usually divided in two terms: systolic and diastolic pressure. Systolic pressure refers to the pressure of blood on the walls of arteries due to ejection of blood during cardiac output. Diastolic pressure occurs during relaxation of the heart after the beat and it measures the resistance of arteries to the passage of blood.

Blood pressure is classified depending on systolic and diastolic terms. The European Society of Cardiology (ESC) and the European Society of Hypertension (ESH) determine 4 subgroups of pressure to classify patients due to their blood pressure, as **table 2.1** shows [13]. In HBP subgroup, 3 grades of hypertension are found, depending on the intervals of pressure. However, the American College of Cardiology and the American Heart Association Task Force have recently reduced the values to sort HBP – or hypertension – and high normal pressure – or pre-hypertension – in 10 mmHg [14].

2.2 Relationship between hypertension and cardiovascular diseases

The American College of Cardiology and the American Heart Association Task Force's own joint report warns about the growing cases of hypertension in the older age groups of the population. The long-term cumulative incidence of this disorder is 0.3%, 6.5%, and 37% for developing hypertension at ages 25, 45, and 65, respectively [14].

Arterial hypertension cannot be considered a disease, strictly speaking. Rather, it is a warning sign for different physiological events [8]. The relationship between hypertension and the increased risk of

Table 2.1: Blood pressure classification values (ESC/ESH).

Category		Systolic (mmHg)		Diastolic (mmHg)
Optimal		< 120	y	< 80
Normal		120 – 129	y/o	80 - 84
High normal		130 -139	y/o	85 - 89
High blood pressure	Grade 1	140 – 159	y/o	90 – 99
	Grade 2	160 – 179	y/o	100 - 109
	Grade 3	≥ 180	y/o	≥ 110

developing a cardiovascular accident is considered to be proven with sufficient scientific evidence [5] [13]. As well as early diagnosis, treatment and effective control of blood pressure helps to reduce the morbidity and mortality rate of major cardiovascular events [14].

For example, chronic high blood pressure leads to constant growth of the left ventricle in order to maintain a continuous ejection load. This growth can lead to what is known as left ventricular hypertrophy, which can eventually lead to heart failure [12].

2.3 Cardiac physiology, properties of the sinus rhythm and its disturbances

The heart is four-chambered, two atria and two ventricles, and is symmetrical, with an atrium-ventricle pair located on each side. The repetitive coordination in the contraction of these chambers forms what is called the cardiac cycle. An isolated event in this cycle is known as a heartbeat.

The cardiac cycle for the atrium or ventricle can be analysed as the set of two phases: systole and diastole. While systole is the period of contraction of the chamber muscle or myocardium, diastole is its period of relaxation. In most cases, the events of the cardiac cycle are represented for the left side, since the ejection of the blood flow that will travel through virtually the entire body occurs via the left ventricle. **Figure 2.1** shows the cardiac cycle for the left ventricle, as well as pressure changes in the left ventricle and atrium, and the aorta. It also shows the relationship of these events to various physiological techniques such as ECG or phonocardiogram [15].

The cardiac cycle is coordinated by the cardiac electrical conduction system. This system is composed of the sinoatrial (SA) node, which continues into the atrioventricular (AV) node, and branches into the bundle of His and the Purkinje fibre network. These components contain cells capable of contracting and autonomously generating an electrical impulse that excites the heart muscle. These cells are known as cardiomyocytes.

Understanding the mechanisms of this system is important to understand how it works in HRV analysis and how it is calculated. In a healthy heart, the electrical impulse is generated from cardiomyocytes in the sinoatrial node. This node is located in the atrial wall. The electrical impulse is then transmitted through the remaining sections as shown in **Figure 2.2**. This organised conduction causes the coordinated contraction of first atria and then ventricles, allowing for a healthy heart rhythm. The rhythm coordinated by the sinoatrial node is known as sinus rhythm and when it is within specific heart rate intervals it is considered the reference healthy heart rhythm.

However, cardiomyocytes in the atrioventricular node and sometimes Purkinje fibres can also have

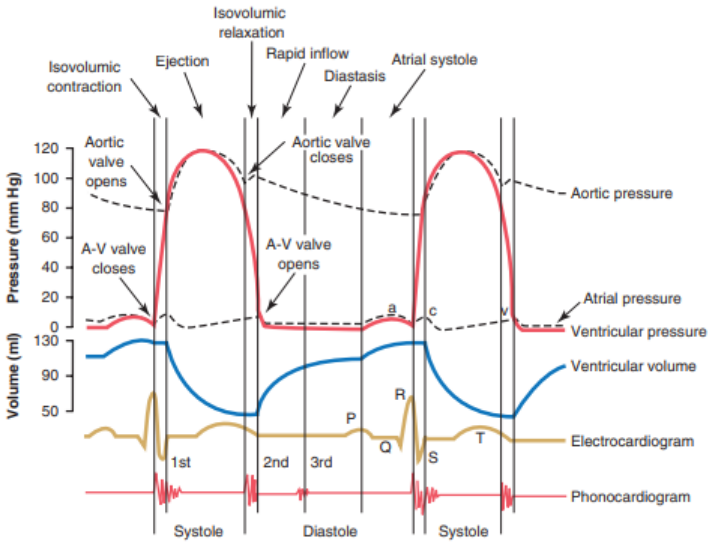


Figure 2.1: Events of cardiac cycle on the left side of the heart alongside ECG and PCG. Derived from 'Guyton and Hall Textbook of Medical Physiology' (13th edition) [15].

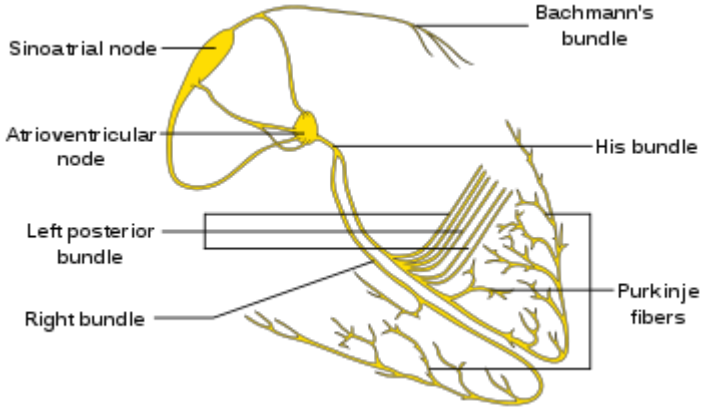


Figure 2.2: Electrical conduction system of the heart.

spontaneous contractions. Normally these contractions do not overpower the sinus rhythm due to the constant predominance of the sinus rhythm, which does not allow enough range for one of these contractions to occur. However, it may happen that a contraction is generated that is not dictated by the sinoatrial node. When this happens, what is known as an ectopic beat may occur. Identifying these beats is important so as not to alter the true measurements of HRV analysis [16], which will be explained in their own section.

Ectopic beats originate locally at a specific point in the heart tissue, known as an ectopic focus. Depending on their location, these beats can be classified as supraventricular beats, like atrial premature contraction (APC), or ventricular premature contraction (VPC). A ventricular premature ectopic beat introduces a new heartbeat between two healthy sinus beats. This ectopic beat has a very characteristic morphology distinct from the beat given by the sinus node. However, an atrial premature beat involves a very similar morphology. In addition, it can restart the sinus rhythm [17] [18]. The following **figure 2.3** illustrates normal heartbeats, the presence of an atrial premature contraction and finally, the presence of ventricular premature contraction.

Other heart rhythm disturbances are abnormal reduction of the heart rate, bradycardia or, inversely, tachycardia. In addition to various arrhythmias and fibrillations, which define an irregular heart rhythm and alter its frequency.

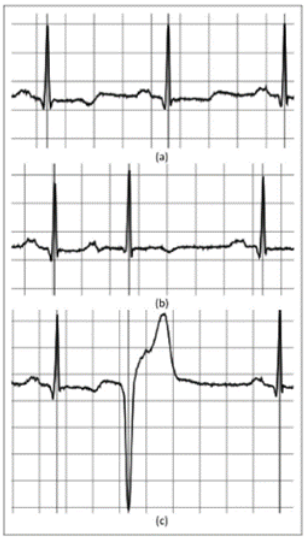


Figure 2.3: Representation of (a) normal heartbeats, (b) APC and (c) VPC. Obtained from [18].

2.4 Electrocardiogram. Measurement of cardiac physiology

As can be seen in **figure 2.1** in the previous section, the electrocardiogram, or ECG, signal is composed of several curves and slopes that represent different events of the cardiac cycle and is coordinated with these events. These curves and slopes are the graphical representation of the electrical conduction of cardiac depolarisation and are obtained using certain electrodes placed at previously defined points on the skin. The combinations of these electrodes are known as leads and depending on the lead, the shape of the ECG signal can vary considerably [15].

These differences between variations, as illustrated in **figure 2.4** - referring to standard chest leads - depend on the location of the electrodes, because the recording of electrical conduction will change depending on the orientation. This is relevant for the computational algorithms used to process and extract features from the ECG signal.

If we focus on a particular lead, for example lead V5 as shown in **figure 2.5**, the signal can be divided into the P wave, the QRS complex and the T wave. The ECG sample may vary with new components but these three are usually observed. The P wave is produced by depolarisation of the atrium and the T wave by repolarisation of the ventricles. The QRS complex corresponds to the depolarisation of the ventricles, just before their contraction. By convention, the R wave of the QRS complex - the first positive peak of the complex - is used to label the heartbeat.

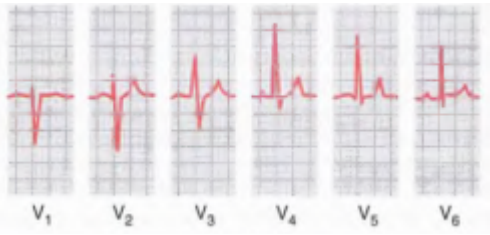


Figure 2.4: Standard chest leads. Obtained from 'Guyton and Hall Textbook of Medical Physiology' (13th edition) [15].

There are a wide range of devices to monitor ECG. Some devices record ECG signals from heart during a few minutes and other ECG devices can perform recordings which last for several hours. One device usually common in heart research is the ECG Holter device. This device allows to monitor the ECG signals during 24 hours or more, recording and saving the information for later processing. A

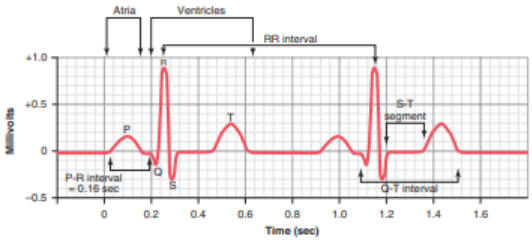


Figure 2.5: Elements of a normal electocardiogram. Obtained from 'Guyton and Hall Textbook of Medical Physiology' (13th edition) [15].

Holter device is composed with the required hardware to record heart signals (electrodes), the hardware to store it and the software to process it. Moreover, some ECG devices can be also monitors, allowing to a continuous follow-up of the recording, **figure 2.6**. The ECG samples which will be used in this project were obtained using this device. However, it is necessary to emphasise that due to the long duration of recordings, ECG Holter devices usually have a low sampling frequency.

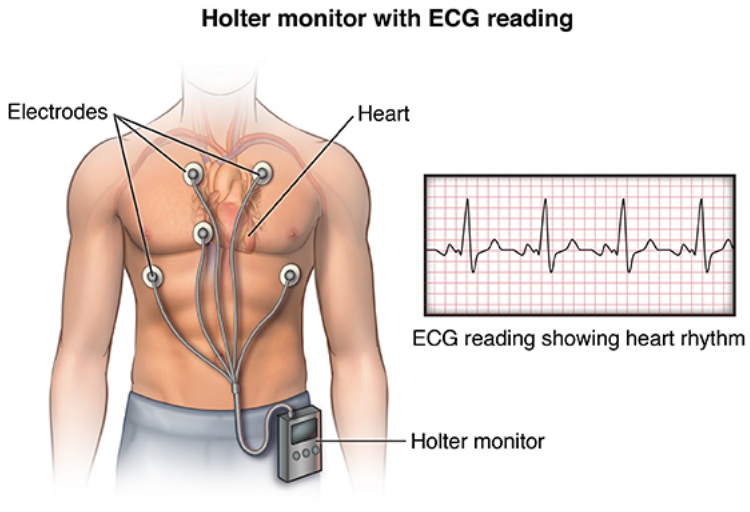


Figure 2.6: ECG Holter system.

2.5 Autonomic Nervous System

As defined by Guyton's 13th edition, "The autonomic nervous system is the portion of the nervous system that controls most visceral functions of the body. One of the most striking characteristics of the autonomic nervous system is the rapidity and intensity with which it can change visceral functions. For instance, within 3 to 5 seconds it can increase the heart rate to twice normal, and within 10 to 15 seconds the arterial pressure can be doubled. At the other extreme, the arterial pressure can be decreased low enough within 10 to 15 seconds to cause fainting" [15]. We can find this section of the nervous system controlling the heart rate as one of its most important roles. This is done through sinus rhythm regulation, managing how sinoatrial node is activated and depolarised.

However, this system not only regulates the sinoatrial node, but also other nodes, as can be seen in **figure 2.7**. While sympathetic nerves stimulate the cardiac nodes - and so increase heart rate and strength of ventricular muscle contraction -, parasympathetic nerves (vagus nerves) do the opposite. The struggle

of these two systems regulate the heart output and its pumping effectiveness. Further, when vagus nerves are strongly stimulated, even the heart can stop for a few seconds.

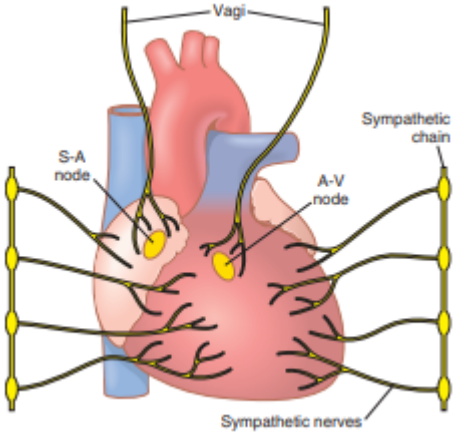


Figure 2.7: Cardiac sympathetic and parasympathetic nerves. (The vagus nerves to the heart are parasympathetic nerves.) A-V, atrioventricular; S-A, sinoatrial. Obtained from 'Guyton and Hall Textbook of Medical Physiology' (13th edition) [15].

The cardiac role of autonomous nervous system has proven to be very important due to its nature of rapidly increasing or decreasing cardiac output with an incredibly fast adaptation, for example, to haemodynamic changes. But, as it is exposed first in motivation section, hypertension could reduce autonomic nervous system adaptation. If we combine this with a sympathetic nervous system dominating over parasympathetic, this could lead to a certain risk of undergoing a cardiovascular disease [2] [19]. Therefore, monitoring the activity of autonomic nervous system on the heart in hypertensive patients might be a powerful tool to reduce risks. The balance of the autonomic nervous system can be evaluated by HRV analysis metrics. So, we can relate the result of this analysis to the state of the autonomic nervous system and whether this can lead to some risk or not.

2.6 Heart Rate Variability

Before the beginning of the section, most of theory definitions, methods and keys has been obtained from Sörnmo & Laguna's book 'Bioelectrical signal processing and neurological application', in particular from chapter 8: 'ECG Signal Processing: Heart Rate Variability' [20]. This theory involves complex physiological and mathematics insights as well as key issues which will be summarised so as not to extend the section.

Heart rate modulating signals involving the brain and cardiovascular systems act upon the sinoatrial (SA) node of the heart influencing sinus rhythm. So, assessments of autonomic function reflect the ability of the system to stimulate the SA node. Heart Rate Variability (HRV) analysis is used to investigate in a non-invasive way the influence of autonomic nervous system activity on the SA node and, consequently, the control of autonomic nervous system on electric heart activity.

This analysis is based on measuring the heart rate, for which it is necessary to tag a timestamp for each heartbeat. Therefore, a series of occurrence times $a_0, a_1, ..., a_M$ is analysed by different techniques to inquire into the control of autonomic activity on the sinoatrial node. These techniques can measure different features in time, spectral or time-frequency domains.

The separate contributions of sympathetic and parasympathetic autonomic activity modulate the heart rate intervals (RR) of the QRS complex in the electrocardiogram (ECG) at different frequencies. Sympathetic activity is associated with the low-frequency range (0.04-0.15 Hz) of the variation frequencies

while parasympathetic activity is associated with the higher frequency range (0.15-0.4 Hz) of heart rate variation frequencies. Thus, it is possible to discriminate between sympathetic and parasympathetic nervous systems.

Since the aim is to investigate the sinoatrial node, the onset of the P wave could be an accurate fiducial point of the heartbeat. However, the onset of this wave is really complicated to determine accurately, many times the P wave is missing or showing very tiny. An alternative is the use of QRS complex, assuming that R timestamps is relatively fixed with P timestamps. Because QRS complex is the result of electrical ventricle depolarisation, some problems are introduced about the detection of sinoatrial activity using R labelling. Problems such as ectopic beats or noise artefacts can affect the detection of heartbeats and therefore the assessments of sinoatrial activity, as **figure 2.8** shows. RR timestamps are often represented as $t_0, t_1, ..., t_M$.

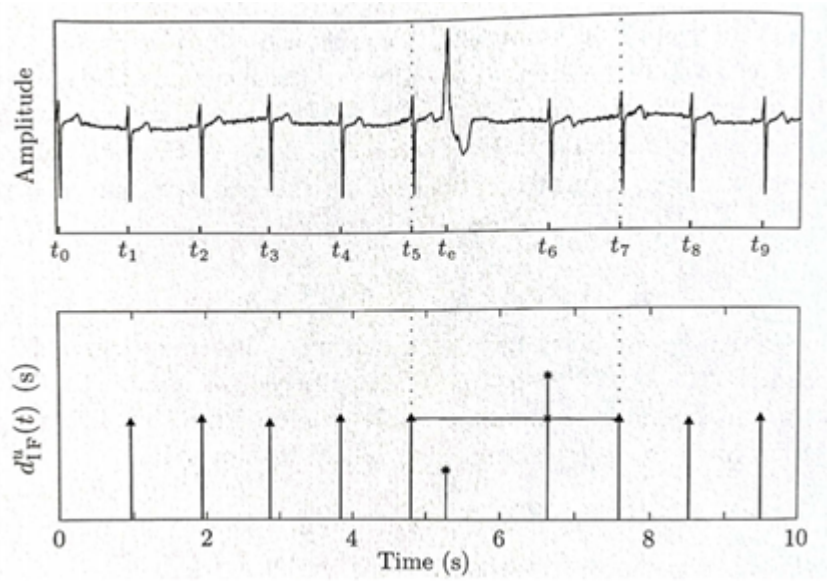


Figure 2.8: . Illustration of an ectopic beat as outlier of sinus beats and its correction of RR series. Obtained from 'Bioelectrical signal processing and neurological application' [20].

Once these problems are overcome, HRV methods analyse the so-called Inter-beat Interval (IBI) or NN interval as well, when NN denotes the interval between 2 sinus beat timestamps. When referring to interval between 2 heartbeats labelled in ECG before different corrections, it is named RR interval.

As far as data acquisition concerns, the sampling rate of ECG monitoring should be at least in 250 Hz. Lower sampling rate could lead to an inaccurate measure of beat-to-beat variations. However, certain databases like Holter recordings are digitalised in a lower rate. When there is such a lower rate, the recordings can be resampled to increase the frequency.

Due to the fact that HRV requires the location – or rather labelling – of sinus beats and outliers like ectopic beats, heartbeats morphologies must be clustered to separate sinus beats from others and to correct the missed sinus beats. Usually, the cluster for sinus beats is the one with the largest number of beats and so it is assumed. There are many different methods and those employed will be explained later in methodology section.

Once sinus beats and outliers have been labelled, beats labelled as 'erroneous' have to be handled. If not dealt with, HRV analysis could lead to wrong results. For example, spectral analysis like power spectrum may manifest disturbances similar to white noise, altering high frequencies (**figure 2.9**).

Many techniques can be used to deal with ectopic beats and correct them, such as modifying by median, interpolating or just deleting the outliers. It should be noted that usually the ectopic beat implies

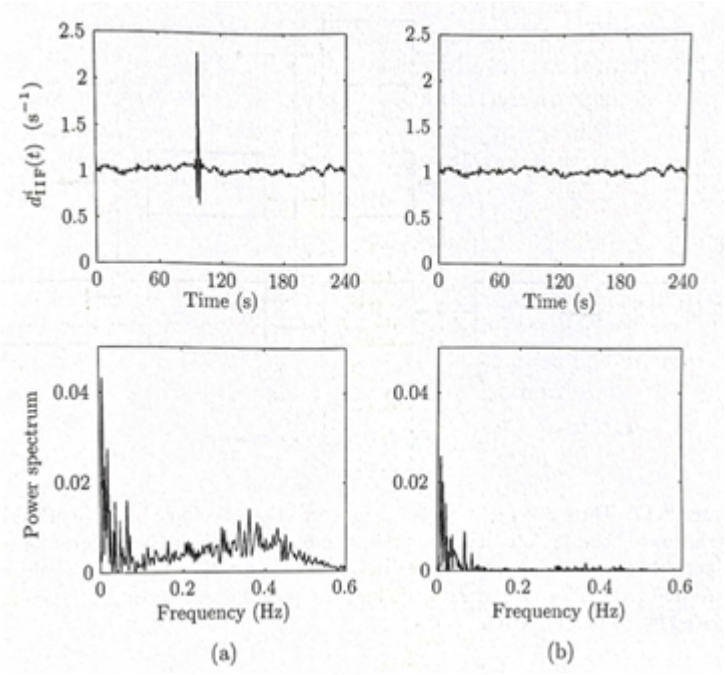


Figure 2.9: Perturbation of spectral features due to an ectopic beat. Obtained from 'Bioelectrical signal processing and neurological application' [20].

an abnormal RR interval but the next RR interval as well, as **figure 2.8** suggests.

Chapter 3

Material and Methodology

The following sections present the methodology used to corroborate whether a short time HRV analysis can be as valid as a long time HRV analysis for classifying hypertensive individuals into high or low risk groups of developing cardiovascular events.

The preparation of the data to train the different Machine Learning models was carried out using Matlab mathematical software [21]. These data correspond to different HRV analyses from the ECG-Holter samples used, which are explained below. The training, validation and testing of the different Machine Learning classification models as well as their statistics were performed using at first Matlab software for a quick approach and later Python software [22], specifically the packages Numpy and Scikit-Learn.

The same dataset will be used for both goals explained in Intro section, however, some steps in the methodology will differ. The application of machine learning techniques will be slightly different between the first section - replicating and validating the hypothesis and methodology in [3] - and the second one - comparing short-term HRV prediction vs other times in HRV analysis. Forward details are provided along this chapter and chapter for results and discussion.

3.1 Database and Physionet

For the elaboration of this project, a database of nominal 24-h electrocardiographic (ECG) Holter recordings was obtained from Phisionet. Physionet is an online resource launched in 1999 by Beth Israel Deaconess Medical Center (BIDMC) in Boston, Boston University, McGill University and the Massachusetts Institute of Technology (MIT) under the auspices of the National Center for Research Resources of the National Institutes of Health. This includes a large collection of different medical data, an open-source library with a large number of functions called PhysioToolkit and a forum to connect and exchange information among the scientific community in medicine [23].

The database - Smart Health for Assessing the Risk of Events via ECG (SHAREE) [3] - consists of 139 samples of ECG Holter recordings, plus other details such as anonymised patient information and details about vascular characteristics evaluated by cardiac and carotid ultrasonography. Patients were treated and samples were obtained in the Centre of Hypertension of the University Hospital of Naples Federico II, Naples, Italy.

These are the details for the database used in this project:

- Patients aged over 55 years - 49 female and 90 male - followed up for 12 months after the record-

ings and labelled whether they suffered major cardiovascular and cerebrovascular events: fatal or non-fatal acute coronary syndrome including myocardial infarctions, syncopal events, coronary revascularization, fatal or non-fatal stroke and transient ischemic attack.

- 17 patients experienced a recorded event (11 myocardial infarctions, 3 strokes, 3 syncopal events). Because only 17 patients out of 139 experienced a cardiovascular event, the information from this dataset is highly unbalanced.
- Each Holter recording contains three ECG signals: derivations III, V3 and V5, each sampled at 128 samples per second with 8-bit precision. While they are nominally 24 hours, it has an actual duration of about 22 hours. Also, it was performed after a one-month anti-hypertensive therapy wash-out.
- The recordings are accompanied by QRS annotations of the III ECG derivation obtained by an automated detector (WQRS) [24] but not corrected manually. The start of the recordings are also indicated.
- Recordings are supplied with demographic and clinical information (**table 3.1**) age, gender, weight, height, body surface area, body mass index, smoke or not, values of systolic and diastolic blood pressure as well as eventual vascular event.

Table 3.1: Demographic and clinical information for patients (mean and std).

Gender	Event	Number (n)	Smoker (y/n)	Age	BSA	BMI	SBP	DBP	IMT MAX	LVMi	EF
Male	Normal	81	27/54	72 ± 7	1.94 ± 0.14	27.53 ± 3.48	135.70 ± 17.95	77.80 ± 8.90	2.39 ± 0.69	133.77 ± 23.83	57.52 ± 11.29
	MI	6	3/3	77 ± 7	2.00 ± 0.15	28.46 ± 3.40	143.17 ± 20.93	74.00 ± 10.10	2.98 ± 1.40	151.33 ± 23.37	56.00 ± 7.77
	stroke	2	1/1	70 ± 4	1.71 ± 0.05	22.72 ± 0.44	127.50 ± 10.61	75.00 ± 7.07	2.65 ± 0.21	122.00 ± 5.66	43.50 ± 26.16
	syncope	1	1/0	76	2.02	28.73	120.00	70.00	3.80	184.00	33.00
	Overall	**90**	**32/58**	**72±7**	**1.94±0.14**	**27.49±3.48**	**135.84±17.98**	**77.40±8.89**	**2.45±0.76**	**135.38±24.36**	**56.77±11.72**
Female	Normal	41	8/33	71±6	1.81±0.17	27.81±4.71	138.40±22.50	73.40±8.93	2.24±0.75	122.95±29.21	62.82±9.47
	MI	5	0/5	71±8	1.81±0.20	29.47±6.63	151.00±29.24	77.00±4.47	1.55±0.49	133.00±23.09	64.00±5.24
	stroke	1	0/1	74	2.02	34.67	140.00	80.00	2.30	101.00	72.00
	syncope	2	0/2	78±1	1.62±0.10	23.66±1.90	140.00±42.43	60.00	1.90±0.57	141.00±7.07	67.00±4.24
	Overall	**49**	**8/41**	**71±6**	**1.80±0.18**	**27.95±4.92**	**139.81±23.37**	**73.35±8.82**	**2.17±0.74**	**124.35±28.00**	**63.33±8.90**
Dataset		**139**	**40/99**	**72±7**	**1.89±0.17**	**27.66±4.04**	**137.22±20.01**	**75.99±9.02**	**2.35±0.76**	**131.41±26.09**	**59.13±11.16**

BSA: body surface area (m^2), BMI: body mass index (kg/m^2), SBP: systolic blood pressure ($mmHg$), DBP: dyastolic blood pressure ($mmHg$), IMT: intima media thickness (mm), LVMi: left ventricular mass index (g/m^2), EF: ejection fraction (%), Weight (kg), Height (cm).

3.2 Extraction and preprocessing of IBI series

Samples of the database were identified and sorted using the identifier number given by the collectors and researchers of University Hospital of Naples Federico II. ECG lead III was the one chosen to calculate NN series of every time section analysed by using R peaks as references. According to the time span we would like to analyse, a 'stationary'[1] time section of this register was randomly chosen and extracted to work with. Next, since recordings are sampled at 128 Hz and guidelines about HRV analysis suggest a minimum frequency of sampling at 250 Hz [25], this signal was resampled to double the frequency of sampling. Therefore, the new frequency was 256 Hz. While the signals were randomly chosen between the 24 hours Holter recording in [3], and this is how it will be done in the first section, in the second section only signals between sleep time will be appreciated.

[1]This term is used here to refer to a time vector that does not present large variations or interferences due to a disturbed context. So, restless time in patient's day were avoided.

Once the signal had been resampled, R peak indexes were detected by Pan-Tompkins algorithm. Then, RR time series were obtained calculating the difference between adjacent R points, dividing by the frequency and multiplied by 1000 to obtain a vector of time differences between adjacent R peaks in milliseconds. However, this vector is not a valid representation of NN series. As previously explained in HRV theory section, RR time vector must be corrected from outliers, missing values and ectopic beats. This was mostly made manually, using only a few tools to visually identify possible wrong values. After these values were identified, they were corrected by linear interpolation of neighbours and saved to later HRV analysis. Further explanation of these methods is provided in the following subsections.

3.2.1 Pan-Tompkins algorithm

This method is highly appreciated due to its robustness, with a high precision. Pan-Tompkins algorithm can be easily used thanks to an implementation in Matlab [26]. To detect R peaks, Pan-Tompkins algorithm focus on frequencies in the interval 5-15 Hz. This bandwidth is related with the maximum spectral power of QRS complex **figure 3.1**. Thus, filtering the ECG signal between 5-15 Hz helps to better discrimination of R peaks. Pan-Tompkins algorithm is implemented by default to a sampling frequency of 200 Hz, nevertheless, this is not the case. For other sampling frequencies, the implemented function of the algorithm calls a zero-phase forward-and-reverse digital IIR filtering, known as filtfilt, using a 6th-order bandpass Butterworth filter.

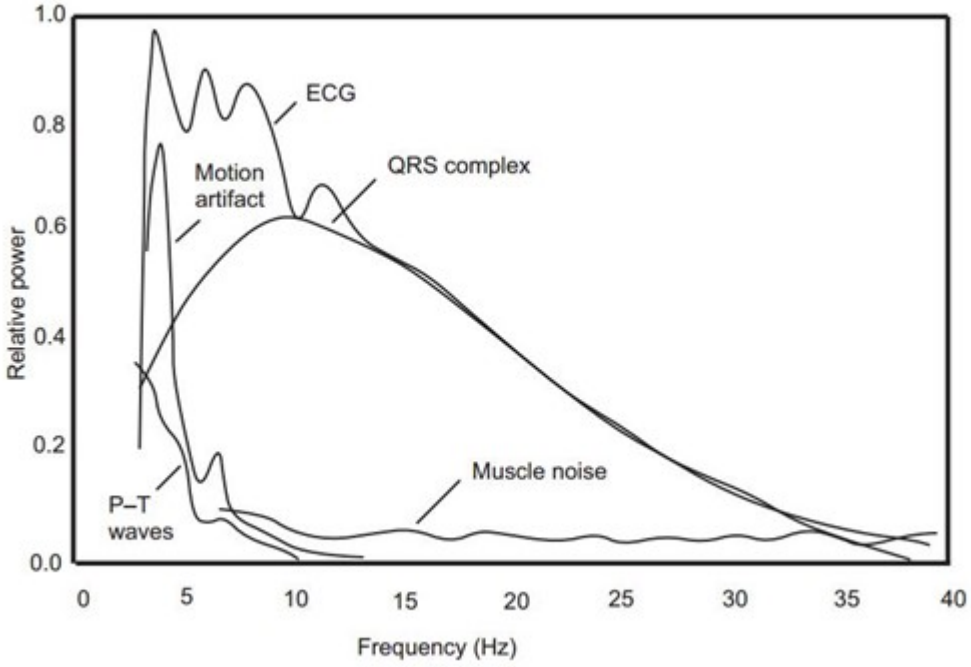
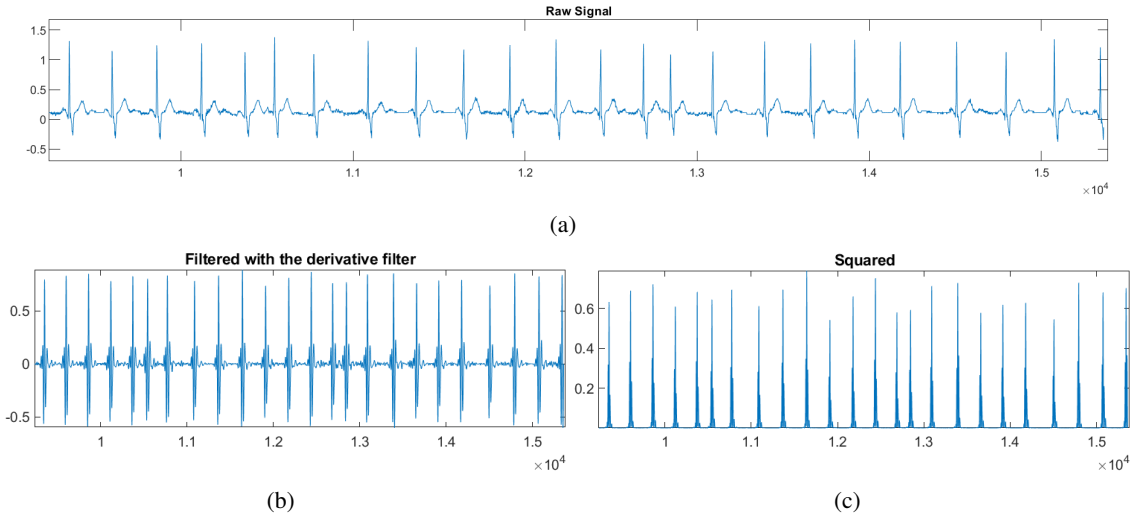


Figure 3.1: Relative power spectra of QRS complex, P and T waves and muscle noise and artifacts.

Later, QRS complex are highlighted by a 7th-order derivative filter due to our sampling frequency. The previous signal is filtered with 'filtfilt' filtering process. Then, signal is squared, so high amplitudes are more emphasised. The following **figure 3.2** shows the resulting signal from this process.

At this point, previous steps have generated a roughly pulse-shaped waveform. The decision to whether a pulse corresponds to a QRS complex (as opposed to a high-sloped T-wave or a noise artefact) or not is performed with an adaptive thresholding operation and other decision rules.

First, the waveform is processed to a set of weighted of unit samples in order to localise the QRS



(a) (b) (c)

Figure 3.2: (a) ECG signal before preprocessing (raw). (b) Filtered signal by bandpass filter and derivative filter. (c) Resulting signal after filtered and squared.

complex to a single instant of time to act as fiducial marks. Later, two threshold values are calculated - THR_{SIG} and THR_{NOISE} – which are constantly adapting. When a peak is detected, it must be classified as either a noise peak or a signal peak. To be considered as a signal peak, it must be greater than THR_{SIG}; otherwise, greater than THR_{NOISE} if it requires a searchback to find the QRS complex. If a detected peak is greater than THR_{NOISE} but lesser than THR_{SIG}, then it is considered noise. The thresholds are adapted by some values which are updated every time the moving window classifies a new peak. An example is shown in **figure 3.3**

Because all this conditions, it might be possible QRS peaks are not detected in a unreasonably long period, if so, then the algorithm performs a searchback for missed QRS complexes within a section of 1.66 times the current RR time period. The highest peak between both thresholds is considered the true QRS complex.

The implemented Pan-Tompkins function also can eliminate other possible fake detected QRS complex if they are in the refractory period of 200 millisecond after the previous one. Also, it can perform some calculations to discriminate if a detected peak is an abnormally prominent T wave. Further information can be found in [26].

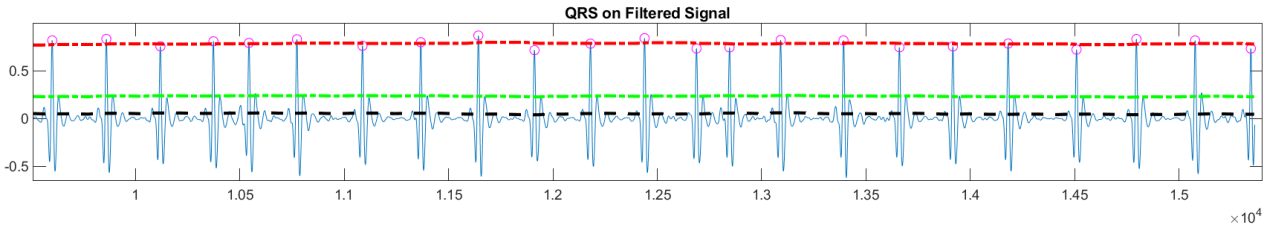


Figure 3.3: Detection of R peaks using the thresholds which are constantly adaptating throughout the moving window.Red line: THR_{SIG}; Black line: THR_{NOISE}; Green line: an adaptive threshold combination of the above two.

3.2.2 Preprocessing RR series

Once R peaks were detected and their indexes were known, a vector of R timestamps was calculated dividing R indexes by the sampling frequency of 256 Hz. Then, RR series could be visualised, and possible erroneous values detected. There are many automatic methods that can be used, however, they are only a mere approximation of a properly detection and manual supervision of the results is mandatory.

First, if a value was higher than 2000 ms or lower than 345 ms, it was considered physiologically wrong. A detection algorithm was also created to detect possible outliers in two modes, depending on the shape of the RR series.

The first mode was detrending the RR series - so a possible trending line could not alter the location of outliers - and detecting all values which were higher or lower than three times the standard division of the set. The detrending were made by interpolating the values with a high grade (15th) polynomial to extract an accurate trending of the time series and removing it from the series.

However, some RR series had different issues that could likely distort the use of the standard division, for example, many outliers or being very flat. Then, other mode could be used. Instead of removing the trending line from the series, it was used to calculate both positive and negative threshold of 10% of the interpolated trending line. Next, values outside this area would be considered erroneous and labelled as outliers.

Also, a tool to detect premature ventricular contractions (PVC) were used from Physionet Cardiovascular Signal Toolbox [27]. The detection of PVCs is based on the application of a convolutional neural network (CNN) to the wavelet transform (WT) of the raw ECG channel [28]. Segments of the ECG channel are turned into a 2D time-frequency image. A CNN model had been trained and provided by authors to detect ectopic beats. The authors claim to have a results of 85% F1 score and 97% accuracy. However, PVCs detection was overseen manually and compared with the other used tools, because reliability of the method decreases in short-term signals.

After having identified the erroneous values, they were manually revised to state if they were truly outliers. It a outlier was found higher than 2000 ms, it was usually due to a missing R peak. Then, R peak was manually detected and added to the vector. Values next to detected outliers due to ectopic beats were also labelled as outliers. As it is explained in HRV theory section, ectopic beats alter the next beat, usually blocking the normal conduction of electricity; therefore, this beat must be corrected too.

Outlier timestamps were declared as NaN value and corrected by linear interpolation of their two closer true values. Alongside this process, details of the supervision and editing of the RR timestamps were saved in different arrays in case of being necessary at a later steps.

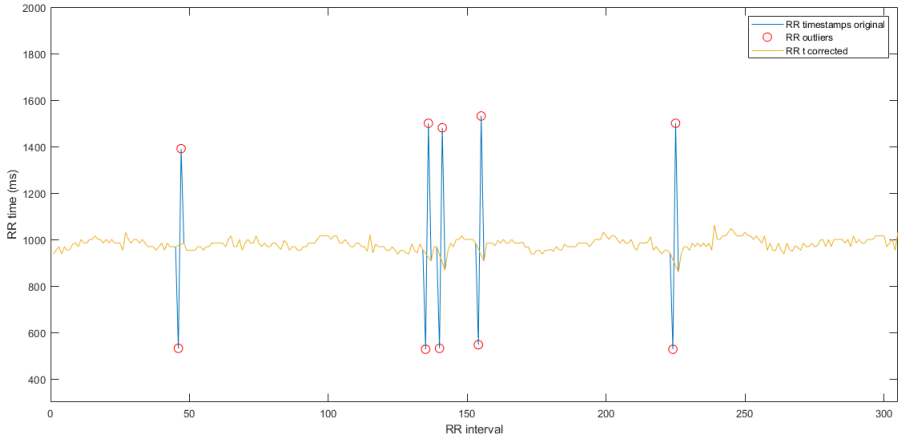


Figure 3.4: Correction of erroneous RR intervals due to ectopic beats by linear interpolation.

3.3 HRV analysis parameters

Several types of HRV analysis were performed to study the randomly chosen signal sections. HRV analysis could be classified as linear analysis in time and frequencies domain as well as non linear analysis. The different analysis may change whether studying short or long-term HRV as according to International Guidelines [25].

3.3.1 Time domain HRV measures

Standard linear HRV analysis in time-domain was performed according to methodology in [3] and standard guidelines [25]. Time domain methods could be grouped into statistical or geometrical. Since many methods are correlated with others, the most used are the statistical methods SDNN and RMSDSD, which represent overall components of HRV and short-term components of HRV, respectively; geometrical methods such as HRV triangular index are also highly appreciated [25].

- **AVNN**. The average of all NN intervals.
- **SDNN**. Standard deviation of all NN intervals.
- **RMSSD**. The square root of the mean of the sum of the squares of differences between adjacent NN intervals.

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N-1}((R-R)_{i+1}-(R-R)_i)^2} \tag{3.1}$$

- **NN50**. Number of pairs of adjacent NN intervals differing by more than 50 ms in the entire recording.
- **pNN50**. NN50 count divided by the total number of all NN intervals.
- **HRVTi**. HRV triangular index (HRVTi) is the total number of all NN intervals divided by the height of the histogram of all NN intervals measured on a discrete scale with bins of 1/128 seconds. **Figure 3.5**.

$$HRVTi = total\,number\,of\,all\,NN\,intervals/Y \tag{3.2}$$

- **TINN**. The triangular interpolation of NN interval histogram is the baseline width of the distribution measured as a base of a triangle. This triangle is found by minimum square difference. As seen in **figure 3.5**, values N and M are established on the time axis and a multilinear function q constructed such that q(t)=0 for t $\leq$ N and t $\geq$ M and q(X)=Y, where the following integral is the minimum among all selections of all values N and M.

$$q(t) = \int_0^\infty (D(t)-q(t))^2\,dt \tag{3.3}$$

Then, the triangular interpolation is calculated as:

$$TINN = M - N \tag{3.4}$$

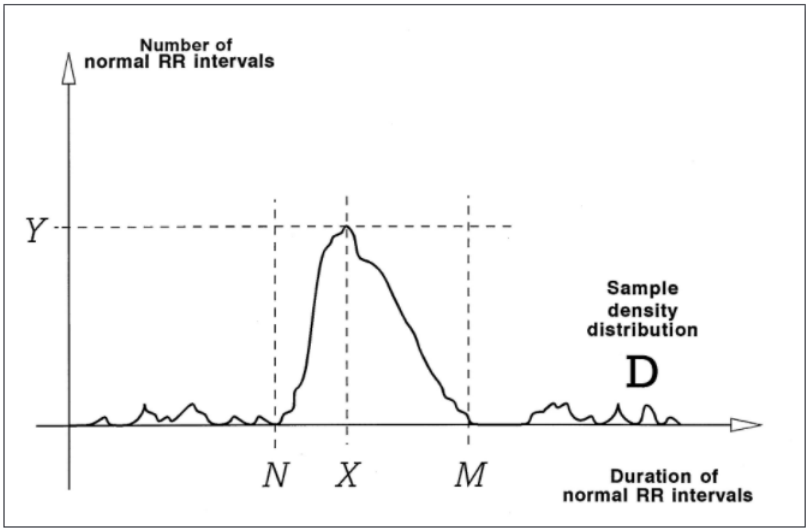


Figure 3.5: Geometric measures of NN interval histogram: HRVTi and TINN. Where D is the sample density distribution of NN interval. HRVTi is obtained dividing the area D by the maximum Y. Obtained from [25].

3.3.2 Frequency-domain HRV measures

In short-term HRV analysis,three main spectral components are distinguished. This components or frequency bands are the very low frequency (VLF, 0–0.04 Hz), low frequency (LF, 0.04–0.15 Hz) and high frequency (HF, 0.15–0.4 Hz). In long-term HRV analysis, another spectral component is distinguished, the ultra low frequency (ULF, 0-0.003 Hz). The power spectral density (PSD) is computed with Lomb-Scamble periodogram.

- **VLF measures**. Absolute power, relative power among all ranges and peak frequency in very low frequency band.
- **LF measures**. Absolute power, relative power among all ranges and peak frequency in low frequency band.
- **HF measures**. Absolute power, relative power among all ranges and peak frequency in high frequency band.
- **Normalised power**. Also, LF and HF band powers are presented in normalized units. This emphasizes the controlled and balanced behaviour of sympathetic and parasympathetic nervous system.
- **LF/HF power ratio**. Calculated as the division between power values of LF/HF.
- **TP**. Total power of the signal.

3.3.3 Non-linear HRV measures

- **Poincaré Plot measures**. Poincaré Plot is a return map used usually to quantify periodic functions. As a scatter plot of successive RR intervals (**figure 3.6**), it produces two standard deviation measures, SD1 - describing short-term variability - and SD2 - describing long-term variability analysis. This features can be used to describe NN series as well as parasympathetic nervous activity [29].
- **Approximate Entropy**. Approximate Entropy (ApEn) quantifies the randomness of time-series data without any previous knowledge [30]. ApEn algorithm needs of a equally sampled data. The

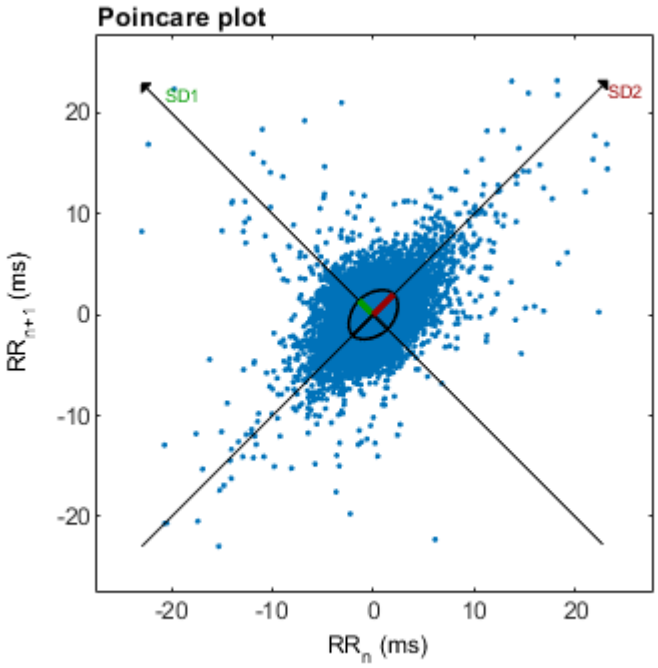


Figure 3.6: Poincaré Plot. Beat-to-beat intervals are represented in relation to the next beat-to-beat interval. SD1 and SD2 are the standard deviations in both orthogonal directions of the plot.

defined embedding dimension was 2 and its tolerance 0.2 times the standard division (SDNN) of NN interval [3].

- **Sample Entropy**. Similar to Approximate Entropy method, Sample Entropy (SampEn) is an algorithm for determining the regularity of series of data based on the existence of patterns. In contrast to ApEn, the result of this algorithm does not widely vary depending of the tolerance used and it does not depend on the length of the data serie, so this is less biased. The embedding dimension and tolerance was also 2 and 0.2 times SDNN.

- **Correlation Dimension**. The correlation dimension (CD) of a data series is the measure of dimensionality of the space occupied by their points. It is useful to measure the complexity of a signal system and provide information of the minimum number of dynamic variables needed to model that system. CD values for HRV have been reported between 4 and 10 and the needed data points (N) to calculate it must be at least $10^{CD/2}$. Computation of CD is defined by two main parameters, the embedding dimension (m) and a time delay (τ). Embedding dimension is defined as at least $2CD+1$, hence, it is set usually at least at 10. Meanwhile, higher time delay values increase CD outputs, but this values tend to saturate at a $\tau \geq 5$ [31]. Therefore, the embedding dimension is set at 10 and the time delay at 5.

- **Detrended Fluctuation Analysis**. This method measures the correlations within the signal at different time scales, determining the statistical self-affinity of a signal expressed as $F(n) \propto n^{\alpha}$. The value alpha is obtained for short-term fluctuations within range 4–16 beats (α_1) and long-term fluctuations within range 16-64 beats (α_2) [32].

- **Recurrence Plot measures**. The recurrence plot technique projects a time series in a higher dimensional space, called phase space, creating a matrix where each row and each column is a point in

the phase space, and each element of the matrix is the respective distance [3] [33]. Several features are studied using this technique.

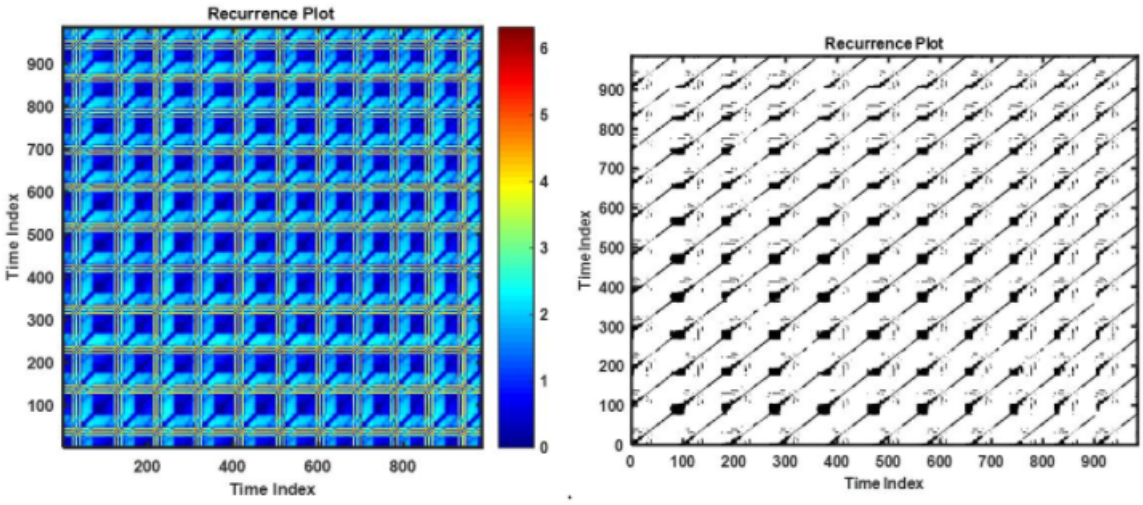


Figure 3.7: Recurrence plots for a ECG using Recurrence Quantification Analysis Matlab toolkit [34].

– **Recurrence rate**. When the distance between each point is smaller than a given threshold r, that means two points are close in the phase space and a so called recurrence (R) occurs. The recurrence rate (REC) is calculated then as the density of recurrence points in the recurrence plot:

$$REC = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} R_2(i,j) \tag{3.5}$$

– **Maximal length of lines**. This feature (L_{max}) is an easily computated approximation of Lyapunov exponent, which characterizes the rate of separation or divergence of a system sensitive to its initial conditions. In recurrence plots, lines can be defined as series of diagonally adjacent recurrences, where the length l of a line is the number of points which the line consists of. Then, (L_{max}) is the maximal length of lines.

– **Mean of the length of lines**. A similar measure is the mean of length values (L_{mean}).

– **Determinism**. The determinism (DET) is related with the predictability of the dynamical system. This helps to distinguish from noises, such as white noise. This feature shows the percentage of recurrence points which form diagonal lines in the recurrence plot of minimal length.

$$DET = \frac{\sum_{l=2}^{l_{max}} l \cdot N_l}{\sum_{i=1}^{N} \sum_{j=1}^{N} R_2(i,j)} \tag{3.6}$$

– **Shannon Entropy**. Shannon Entropy (ShEn) is the entropy method used in information theory and it measures the uncertainty of a system. In this case, the sum of the percentage of N_l over all the number of lines (n_l) in the recurrence plot.

$$ShEn = \sum_{l=l_{min}}^{L_{max}} n_l * \ln n_l \tag{3.7}$$

3.4 Machine Learning for classification

Nowadays, there are a multitude of mathematical methods for finding statistical differences to group and label data according to their patterns. This models usually are considered in 4 groups based on the method of learning: supervised, unsupervised, semi-supervised and reinforcement learning.

This Master's dissertation aims to classify different samples using their patterns. For this purpose, the best type of models are the ones prepared by supervised learning. Supervised learning models are algorithms trained with previously labelled data, which labelling is usually manual. Using this labelled data, algorithms analyse different patterns to distinguish among labels and later are tested with more labelled data (whose labels they do not know) to evaluate their performance. There are different metrics to evaluate the performance of a classification models. These metrics alongside different classifier models are explained in this section.

3.4.1 Developing M-L models

The process of preparing a classification supervised-learning model consists of various steps. Because this technique uses labels to improve the classification models, models cannot be tested using the same samples with which they were trained. Otherwise, it will bring wrong results due to a biased evaluation. Therefore, the dataset usually is split between data to train the models and data to test them. The most simple technique is called hold-out split and the data is randomly separated between both set once (in **figure 3.8** the hold-out technique split the dataset in a proportion 80%-20%).

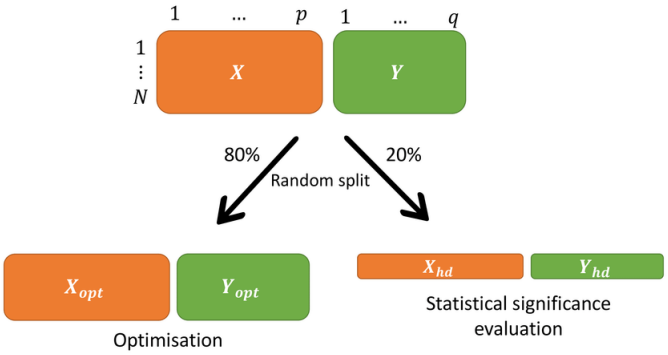


Figure 3.8: Dataset split between train and test groups by hold-out technique.

However, sometimes it is important to find the best features o hyperparameters for a M-L model. In this case, the set of models with different hyperparameters cannot be evaluated using the test set, because then the final model would be biased and the outcome performance metrics wrong. To find the best hyperparameters, the training set is split again between training and validation set. This can be done by hold-out technique, but to ensure that the performance of the models is not altered by the random data they are trained on, cross-validation technique is usually recommended.

In cross-validation technique, the dataset is randomly split into 'k' groups or folds, usually 5 or 10. Then, the M-L model is trained with k-1 groups and validated with the other group. This is applied k times, until every group was used to test the performance of the trained model (**figure 3.9** shows a 5-fold cross validation). Then, the performance metrics are averaged and the final tuned model with the final hyperparameters can be tested with the test set. When the value of k-folds is equal to the number of training samples, this type of validation is also called leave-one-out. Then, every fold or group consists of a single sample.

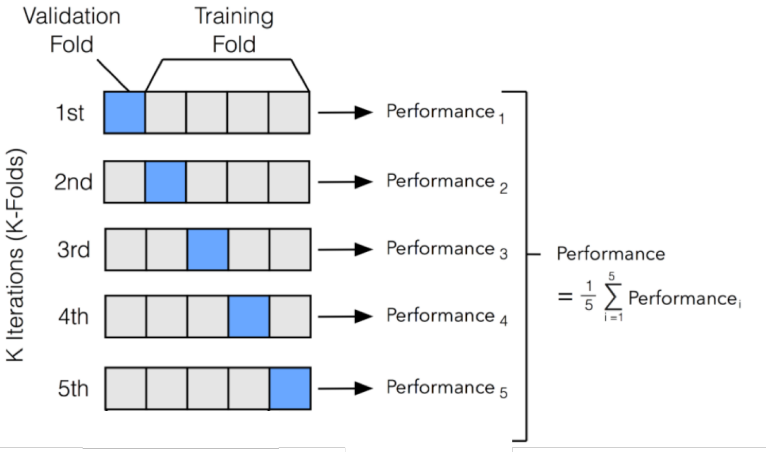


Figure 3.9: Cross validation technique.

3.4.2 Classification models

Different classification models were trained to predict an accurate classification between high or low risk in hypertensive people. First, Random Forest and Support Vector Machine algorithms will be used to test the viability of the methodology proposed in [3] as they were the two best models proposed. Also, Gaussian Naive Bayes approach, K-N Neighbours and Logistic Regression will be used. It should be noted that some machine learning models require data to be pre-processed to avoid possible statistical imbalances, also known as feature scaling.

- **Random Forest Classifier.** This algorithm was proposed by Leo Breiman [35] who has contributed statistics and machine learning fields with boostrap aggregation methods and random forest algorithm.

  Boostrap aggregation, also called bagging, is an ensemble learning method for approaching results of a repeated model, reducing their variance and overfitting. It is common to use with decision trees (**figure 3.10**), where several decision trees are trained and their classification averaged by different methods, such as majority votes. Decision trees are methods which results are calculated by simple rules of limits such as greater or less than a value, or categorical rules [36].

  However, this reduction of variance and overfitting reach a limit because all decision trees are trained the same way, so are they strongly correlated. Random forest was designed to solve this problem. Opposite to boostrap aggregation, every decision tree in random forest chooses a random subset of features, so results from different decision trees are weaker correlated because of using different subsets, **figure 3.11**.

- **Support Vector Machines.** Shortened with the acronym SVMs, this method applies the Statistical Learning Theory. SVM models are very useful when it is not possible to distinguish accurately different classes of a dataset which features are represented in a plane but when this features are represented in a hyperplane, the classification improves [37], **figures 3.12**.

  The extent of features to a higher dimension is calculated by a Kernel functions, like Polynomial, Gaussian or Radial-Basis Function. Moreover, SVM models are notable for their robustness to high dimensional data. Besides, SVMs also standout for being difficult and sensitive to a proper parameter tuning. Linear, 2nd and 3rd degree Polynomial, Radial-Basis and Sigmoid Kernel functions were evaluated for prediction in the different time series used in this project.

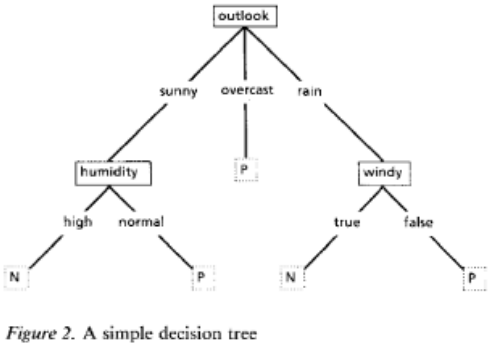


Figure 3.10: Example of a simple small decision tree of 3 leaves. Retrieved from [36].

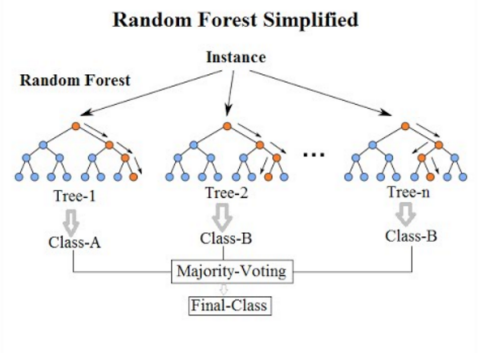
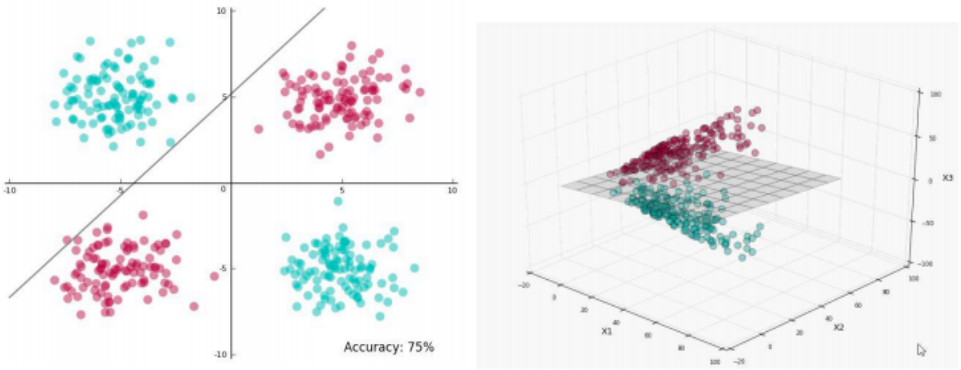


Figure 3.11: Scheme of the aggregation of different decision trees to create a random forest.



(a) (b)

Figure 3.12: (a) Discrimination of data in a 2D-plane using linear function. (b) Stronger discrimination of data in a 3D-hyperplane using linear kernel function.

- **Gaussian Naive Bayes**. Naive Bayes classifier is a probabilistic model based on applying Bayes' theorem while assuming all predictors are independent. A sample is classified by using either a plurality vote approach or some averaging rules to calculate the probability of belonging to the different classes in the dataset [38]. This probability is compared with a threshold to classify the sample into a class.

  For this methodology, due to features are continuous data, it is assumed their distribution corresponds to a normal distribution, also called Gaussian. Then, data is segmented by classes and each mean and variance of segments are computed. Each test sample is classified by their probabilities or scores in the normal distribution.

- **K-Nearest Neighbours**. A non-parametric method which compute the distance between a sample and their k nearest neighbours samples and adopt their predominant value or label as the prediction. There are different types of distances and algorithms to calculate which samples are the nearest neighbours. The most importat key in this sort of models is the value of k - the number of neighbours to use - and how the distance is computed: Manhattan, Euclidean, Minkowski ... [38].

- **Logistic Regression**. This algorithm transform the ability of prediction continues variables by polynomial regression to a probability in a binary discrete space, 0 or 1. Therefore, this model is usually used for binary classification, like this project. By a logistic function, the probability of one class or another tends to 0 or 1 depending on the input, **figure 3.14** .

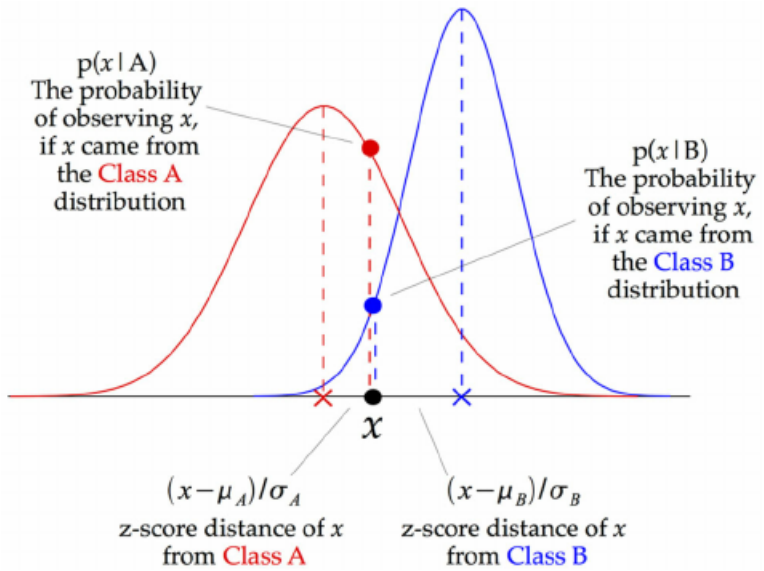


Figure 3.13: Labelling a binary problem using Gaussian distribution.

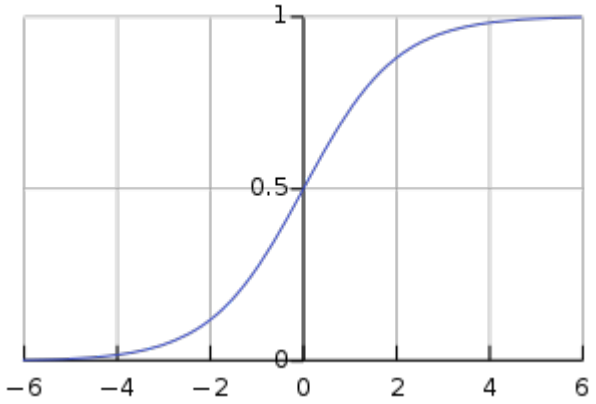


Figure 3.14: Increase in the outcome of two probabilities using a logistic function.

3.4.3 Statistical analysis, feature selection and performance of classification models

Due to the high number of HRV features (32), it might be likely that some of them are correlated. Then, some feature selection methods were applied to filter out redundant information. Some methods works best for estimating feature importance for specific machine learning models. These methods are available using MATLAB software.

Feature selection is a procedure to select the minimum number of attributes needed to represent the data accurately. The output is a set of features highly significant with the required prediction (labels in classification problem), yet uncorrelated with each other. By using relevant features, classification algorithms can in general improve their predictive accuracy, shorten the learning period and avoid possible overfitting of models, increasing generalisability of their predictions [39].

- χ^2 **statistics method [40].** It sorts the features by computing the p-values of the chi-square test statistics. The χ^2 value is set at a high significance level at the beginning, and then it is iteratively recalculated until the features are ranked by the following **ecuation 3.8**. This method ranks predictors individually.

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \tag{3.8}$$

where k: nº of classes; A_{ij}: nº patterns in i_{th} interval, j_{th} class; E_{ij}: = expected frequency of A_{ij}. The degree of freedom of the χ^2 statistic is one less the number of classes. Further explanation is provided in the original article [40].

- **Minimum redundancy maximum relevance (mRMR) method [41].** This method is based on the previous maximal relevance (Max-Relevance) feature selection algorithm. These Max-Relevance selected features which individually were the most relevance to the target class. However, this does not deal with redundancy problems. The proposed minimal-redundancy-maximalrelevance (mRMR) framework promises to select features with minimal redundancy but strong relevance to the target class for both continuous and discrete data sets.

- **Out-of-Bag, Predictor Importance Estimates by permutation using Random Forest**. This method measures how influential the predictor variables in the model are at predicting the response using RF classification model. Different trees are created with randomly permuted values of a predictor and samples, then the model error is estimated using the out-of-bag observations containing the permuted values. If a predictor is relevant for an accurate outcome, then the model error is increased due to permutation of its values. If the predictor is not important, the model error is little altered.

Once machine learning models are trained, there are different methods for assessing the performance and predictive ability of each classification model. A widely used method in classification problems is the use of confusion matrices and the metrics derived from it (**figure 3.15**). A confusion matrix represents de number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values predicted for a target class or label. By means of these outputs, metrics such as accuracy (acc), sensitivity (sen), specificity(spe) or F1-score represent the performance of the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.9}$$

$$Sensitivity = \frac{TP}{TP + FP} \tag{3.10}$$

$$Specificity = \frac{TN}{TN + FN} \tag{3.11}$$

$$F1 - score = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.12}$$

However, depending on the nature of the problem, some metrics may be more important than others in evaluating the model. Other useful method is graphing the Receiver Operating Characteristic (ROC) curve of the classification. The graph is related to the confusion matrix as this curve shows the relation between sensitivity and specificity. The area under the curve (AUC) is sometimes used as a metric of the performance of the model. From 0 to 1, a random model dependent on the luck will show a AUC approximated to 0.5, the highest the value the best the model is considered (**figure 3.16**).

$$AUC = \int_0^1 ROC(t)\, dt \tag{3.13}$$

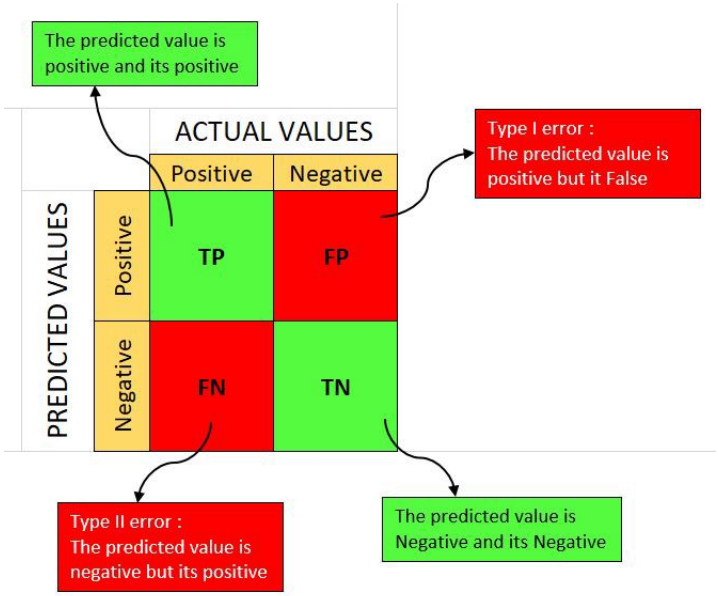


Figure 3.15: Representation of predicted outcomes by a classification model using a confusion matrix.

3.5 Part 1: Validating methodology applied in SHAREE research

Firstly, the conclusions published by The Multidisciplinary Department of Medical, Surgical and Dental Sciences of Second University of Naples in [3] were validated by a review of the dataset and methodology applied. Due to the characteristics of the dataset previously exposed at the beginning of this chapter, there could be some concerns about the feasibility of the results in other studies.

This project tried to replicate these results with the methodology proposed. 5 minutes short-term HRV analysis in ECG Holter signals were used to train the best two classification models exposed in [3]. These ML models were a RF and SVM classification models without feature selection (**table 3.2**). Hence, feature selection was not applied in this part of the project.

Table 3.2: Performance measurements estimated on the test set (hold-out estimation) of the best classifiers based on HRV features in SHAREE paper.

Model	Hyperparameters	Feature selection	AUC	ACC	SEN	SPE
RF	NT: 300, NF:5	None	88.8%	85.7%	71.4%	87.8%
SVM	G:1.4	None	90.1%	93.9%	71.4%	85.7%

Due to the size of the SHAREE dataset, the low-risk and high-risk classes are highly unbalanced, moreover, the high-risk class is not enough large to ensure a good generalisability to other cases or dataset. Methodology in the original publication [3] suggests using a technique of oversampling to balance classes (annex 2 of the paper), called SMOTE. Oversampling means creating 'synthesised' samples from the actually known samples in dataset. Oversampling the dataset could alter the results on the previous **table 3.2** and therefore how this technique was applied had to be reviewed. However, how this technique was used is only briefly explained in a annex associated with the paper.

This technique is named 'Synthetic Minority Over-sampling Technique' (SMOTE) and it averages the features of a labelled class by randomly selecting a subset of samples. To understand the possible risk of applying oversampling in this case, it is necessary to understand the proposed technique and how it was used.

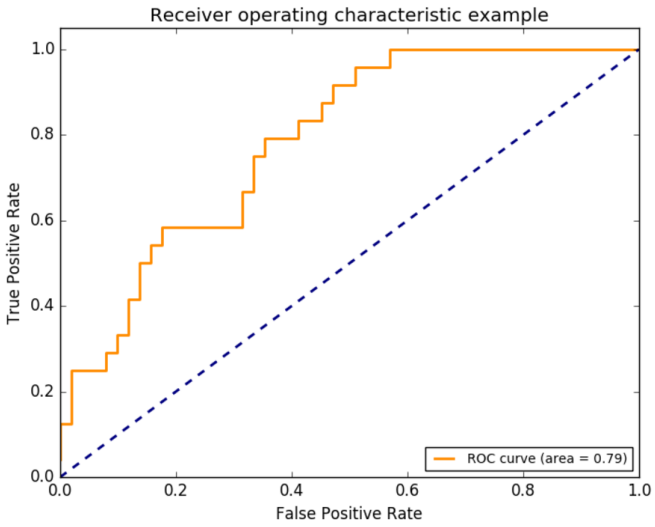


Figure 3.16: ROC curve graph and AUC metric for a example model and luck standard (dotted line).

3.5.1 SMOTE: oversampling technique

SMOTE needs two parameters: the number 'k' of nearest neighbours to randomly select and 'N' (N = n-1) where n is the number of times for oversampling as an integer. For example, if the required oversampling is 300% (3 times the size of the dataset for the specific class): N=2 and k=5, the algorithm randomly select 2 of the 5 nearest neighbours. Then, for every feature, takes the value of the sample and average calculating the vectorial distance with the selected neighbours and multiplying by a random number between 0 and 1, so 2 new synthesised samples of data are created [42]. It is also suggested to perform an undersampling of the majority class (reducing its samples) in the case it is required a big percentage of oversampling comparing to the number of samples of the minority class.

The outcome is a minority class which is oversampled, so the decision regions of the dataset which are useful to discriminate the classification are larger and less specific. This decision regions are more weighted by the previously minority class and could improve the results. Nevertheless, these decision regions or boundaries should be highly specific for the different classes and the minority class should be highly concentrated or the resulting oversampled minority class could turn out to be a propagation of the minority class through all the distribution, creating a fake result.

Unfortunately, how SMOTE technique was applied to the dataset in [3] is not explained, different approaches could have been done. For example, it could not be the same outcome to apply SMOTE technique first and later splitting the dataset between training and testing set as the other way round. Obtaining some synthesised samples from various real samples and later split randomly them to train and test the model could lead to great performance metrics but due to overfitting. It might be possible that the real decision boundaries to distinguish samples between classes are under-represented and the classification model only evaluate some of them. Also, decision regions could be blurred if possibles statistical outliers are not removed, so they are used to draw their supposed boundaries.

3.5.2 Summary of the approaches followed.

The hypothesis of using only 5 minutes of HRV analysis to evaluate risk of cardiovascular events in hypertense people was validated as feasible in the SHAREE publication [3] when it was compared with using cardiovascular anatomical measures from images to train predictive models. However, this conclusion was made assuming that oversampling the 'high-risk' class was right, no matter it was only 17

samples for this class. Then, the first part of the project aimed to study the distortion of this technique in the dataset. Therefore, the following steps were taken to evaluate whether using this oversampling technique in methodology was valid or not.

1. A first attempt was made, applying SMOTE technique to oversample every sample of the target ('high-risk') class. After oversampling, the new dataset was split into training and testing set.
2. Second attempt was made in the inverse manner. First the original dataset was split into training and testing set and then, both groups were oversampled individually. So, the origin of synthesised samples was not a mix.
3. Both oversampling approaches were tested using a new different dataset of only target class samples to validate the sensitivity of trained models using this methodology in new samples.
4. Multiples 5-minutes samples in similar context were obtained from every patient, collecting more samples from target class to balance the dataset. In this case, oversampling methodology was not applied. This approach is referred as 'manually oversampling'.
5. Finally, a comparison was made for the distribution of the values of the predictors (variables) of the samples between the SMOTE oversampled dataset and the manually oversampled one using whisker box-plots. Due to the large number of predictors, only the top-6 best ranked predictors in the original paper [3] were used.

3.6 Part 2: Comparing predictive ability of 5', 30' and 1 hour

Later, 5 minutes, 30 and 1 hour of HRV analysis will be compared. Because of the highly unbalance of the dataset as well as reasonable doubts about using oversampling techniques in this data, a different approach will be tried. Not oversampling technique will be used. However, to reduce the unbalance between classes, only 34 samples will be used, 17 for 'low-risk' class and 17 for 'high-risk' class.

In this case, all samples will be obtained during patient's sleep at night, which is assumed to be between 00 a.m and 06 a.m. HRV analysis during sleep stages is highly valued, due to HRV is relatively higher while different cardiovascular diseases can make HRV lower due to a loss in activation capacity of vagus nerves. [43].

Due to this approach, there is not a great expectation of high predictability in the classification models. The amount of samples will not be enough for the generability of their real distribution. But the models will not be highly-biased to one class or apparently faked. Nevertheless, generability of the predictive models is not what it is been looking for in this project. Applying this same methodology for the three lengths of time can establish a reasonable way to compare them. Hence, feasibility of predicting with 5 minutes HRV analysis can be ruled out or not for this framework and this dataset.

3.6.1 Summary of the methodology followed.

1. 17 patients from 'low-risk' class were randomly choosen from the 122 available patients.
2. 5 minutes, 30 minutes and 1 hour of ECG signal were extracted from Holter ECG of these patients and the 17 patients from 'high-risk' class. Extracted ECG signals were resampled to 256 Hz.
3. Each beat in ECG signals was labelled using Pan-Tompkins algorithm and manually overseen. Differences in the position/time of each label was save in a vector array and later preprocessed to remove ectopic beats, noise and outliers as explained previously.

4. Different time, frequency and non-linear HRV analysis were performed to calculate the 32 analysis features - predictors for the machine learning models - for every sample.

5. 'χ^2 statistics', 'mRMR' and 'Out-of-bag RF Permutation' feature selection methods were used to rank the top 6 features to train the models.

6. The 5 different classification models were tuned using the 34 samples by a cross-validation of 17 folds - each fold containing 1 sample of each class. Then, the best models with the best hyper-parameters were tested using a leave-one-out approach. This was done for every feature selection method.

Chapter 4

Results and Discussion

In this chapter, results to validate or reject the hypothesis are presented and it is discussed whether it is feasible or not to obtain an accurate classification of high and low risk in developing cardiovascular diseases using HRV analysis of only 5 minutes short-term ECG signals.

First, results presented in SHAREE paper [3] are evaluated and the same results are intended to be achieved. However, an important part of the methodology used is the oversampling of the data to obtain 'synthesised samples' using SMOTE oversampling technique [42] and the methodology has to be validated whether this technique might distorted results or not.

Then, using the same dataset, the hypothesis will be assessed comparing HRV analysis on different length of ECG signals and performing different data science methodology to reduce as maximum as possible the weaknesses of this dataset.

4.1 Results

First, an attempt was made to replicate the results obtained in SHAREE paper. A set of samples consisting of 32 features of HRV Analysis from randomly selected 5-minutes segments of ECG were split in 60% train group (122 samples) and 40% test group (82 samples). The high-risk class was oversampled 500% from 17 to 102 samples using SMOTE technique [42] following the procedure in [3]. At the mean time, the low-risk class - the major class - were undersampled to 102 samples deleting randomly some samples, as suggested in original oversmapling technique publication [42]. Then, both classes are equally balanced.

The two best models proposed in [3] were trained by 60% dataset and then tested by the other 40%. As **table 4.1** shows, results are similar as the original SHAREE research. For both models, AUC, true positive rate and false positive rate are plotted in **figure 4.1** as well as their confusion matrix.

Table 4.1: Performance measurements estimated on the test set (0.6-0.4 hold-out estimation) for 5' short-term HRV with SMOTE technique proposed in [3].

Model	Hyperparameters	AUC	ACC	SEN	SPE	F1_score
RF	NT: 300, NF:5	97%	89.02%	95.12%	82.93%	89.66%
SVM	G:1.4, k:poly, d=3	89%	84.15%	100%	68.29%	86.32%

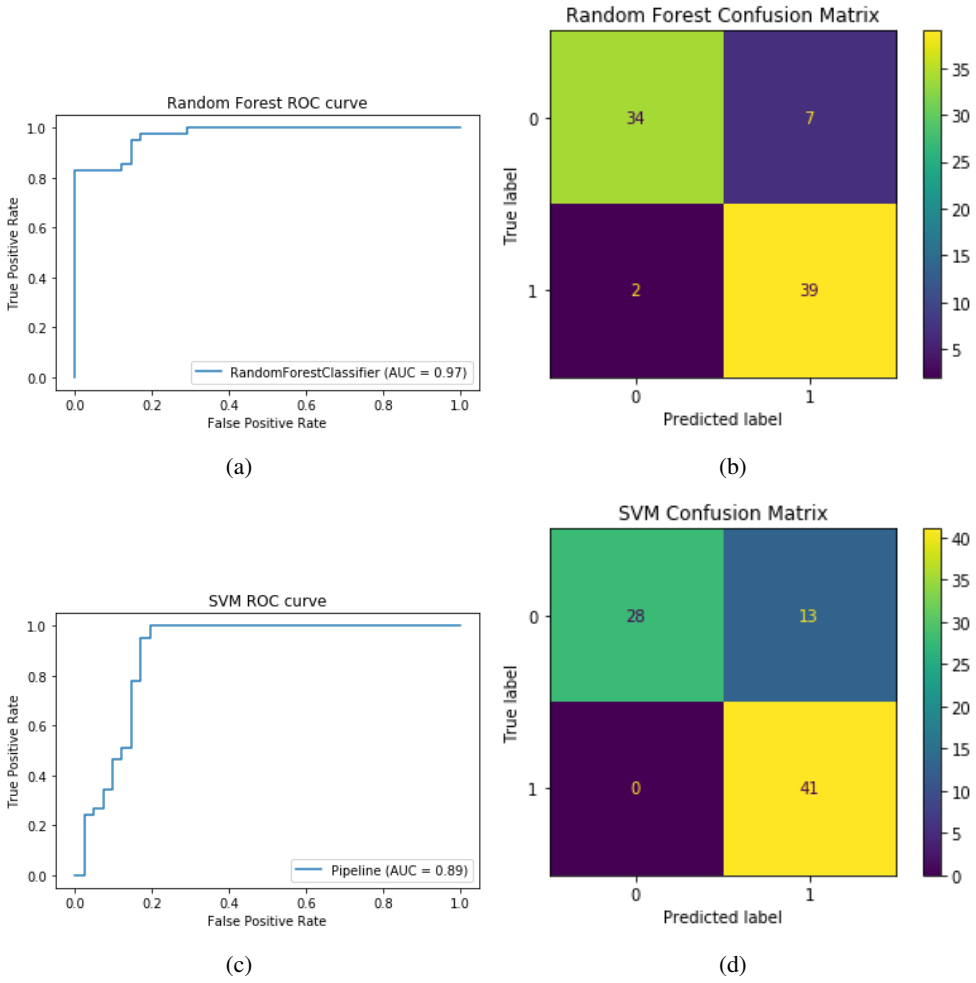


Figure 4.1: (a) AUC and ROC curve for Random Forest classifier model in [3]. (b) Confusion matrix for the same RF classifier model. (c) AUC and ROC curve for Support Vector Machine classifier model. (d) Confusion matrix for the same SVM classifier model. The dataset was oversampled by SMOTE technique and then split using hold-out 0.6-0.4 technique.

However, the use of SMOTE technique to oversampling the classes, hence, modifying the dataset, could be risky due to the possibility of mixing 'synthesised copies' from the same sample in both training and testing groups. In the original paper, no comment has been made on how this technique was applied. The only results that coincide with the published ones are those obtained by the previous methodology. If synthesised predictors from a sample is used to train the classification model and later other synthesised predictors from the same sample is used to test the same model, the results might be great due to significantly similarities between oversampled dataset and likely their results could not be generalised, then it might be a case of possible overfitting.

The use of SMOTE were validated using different approaches. For example, other 5-minutes segments were obtained from the 'high-risk' group of patients with similar context and characteristics as the used to train the previous classification models. This models were tested using this new dataset. Because this test group was only a set of high-risk class, it only could validate the sensitivity of the models, but it is a quick method to test them. The results are shown in the following **table 4.2**.

Also, other example to validate if SMOTE technique distorted the results was using an approach in which synthesised samples from same patients could not being in both training and testing groups. Firstly, dataset was split in training and testing groups and later the SMOTE technique was applied to balance classes in both groups. Due to the difficult to 'synthesise' by oversampling enough data with such a small number of samples, the number of oversampled samples was lower this time, but ensuring a balance between classes. The testing metrics are in the following **table 4.3**. This technique was also tested using the other set of 47 new samples of 5-minutes segments **table 4.4**. **Figure 4.2** shows the resulting ROC curve and confusion matrix of the test set.

Table 4.2: Sensitivity of classification models when applied to other test set for 5' short-term HRV with SMOTE technique proposed in [3].

Model	Hyperparameters	TP	FP	SEN
RF	NT: 300, NF:5	11	37	23.40%
SVM	G:1.4, k:poly, d:3	34	13	72.34%

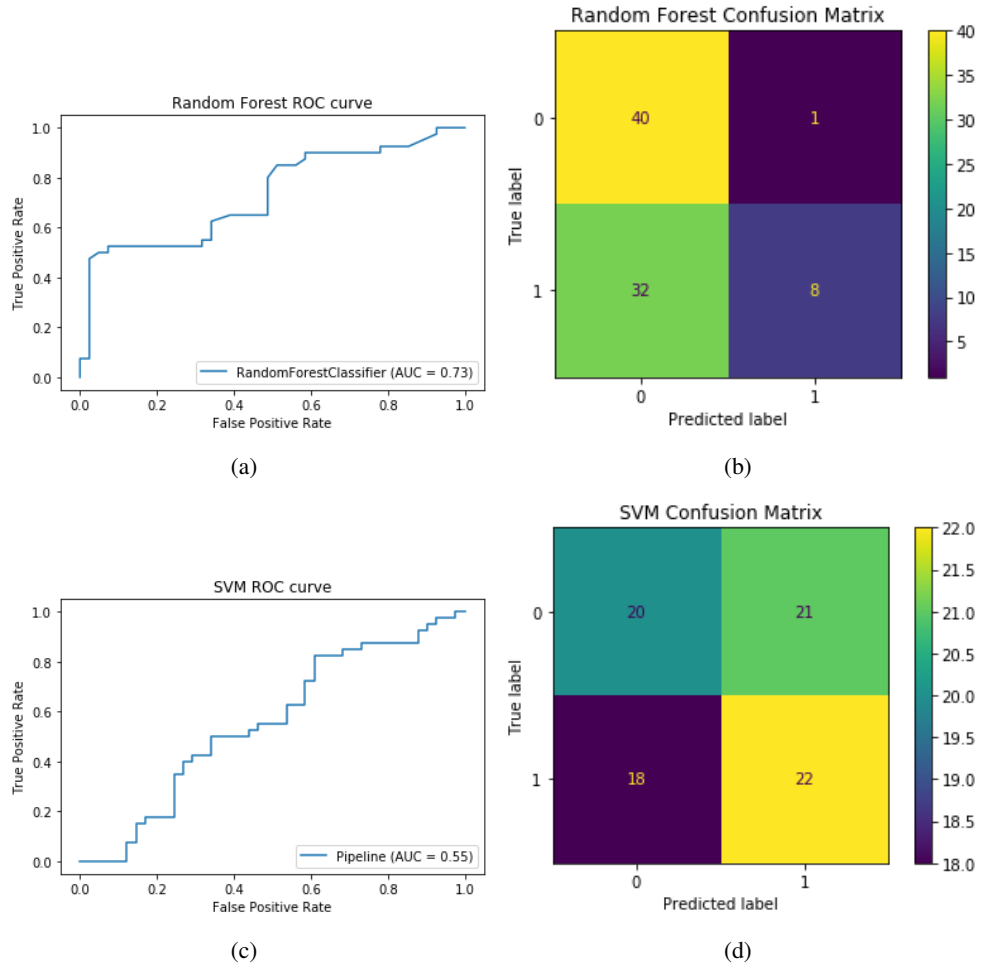
Table 4.3: Performance measurements estimated on the test set (06-04 hold-out estimation) for 5' short-term HRV splitting before SMOTE technique.

Model	Hyperparameters	AUC	ACC	SEN	SPE	F1_score
RF	NT: 300, NF:5	73%	59.26%	20%	97.56%	32.65%
SVM	G:1.4, k:poly, d:3	55%	51.85%	55%	48.78%	53.01%

Table 4.4: Sensitivity of classification models when applied to other test set for 5' short-term HRV with training and testing groups before SMOTE technique.

Class	Hyperparameters	TP	FP	SEN
RF	NT: 300, NF:5	8	39	17.02%
SVM	G = 1.4, Kernel = poly, Degree = 3	30	17	63.83%

A longer dataset with multiples samples from each patient were used to train and test the models. This samples were obtained manually, without using the SMOTE technique. Samples obtained from the same patients were within similar context and with similar characteristics to avoid faking the results due to patient's circumstances. The data is class-balanced, low-risk class is represented by 60% of the set and high-risk class is represented by 40%. **Table 4.5** and **figure 4.3**.



(a) (b)

(c) (d)

Figure 4.2: (a) AUC and ROC curve for Random Forest classifier model in [3]. (b) Confusion matrix for the same RF classifier model. (c) AUC and ROC curve for Support Vector Machine classifier model. (d) Confusion matrix for the same SVM classifier model. The dataset was firstly split using hold-out 0.6-0.4 technique and later oversampled separately training and testing groups.

Table 4.5: Performance measurements estimated on the test set (0.6-0.4 hold-out estimation) for 5' short-term HRV manually balanced without SMOTE technique

Class	Hyperparameters	AUC	ACC	SEN	SPE	F1_score
RF	NT: 300, NF:5	62.29%	66.13%	38.46%	86.11%	48.78%
SVM	G:1.4, k:poly, d:3	51.71%	53.23%	42.31%	61.11%	43.14%

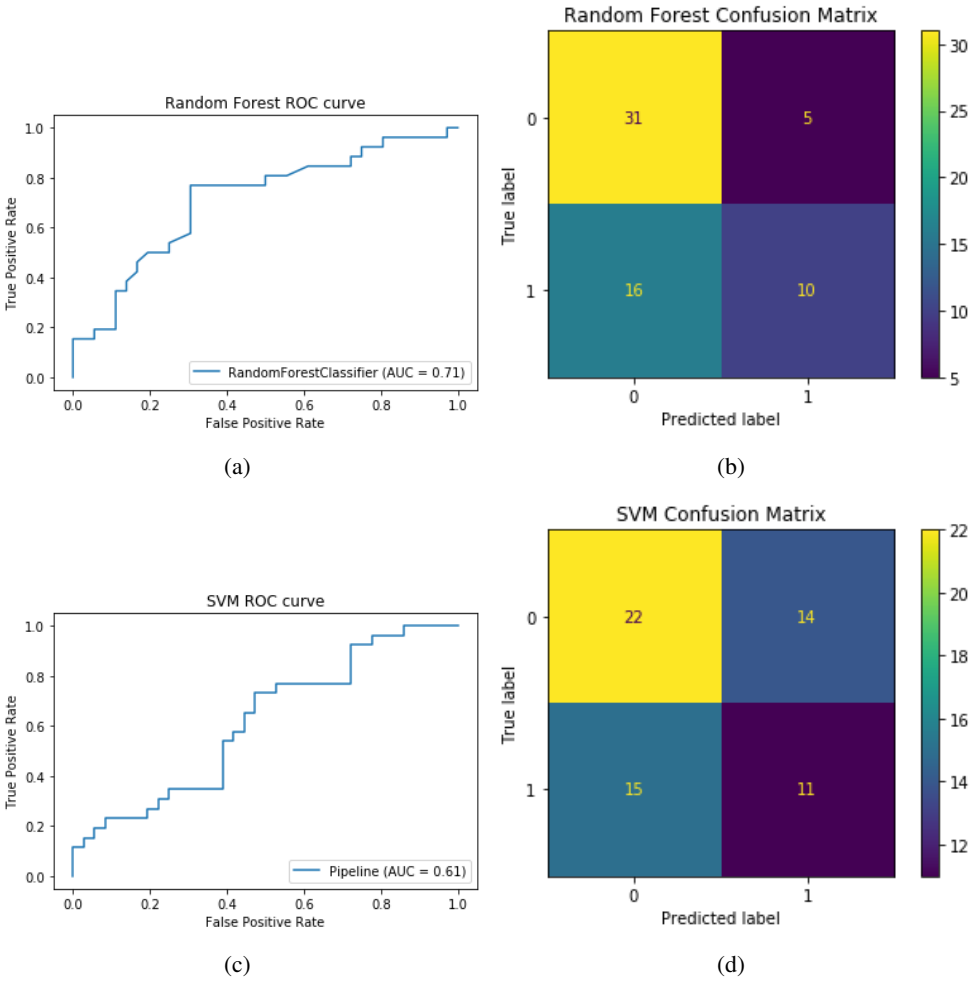
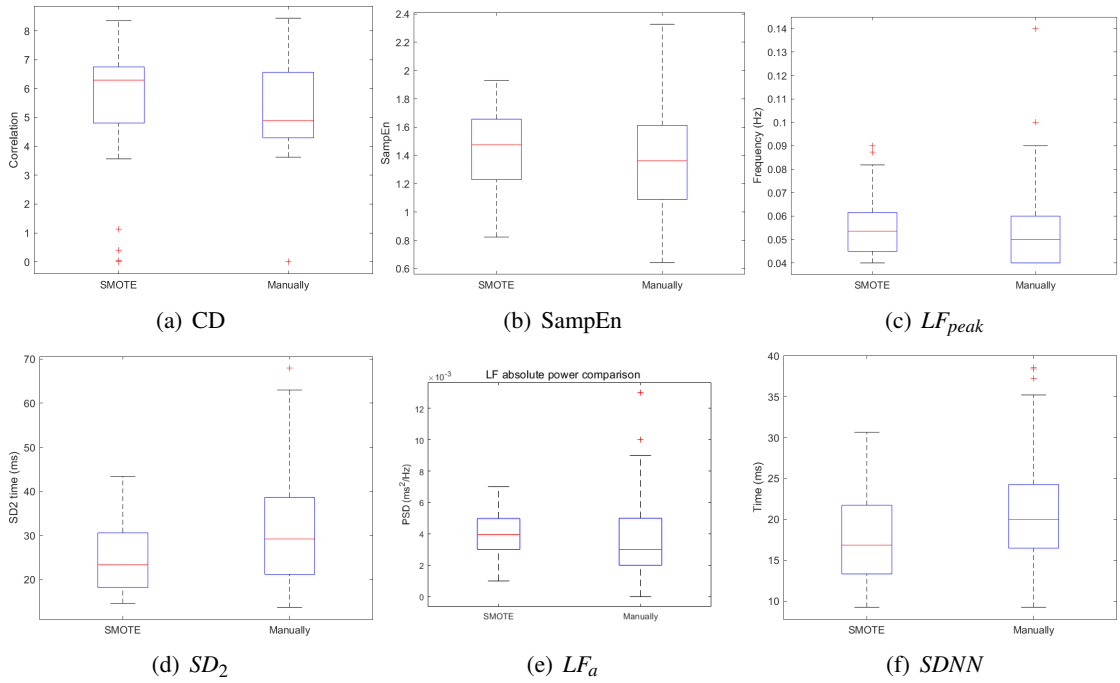


Figure 4.3: (a) AUC and ROC curve for Random Forest classifier model in [3]. (b) Confusion matrix for the same RF classifier model. (c) AUC and ROC curve for Support Vector Machine classifier model. (d) Confusion matrix for the same SVM classifier model. The dataset was split using hold-out 0.6-0.4 technique. Low-risk class was not oversampled, while the high-risk class was oversampled to be 60% of the dataset and high-risk the 40% .

Finally, in **figure 4.4** a whisker box-plot comparison of the top 6 features in [3] is shown below to illustrate the large differences between the performance metrics obtained by using the SMOTE technique (**table 4.1**) and not using it or altering the way it is applied (**other tables**). 'High-risk' class (target) values are represented in the box-plots and 'low-risk' class values were discarded because these values were not oversampled. Also, decision boundaries for the samples plotted using SD2 and SampEn as an example is showed in **figure 4.5**. This figure illustrates the problem of decision boundaries when a oversampling technique is not applied in all the distribution of one problem. There, extra samples which were not oversampled are not represented by the new synthesised data.

The 6 features ranked as the most relevant were Correlation Dimension (*CD*), Sample Entropy (*SampEn*), peak frequency in LF band (LF_{peak}), SD2 of Poincaré Plot (SD_2), absolute power frequency of LF band (LF_a), and the standard deviation of NN intervals (*SDNN*). Not all variable box-plots are shown due to the large number of variables.



(a) CD (b) SampEn (c) LF_{peak}

(d) SD_2 (e) LF_a (f) *SDNN*

Figure 4.4: Comparison of the distribution of values for target class of CD, SampEn, LFpeak, SD2, LFa and SDNN descriptors in oversampled dataset by SMOTE technique and manually oversampled dataset.

The box-plots show a large differences in distribution of most of the top rated features: *SampEn*, LF_{peak},SD_2, LF_a and *SDNN*. Also, correlation dimension box-plot shows a completely different median and the central 50% distribution values (percentile 25-75). This may indicate that the oversampling technique used to carry out the study under-represented the possible true distribution of the different features. Therefore, the trained intelligent models only had to predict the tests in very specific decision regions, which could have increased the positive results.

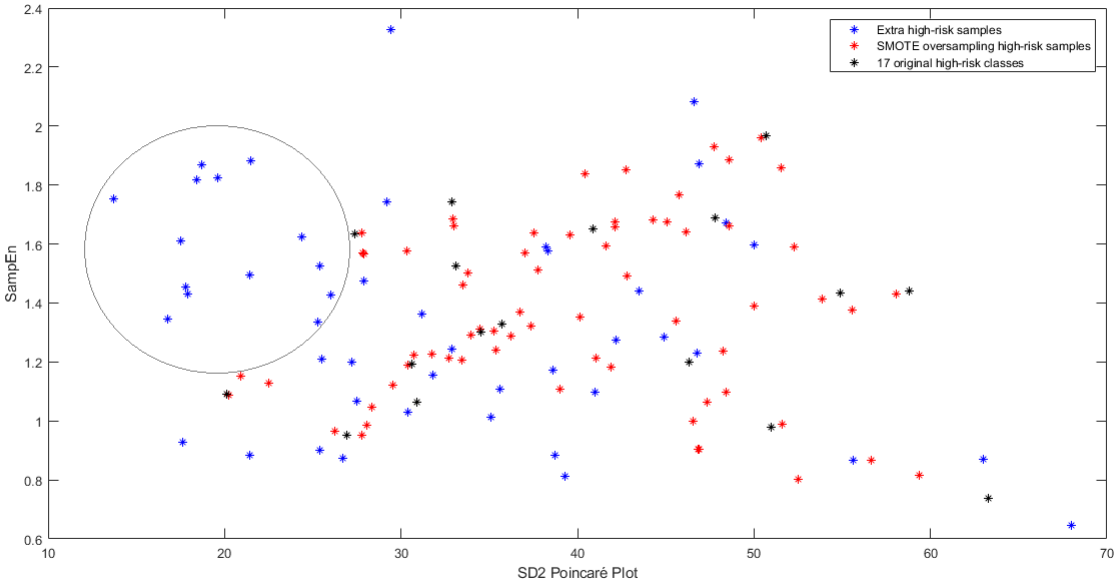


Figure 4.5: Representation of oversampled high-risk classes (samples red), original samples (black) and extra samples obtained from the same patients (blue).

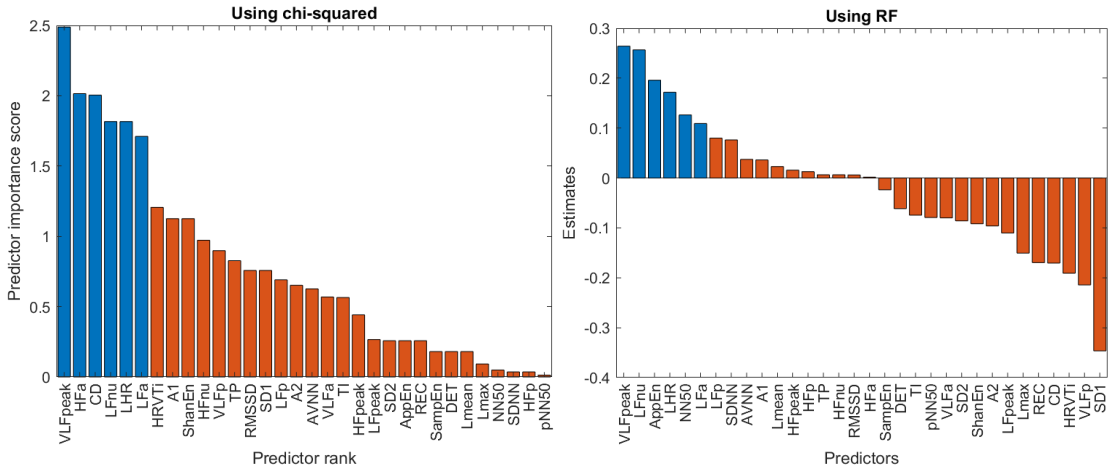
Once it was visualised that the method of oversampling the 'high risk' class could distort the results, different classification models were developed to compare different analysis times. A tiny dataset of 34 samples - 17 samples per class - was used to train and evaluate the models this time. Each random sample was only selected between the time interval from 00 a.m. to 06 a.m to focus on HRV analysis in night sleep. **Tables 4.6**, **4.7** and **4.8** show the predictive results of the 5 classification models used in this methodology, in bold the best model for each time.**Figure 4.3** shows an example of the outcomes of feature selection methods.

Table 4.6: Performance measurements estimated by leave-one-out for 5' HRV analysis

Model	Hyperparameters	Feature selection	AUC	ACC	SEN	SPE	F1_score
RF	NT: 73, NF:'sqrt', mS: 10, mL: 4,	RF-FS(6)	61.8%	61.8%	70.6%	52.9%	64.9%
SVM	G:0.01, k:linear, dFS:'ovr', C:0.1	X2-FS(6)	70.6%	70.6%	88.2%	52.9%	75.0%
NB	**D:'Gaussian'**	**X2-FS(6)**	**70.6%**	**70.6%**	**82.4%**	**58.8%**	**73.7%**
KNN	N:5, alg:'ball_tree', lS:50	X2-FS(6)	52.9%	52.9%	47.1%	58.8%	50.0%
LR	C:1, P: 'l2', s: 'liblinear'	RF-FS(6)	64.7%	64.7%	70.6%	58.8%	66.7%

RF: Random Forest, *SVM*: Support Vector Machine, *NB*: Gaussian Naive Bayes,
KNN: K-Nearest Neighbours, *LR*: Logarithmic regression.

NT: number of trees, *NF*: number of randomly chosen features, *mS*: minimum samples per split,
mL: minimum samples per leaf, *G*: gamma, *k*: Kernel Function,
d: degree of polynomial, *dFS*: decision Fuction Shape, *D*: distribution,
N: n neighbours, *alg*: decision algorithm for nearest neighbours, *lS*: leaf size for 'ball tree' alg,
metric: distance metric, *C*: inverse of regularization strength, *P*: penalty,
s: solver.



(a) X^2 Feature Selection (b) Out-of-bag RF permutation

Figure 4.6: Example of ranking of predictors by feature selection algorithms for 5-minutes dataset. Blue: top 6 ranked predictors. Orange: Not selected predictors.

In accordance with the general metrics - Area Under the Curve (AUC), accuracy (ACC) and F1-Score - results are similar for the 3 time periods. With the exception of KN-Neighbours, classification models trained with only 5 minutes of HRV analysis showed a F1 Score performance of 65% - 75%, with higher sensitivity values than specificity. In this case, target class - 'high-risk' - samples were better classified while 'low-risk' class samples tended to be misclassified more often. Although linear SVM model showed the best value, Gaussian NB was chosen as best performance due to a less due to a smaller imbalance between sensitivity and specificity, and therefore, the best model obtained with this time period. Gaussian

Table 4.7: Performance measurements estimated by leave-one-out for 30' HRV analysis.

Model	Hyperparameters	Feature selection	AUC	ACC	SEN	SPE	F1_score
RF	NT: 400, NF:'sqrt', mS: 5, mL: 4,	X2-FS(6)	64.7%	64.7%	47.1%	82.4%	57.1%
SVM	**G:0.5, k:poly, d:2, dFS:'ovr'**	**X2-FS(6)**	**76.5%**	**76.5%**	**58.8%**	**94.1%**	**71.4%**
NB	D:'Gaussian'	mRMR(4)	67.6%	67.6%	52.9%	82.4%	62.1%
KNN	N:3, alg:'ball_tree', lS:30	RF-FS(6)	61.8%	61.8%	52.9%	70.6%	58.1%
LR	C:1, P: 'elasticinet', s: 'Saga'	RF-FS(6)	70.6%	70.6%	58.8%	82.4%	66.7%

Table 4.8: Performance measurements estimated by leave-one-out for 1 h HRV analysis.

Model	Hyperparameters	Feature selection	AUC	ACC	SEN	SPE	F1_score
RF	NT: 73, NF:3, mS: 5, mL: 2,	RF-FS(6)	67.6%	67.6%	64.7%	70.6%	66.7%
SVM	**G:1.5, k:poly, d:2, dFS:'ovr'**	**RF-FS(6)**	**70.6%**	**70.6%**	**70.6%**	**70.6%**	**70.6%**
NB	D:'Gaussian'	X2-FS(6)	47.1%	47.1%	52.9%	41.2%	50.0%
KNN	N:5	X2-FS(6)	64.7%	64.7%	70.6%	58.8%	66.7%
LR	C:1, P: 'l2', s: ' 'liblinear'	RF-FS(6)	64.7%	64.7%	58.8%	70.6%	62.5%

NB classification model showed slightly good accuracy of 70.6%, a good ability to detect high-risk of 82.4% and a low capacity to discriminate low-risk of 58.8%. F1 score of this model was 73.7% while SVM F1 score was 75%.

For 30 minutes signals, F1 score metrics of classification models were slightly lower than for a 5 minutes period. F1 score showed values between 57% and 71%. The best outstanding model was square polynomial SVM model for this time. Its accuracy was of 76.5%, with a specificity for low-risk samples of 94.1%. However, the sensitivity for the target class was not a high value, only 58.8%. Surprisingly, classification models for this time period performed much better when it came to dismissing 'low-risk' class, but their general ability to detect high-risk samples was poor.

The final period of 1 hour allowed classification models to obtain a F1 score similar as models trained with 30 minutes of analysis. Although, this time their ability for correctly detect both classes was more balanced. The best classification model was a square polynomial SVM model again. Its metrics were incredibly balanced, with a sensitivity and specificity of 70.6%, therefore, their general metrics, AUC, accuracy and F1 score showed 70.6% as well. In general, the average F1 score of all models was similar as 30 minutes models, but with a more balanced predictions, exchanging ability to dismiss low-risk samples for improving detection of high-risk.

Finally, echographic information on the state of the cardiovascular system and hypertension details was added to the classification models. Only when 5 minutes HRV analysis was combined with LVMi - left ventricular mass index -, the classification performance slightly improved in specificity, **table 4.9**.

Table 4.9: Performance of linear SVM model trained with 5' HRV + LVMi.

Model	Hyperparameters	Feature selection	AUC	ACC	SEN	SPE	F1_score
SVM	G:0.01, k:linear, dFS:'ovr', C:0.1	X2-FS(6)	73.5%	73.5%	88.2%	58.8%	76.9%

4.2 Discussion

The main goal of this project was not to obtain incredible high-performance classification models, the characteristics of the dataset did not allow to do so. The purpose of these results was to compare if using only 5 minutes of HRV analysis it was possible to train a classification model with a similar general performance (F1 score) as classification models trained with longer time periods. Previously, in the paper of SHARE project with the University of Naples [3], this hypothesis had been already accepted. However, the characteristics of the dataset and of the methodology used itself raised certain reasonable doubts. The first section of this Master's Dissertation aimed to review this issue to evaluate whether a different approach was necessary or not, due to these circumstances.

First, the dataset was highly unbalanced. It is understandable the difficulty of collecting information for this specific framework: hypertensive patients who sadly suffered a cardiovascular event in the following 12 months. But, the resulting dataset only contained 17 samples of the target class in contrast to the 122 samples of the normal class. This imbalance means a proportion of 86% - 14% between classes. A possible solution is the use of oversampling techniques to reduce this imbalance. However, this approach should be used only in specific frameworks where it is sure that the general distribution of samples of each class can be drawn correctly with the current information. 122 samples could be sufficient to know the general distribution for 'low-risk' class, but 17 samples of 'high-risk' hypertensive people is not for sure enough. This problem could be made worse if the 5-minutes signal extracted from a 24-hours ECG Holter are randomly chosen, without attending the expected variability of HRV values due to activity throughout a patient's day: exercises, sleep...

As **figure 4.4** showed previously, distributions between a artificially oversampled dataset of 'high-risk classes' - 17 real samples and 85 synthesised samples - and a manually oversampled dataset of the same class samples - 102 real samples - were compared. The comparison provides some insights. The manually oversampled dataset has a wider distribution and sometimes different median values. This means the synthesised samples are concentrated around the real samples in the artificially oversampled dataset. Therefore, there is a high chance that the actual real distribution of patients who could be classified as 'high-risk' is under-represented with methodology in the original paper [3].

Besides, as **tables 4.1** to **4.5** showed, when the use of this oversampling technique becomes more and more restricted, the performance metrics plummet. This issue means the oversampling technique is inadequate for this specific problem, because it only highlights the actual under-represented distribution, so classification models trained with this synthesised dataset fail when new samples are added.

A new approach to assess the validity of using only 5 minutes without any oversampling technique was made then. Perhaps only using 34 samples with a leave-one-out approach is not enough to obtain accurate classification models with real performance, but models trained with this approach could be enough to compare the suitability of 5 minutes length comparing to other time periods. Because of the small number of samples, a feature selection of the best predictors was mandatory to avoid overfitting of the classification models.

Although, the average performance of classification models was not incredibly great due to the unbalanced dataset and its small number of samples, sensitivity in classification models trained with 5' short-term HRV analysis showed a high performance to alert of a possibly important risk of developing a cardiovascular event in the incoming year, only a few patients which suffered a cardiovascular event were misclassified. This outcome is quite adequate for this sort of classification problems, where misclassifying sometimes a low-risk patient is preferable to fail in the detection of high-risk patients. Hence, results suggest classification machine learning models could at least be useful to monitor hypertensive people and alert of the possible cardiac risk. Some researches have already assessed the suitability of using short-term HRV analysis in patients to predict different events, such us cardiac sudden death [44]. Moreover, high blood pressure has been already related with cardiovascular diseases [1]. Also, it has

been studied the combination of hypertensive people with anatomic heart problems related with high blood pressure, like left ventricular hypertrophy [45]. **Table 4.9** suggests that clinic information about left ventricular hypertrophy could be combined with HRV analysis to increase prediction, but further analysis are needed.

Finally, feature selection algorithms showed the most important features to predict risk were mostly from frequency domain analysis. Sleep stage has been introduced in several studies as a good conditions to perform HRV analysis, due to an absence of sympathetic activity burst and a more stationary heart rate [46]. Therefore, the variation of spectral power in frequency bands of the inter-beat intervals during sleep could show the adaptation of autonomous nervous system and the competition between sympathetic and parasympathetic branches, which is reduced in hypertensive people [2].

The next steps researching this matter should be collecting a larger number of samples of hypertensive patients in different hospitals, regions and contexts to make sure of representing the real distribution of HRV metrics for these patients.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this Master's dissertation, prediction of risk of develop cardiovascular diseases using short term HRV instead of longer analysis has been corroborated as feasible. First, oversampling synthesis techniques to reduce unbalanced datasets has been proved to not be suitable for this one. Hence, data analysis methodology for the specific characteristics of the SHAREE database collected in University Hospital of Naples Federico II, Naples, was developed without the need of statistical oversampling techniques. Classification models showed a similar performance for the different times, but those trained with short-term HRV analysis showed a higher sensitivity for 5 minutes analysis and a slightly higher F1 score metric. Therefore, this project concludes that it is possible to train machine learning models to predict risk of hypertensive patients using only 5 minutes short-term HRV analysis. Moreover, this project showed the invalidity of applying an oversampling algorithm in the dataset to reduce the imbalance, even if the original research and this study have ended with similar conclusions. Also, results suggest that classification models performance could improve when clinic cardiovascular information is added to HRV analysis, like LVMi from echographic or ultrasound imaging. Finally, feature selection algorithms showed that predictors from frequency domain related with the antagonism between sympathetic and parasympathetic nervous system were the strongest values for this case where HRV analysis was made during night stage. Other non-linear analysis like Correlation Dimension analysis showed helpful for this analysis.

5.2 Limitations and Future work

This work has encountered a number of difficulties that have limited its development, mainly related to the dataset. First of all, the dataset was not sufficiently large and balanced to show a good insight of the feasibility of the developed classification models in other possible datasets. Secondly, some clinic information was missing from several patients. So, not the clinic information could have been used to perform other strategies. Also, the QRS annotation provided by the dataset was erroneous in different sections of the ECG, usually when strong noises was presented. Own QRS extraction methodologies had to been implemented.

In future, the same study should be replicated with a longer database. Samples should be collected in different hospital in different regions to be able to know the real distribution of possible samples from hypertensive patients. With a better dataset, better statistical exploratory analysis and new, more robust

machine learning approach could be done, like different ensembles algorithms using several layers of learning from a combination of different algorithms. In this case, a attempt could be done to train real classification models to predict hypertension and to be tested in a real environment such as hospitals.

Chapter 6

Glossary of Terms

ABP: Ambulatory blood pressure.
AF: Atrial Fibrillation.
APC: Atrial premature contraction.
AUC: Area under the curve.
AV: Atrioventricular.
CNN: Convolutional neural network.
CVD: Cardiovascular disease.
DALYs: Disability-adjusted life year.
ECG: Electrocardiogram.
ESC: European Society of Cardiology.
ESH: European Society of Hypertension.
FN: False Negative.
FP: False Positive.
G: For Support Vector Machine, kernel coefficient 'gamma'.
HBP: High blood pressure.
HRV: Hear Rate Variability.
k: For Support Vector Machine, Kernel Function.
ML: Machine Learning.
mL: For random forest, minimum samples per leaf.
mS For random forest, minimum samples per split.
NF: For random forest, number of features in the subset to train each forest.
NN: Interbeat interval from the SA - after preprocessing (normal-to-normal).
NT: For random forest, number of trees in each forest.
PPG: Photoplethysmogram.
PSD: Power spectral density.
RF: Random Forest (Machine Learning model).
ROC: Receiver Operating Characteristic.
RR: Interbeat interval before preprocessing.
SA: Sinoatrial.
Se: Sensitivity.
SHAREE: Smart Health for Assessing the Risk of Events via ECG project.
SVM: Support Vector Machine (Machine Learning model).
SW: Search Window.

TN: True Negative.
TP: True Positive.
VPC Ventricular premature contraction.
WHO: World Health Organization.
WT: Wavelet transform.

List of Figures

List of Tables

Bibliography

[1] Ramachandran S Vasan, Martin G Larson, Eric P Leip, Jane C Evans, Christopher J O'Donnell, William B Kannel, and Daniel Levy. Impact of high-normal blood pressure on the risk of cardiovascular disease. *New England journal of medicine*, 345(18):1291–1297, 2001.

[2] Jacqueline M Dekker, Richard S Crow, Aaron R Folsom, Peter J Hannan, Duanping Liao, Cees A Swenne, and Evert G Schouten. Low heart rate variability in a 2-minute rhythm strip predicts risk of coronary heart disease and mortality from several causes: the aric study. *Circulation*, 102(11):1239–1244, 2000.

[3] Paolo Melillo, Raffaele Izzo, Ada Orrico, Paolo Scala, Marcella Attanasio, Marco Mirra, Nicola De Luca, and Leandro Pecchia. Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. *PloS one*, 10(3):e0118504, 2015.

[4] World Health Organization. Leading causes of death and disability 2000-2019: A visual summary. Technical report, WHO, 2019.

[5] Alberto Cordero, Vicente Bertomeu-Martínez, Pilar Mazón, Lorenzo Fácila, Vicente Bertomeu-González, Juan Cosín, Enrique Galve, Julio Núñez, Iñaki Lekuona, and José R González-Juanatey. Factores asociados a la falta de control de la hipertensión arterial en pacientes con y sin enfermedad cardiovascular. *Revista Española de Cardiología*, 64(7):587–593, 2011.

[6] Vicente Bertomeu, Alberto Cordero, Juan Quiles, Pilar Mazón, Joaquín Aznar, and Héctor Bueno. Control de los factores de riesgo y tratamiento de los pacientes con cardiopatía isquémica: registro trece. *Revista española de cardiología*, 62(7):807–811, 2009.

[7] Hilary K Wall, Judy A Hannan, and Janet S Wright. Patients with undiagnosed hypertension: hiding in plain sight. *Jama*, 312(19):1973–1974, 2014.

[8] Santos Casado Pérez. Hipertensión arterial. In Fundación BBVA, editor, *Libro de la salud cardiovascular del Hospital Clínico San Carlos y la Fundación bbva / dirigido por Antonio López Farré y Carlos Macaya Miguel*, chapter Capítulo 1, pages 121–130. Bilbao, 1.ª edition, 2009.

[9] Jagmeet P Singh, Martin G Larson, Hisako Tsuji, Jane C Evans, Christopher J O'Donnell, and Daniel Levy. Reduced heart rate variability and new-onset hypertension: insights into pathogenesis of hypertension: the framingham heart study. *Hypertension*, 32(2):293–297, 1998.

[10] Sangthong Terathongkum and Rita H Pickler. Relationships among heart rate variability, hypertension, and relaxation techniques. *Journal of Vascular Nursing*, 22(3):78–82, 2004.

[11] Hongbo Ni, Sunyoung Cho, Jennifer Mankoff, Jun Yang, et al. Automated recognition of hypertension through overnight continuous hrv monitoring. *Journal of Ambient Intelligence and Humanized Computing*, 9(6):2011–2023, 2018.

[12] Michael O'Rourke. Mechanical principles in arterial disease. *Hypertension*, 26(1):2–9, 1995.

[13] Bryan Williams, Giuseppe Mancia, Wilko Spiering, and I Pörsti. 2018 esc/esh guidelines for the management of arterial hypertension. *European heart journal*, 39(33):3021–3104, 2018.

[14] Paul K Whelton, Robert M Carey, Wilbert S Aronow, Donald E Casey, Karen J Collins, Cheryl Dennison Himmelfarb, Sondra M DePalma, Samuel Gidding, Kenneth A Jamerson, Daniel W Jones, et al. 2017 acc/aha/aapa/abc/acpm/ags/apha/ash/aspc/nma/pcna guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Journal of the American College of Cardiology*, 71(19):e127–e248, 2018.

[15] John E. Hall. *Guyton and Hall Textbook of Medical Physiology*. W B Saunders, London, England, 13 edition, 2015.

[16] N. Lippman, K. M. Stein, and B. B. Lerman. Comparison of methods for removal of ectopy in measurement of heart rate variability. *American Journal of Physiology - Heart and Circulatory Physiology*, 267(1 36-1), 1994.

[17] Dib Nabil and F. Bereksi Reguig. Ectopic beats detection and correction methods: A review. *Biomedical Signal Processing and Control*, 18:228–244, apr 2015.

[18] Sebastiano Massaro and Leandro Pecchia. Heart Rate Variability (HRV) Analysis: A Methodology for Organizational Neuroscience. *Organizational Research Methods*, 22(1):354–393, jan 2019.

[19] Ahmad Sajadieh, Verner Rasmussen, Hans Ole Hein, and Jørgen Fischer Hansen. Familial predisposition to premature heart attack and reduced heart rate variability. *American Journal of Cardiology*, 92(2):234–236, 2003.

[20] Leif Sörnmo and Pablo Laguna. *Bioelectrical signal processing in cardiac and neurological applications*, volume 8. Academic Press, 2005.

[21] MATLAB. Version 9.10.0 (r2021a), 2021. Copyright 1994-2021 The MathWorks, Inc.

[22] Python Software. Version 3.9.6, 2021. Copyright © 2001-2021 Python Software Foundation. All rights reserved.

[23] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[24] W Zong, GB Moody, and D Jiang. A robust open-source algorithm to detect onset and duration of qrs complexes. In *Computers in Cardiology, 2003*, pages 737–740. IEEE, 2003.

[25] Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065, 1996.

[26] Hooman Sedghamiz. Matlab implementation of pan tompkins ecg qrs detector. *Code Available at the File Exchange Site of MathWorks*, 2014.

[27] Adriana N Vest, Giulia Da Poian, Qiao Li, Chengyu Liu, Shamim Nemati, Amit J Shah, and Gari D Clifford. An open source benchmarked toolbox for cardiovascular waveform and interval analysis. *Physiological measurement*, 39(10):105004, 2018.

[28] Qichen Li, Chengyu Liu, Qiao Li, Supreeth P Shashikumar, Shamim Nemati, Zichao Shen, and Gari D Clifford. Ventricular ectopic beat detection using a wavelet transform and a convolutional neural network. *Physiological measurement*, 40(5):055002, 2019.

[29] Peter Walter Kamen, Henry Krum, and Andrew Maxwell Tonkin. Poincare plot of heart rate variability allows quantitative display of parasympathetic nervous activity in humans. *Clinical science*, 91(2):201–208, 1996.

[30] Alfonso Delgado-Bonal and Alexander Marshak. Approximate entropy and sample entropy: A comprehensive tutorial. *Entropy*, 21(6):541, 2019.

[31] Raul Carvajal, Niels Wessel, Montserrat Vallverdú, Pere Caminal, and Andreas Voss. Correlation dimension analysis of heart rate variability in patients with dilated cardiomyopathy. *Computer Methods and Programs in Biomedicine*, 78(2):133–140, 2005.

[32] Thomas Penzel, Jan W Kantelhardt, Ludger Grote, Jörg-Hermann Peter, and Armin Bunde. Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea. *IEEE Transactions on biomedical engineering*, 50(10):1143–1151, 2003.

[33] Hubert Dabiré, Denis Mestivier, Jacqueline Jarnet, Michel E Safar, and Nguyen Phong Chau. Quantification of sympathetic and parasympathetic tones by nonlinear indexes in normotensive rats. *American Journal of Physiology-Heart and Circulatory Physiology*, 275(4):H1290–H1297, 1998.

[34] Yun Chen and Hui Yang. Multiscale recurrence analysis of long-term nonlinear and nonstationary time series. *Chaos, Solitons & Fractals*, 45(7):978–987, 2012.

[35] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[36] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[37] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[38] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

[39] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[40] Huan Liu and Rudy Setiono. Chi2: Feature selection and discretization of numeric attributes. In *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*, pages 388–391. IEEE, 1995.

[41] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[42] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[43] Emilio Vanoli, Philip B Adamson, Ba-Lin, Gian D Pinna, Ralph Lazzara, and William C Orr. Heart rate variability during specific sleep stages: a comparison of healthy subjects with patients after myocardial infarction. *Circulation*, 91(7):1918–1922, 1995.

[44] Elias Ebrahimzadeh, Mohammad Pooyan, and Ahmad Bijar. A novel approach to predict sudden cardiac death (scd) using nonlinear and time-frequency analyses from hrv signals. *PloS one*, 9(2):e81896, 2014.

[45] Peter M Okin, Sverre E Kjeldsen, Stevo Julius, Darcy A Hille, Björn Dahlöf, Jonathan M Edelman, and Richard B Devereux. All-cause and cardiovascular mortality in relation to changing heart rate during treatment of hypertensive patients with electrocardiographic left ventricular hypertrophy. *European heart journal*, 31(18):2271–2279, 2010.

[46] Gabrielle Brandenberger, Martin Buchheit, Jean Ehrhart, Chantal Simon, and François Piquard. Is slow wave sleep an appropriate recording condition for heart rate variability analysis? *Autonomic Neuroscience*, 121(1-2):81–86, 2005.

Budget

The following sections present the complete economic estimation for this Master's Dissertation. In last table, costs associated with this Master's Dissertation are summarized and presented, with taxes included. For this project, the work of a biomedical engineer, whose total workload amounts to 311 hours and whose cost has been assumed to be a standard salary of €1200 in 14 payments per year, has been considered sufficient. In addition, the worker's salary has certain extra costs associated with it, such as social security. According to Spanish Ministry of Employment and Social Security, it is: 23.6 % for common contingencies, 5.5% for unemployment insurance, 0.6% for vocational training, 0.2% for FOGASA and 1.65% for IT/IMS (accidents at work and occupational diseases). The hourly wage of the worker has been calculated on the basis of a working day of 8 hours and subtracting the rest days per year it can be calculated as 1,792 working hours per year, so the hourly wage can be rounded up to 12 €/h.

The cost of materials included all the software programmes used to develop the project, although some were free of charge as they were freely available or provided by the university. All these programmes have been included in a separate "installation" section to avoid making the budgets much more complex, although each programme is used in other sections.

Finally, an increase of 13% in the material execution budget has been considered for overheads and 6% in industrial profit. In addition, the corresponding VAT of 21% at the general rate has been applied in this section, as it was not included in the previous sections.

In the following tables, the project budget is presented in detail, broken down into categories of expenditures and costs per unit

1. Table of personal costs

Personal costs

N.	Code	Description of staff	Cost per unit	Time	Total
1	MO.IB	Biomedical Engineer	12.000	424.000 h	5,088.00
				Total for personal costs:	5,088.00

2. Table of materials

Material costs

N.	Code	Description of material	Cost per unit	Amount	Amort. Fact.	Total
1	MAT.MO365	Microsoft Office 365 Business License	126.000	1.000 u	7/12	73.5
2	MAT.MAT	MATLAB R2021a Student License	69.000	1.000 u	7/12	40.25
3	MAT.PYT	Python Software	0	1.000 u	-	0.00
4	MAT.WIF	Internet Provider	420.000	1.000 u	1/12	35.00
5	MAT.PC	Laptop Lenovo 510-1 I SK 80SR	540.000	1.000 u	7/60	63.00
					Total for material costs:	211.75

3. Table of partial budgets

1. Project preparation and review of state of art

N.	Unit	Description		Amount	Cost per unit	Cost
1.1	**H**	Review of scientific literature				
			Total h :	24.000	12.36	**296.64**
1.2	**H**	Review of dataset and methodology proposed for SHAREE project				
			Total h :	6.000	12.36	**74.16**
1.3	**H**	Review of different techniques and software packages for signal pre-processing				
			Total h :	14.000	12.36	**173.04**
1.4	**U**	Installation of the required software				
			Total u :	1.000	211.75	**211.75**
1.5	**H**	Preparation of standard classification machine learning models packages				
			Total h :	3.000	12.36	**37.08**
1.6	**H**	Downloading and preparation of SHAREE database from Physionet				
			Total h :	12.000	12.36	**148.32**
		Total partial budget nº 1 Project preparation and review of state of art :				**940.99**

2. IBI series extraction and pre-processing

N.	Unit	Description		Amount	Cost per unit	Cost
2.1	**H**	Assessing best methodology for HRV preprocessing				
			Total h :	35.000	12.36	**197.76**
2.2	**H**	Preparation of script for beat indexes extraction				
			Total h :	12.000	12.36	**148.32**
2.3	**H**	Final coding of preprocessing script				
			Total h :	20.000	12.36	**247.20**
2.4	**U**	Extraction and preprocessing of time signals				
			Total h :	76.000	12.36	**939.36**
2.5	**H**	----------				
			Total u :	1.000	194.41	**194.41**
		Total partial budget nº 2 Signal HRV extraction and pre-processing :				**1,532.64**

3. HRV analysis

N.	Unit	Description		Amount	Cost per unit	Cost
3.1	**H**	Evaluation of packages for HRV analysis				
			Total h :	16.000	12.36	**197.76**
3.2	**H**	Preparation of HRV analysis functions				
			Total h :	6.000	12.36	**74.16**
3.3	**H**	Preparation of scripts for datasets				
			Total h :	4.000	12.36	**49.44**
3.4	**H**	Creation of 5', 30' and 1 h. datasets				
			Total h :	6.000	12.36	**74.16**
3.5	**H**	SMOTE oversampling processing in datasets				
			Total h :	4.000	12.36	**49.44**
3.6	**H**	Exploratory data analysis of datasets				
			Total h :	14.000	12.36	**173.04**
			Total partial budget nº 3 HRV analysis :			**618.00**

4. Machine learning classification

N.	Unit	Description		Amount	Cost per unit	Cost
4.1	**H**	Preparation of hyperparameter tuning in Sk-Learn				
			Total h :	12.000	12.36	**148.32**
4.2	**H**	Training and testing ML models				
			Total h :	32.000	12.36	**394.56**
4.3	**H**	Evaluation and metrics				
			Total h :	12.000	12.36	**148.32**
4.4	**U**	Preparation of tables and results				
			Total h :	6.000	12.36	**74.16**
		Total partial budget nº 4 Machine learning classification :				**765.36**

5. Writing and defence of Master's Dissertation

N.	Unit	Description		Amount	Cost per unit	Cost
5.1	**H**	Drafting of project documents				
			Total h :	90.000	12.36	**1,112.40**
5.2	**H**	Preparation of the defence exposition				
			Total h :	20.000	12.36	**247.20**
		Total partial budget nº 5 Writing and defence of Master's Dissertation :				**1,359.60**

4. Table of unit prices

N.	Description	Cost	
		Figures format (Euros)	Word format (Euros)
	1 PROJECT PREPARATION AND REVIEW OF STATE OF ART		
1.1	h Review of scientific literature	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
1.2	h Review of dataset and methodology proposed for SHAREE project	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
1.3	h Review of different techniques and software packages for signal pre-processing	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
1.4	u Installation of the required software	218.10 €	TWO HUNDRED AND ELEVEN EUROS AND SEVENTY-FIVE CENTS
1.5	h Preparation of standard classification machine learning models packages	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
1.6	h Downloading and preparation of SHAREE database from Physionet	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
	2 IBI SERIES EXTRACTION AND PRE-PROCESSING		
2.1	h Assessing best methodology for HRV preprocessing	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
2.2	h Preparation of script for beat indexes extraction	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
2.3	h Final coding of preprocessing script	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
2.4	h Extraction and preprocessing of time signals	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
	3 HRV ANALYSIS		
3.1	h Evaluation of packages for HRV analysis	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
3.2	h Preparation of HRV analysis functions	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
3.3	h Preparation of scripts for datasets	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
3.4	h Creation of 5', 30' and 1 h. datasets	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
3.5	h SMOTE oversampling processing in datasets	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
3.6	h Exploratory data analysis of datasets	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
	4 MACHINE LEARNING CLASSIFICATION		
4.1	h Preparation of hyperparameter tuning in Sk-Learn	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
4.2	h Training and testing ML models	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
4.3	h Evaluation and metrics	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
4.4	h Preparation of tables and results	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS
	5 WRITING AND DEFENCE OF MASTER'S DISSERTATION		
5.1	h Drafting of project documents	12.36 €	TWELVE EUROS AND THIRTY-SIX CENTS

5.2	h Preparation of the defence exposition	**12.36 €**	**TWELVE EUROS AND THIRTY-SIX CENTS**

5. Table of disaggregated prices

1 Project preparation and review of state of art

N.	Unit	Description				Costs
1.1	**h**	Review of scientific literature				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
1.2	**h**	Review of dataset and methodology proposed for SHAREE project				
	1.000 h	Biomedical Engineer			12.000 €	12,00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
1.3	**h**	Review of different techniques and software packages for signal pre-processing				
	1.000 h	Biomedical Engineer			12.000 €	12,00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
1.4	**u**	Installation of the required software				
	1.000 u	Microsoft Office 365 Business License			73.500 €	73.50 €
	1.000 U	MATLAB R2021a Student License			40.250 €	40.25 €
	1.000 U	Internet Provider			35.000 €	35.00 €
	1.000 U	Laptop Lenovo 510-1 I SK 80SR			63.000 €	63.00 €
			3.000 %	Indirect expenses	211.750 €	**6.35 €**
				Total cost per unit		**218.10 €**
1.5	**h**	Preparation of standard classification machine learning models packages				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**218.10 €**
1.6	**h**	Downloading and preparation of SHAREE database from Physionet				
	1.000 h	Biomedical Engineer			12,000 €	12,00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**

2 IBI series extraction and pre-processing

N.	Unit	Description				Costs
2.1	h	Assessing best methodology for HRV preprocessing				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
2.2	h	Preparation of script for beat indexes extraction				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
2.3	h	Final coding of preprocessing script				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
2.4	h	Extraction and preprocessing of time signals				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**

3 HRV analysis

N.	Unit	Description				Costs
3.1	h	Evaluation of packages for HRV analysis				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
3.2	h	Preparation of HRV analysis functions				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
3.3	h	Preparation of scripts for datasets				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
3.4	u	Creation of 5', 30' and 1 h. datasets				
	1.000 u	Licencia Microsoft Office 365 para empresa			12.000 €	12.00 €

			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per unit		**12.36 €**
3.5	**h**	SMOTE oversampling processing in datasets				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
3.6	**h**	Exploratory data analysis of datasets				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**

4 Machine learning classification

N.	Unit	Description				Costs
4.1	**h**	Preparation of hyperparameter tuning in Sk-Learn				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
4.2	**h**	Training and testing ML models				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
4.3	**h**	Evaluation and metrics				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
4.5	**h**	Preparation of tables and results				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
4.6	**h**	Downloading and preparation of SHAREE database from Physionet				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**

5 Writing and defence of Master's Dissertation

N.	Unit	Description				Costs
5.1	h	Drafting of project documents				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**
5.2	h	Preparation of the defence exposition				
	1.000 h	Biomedical Engineer			12.000 €	12.00 €
			3.000 %	Indirect expenses	12.000 €	**0.36 €**
				Total cost per hour		**12.36 €**

6. Budget for execution under contract

Chapter	Costs (€)
1 Project preparation and review of state of art	**940.99**
2 IBI series extraction and pre-processing	**1,532.64**
3 HRV analysis	**618.00**
4 Machine learning classification	**765.36**
5 Writing and defence of Master's Dissertation	**1,359.60**
Budget for material execution (BME)	**5,216.59**
13% overhead costs	678.16
6% industrial profit	313.00
Budget for execution under contract (BEC = BME + OC + IP)	**6207.75**
21% IVA	1303.63
Budget for execution under contract + IVA (BEC = BME + OC + IP + IVA)	**7511.38**

The total budgeted cost for the contract execution plus IVA tax is estimated at the amount of SEVEN THOUSAND FIVE HUNDRED AND ELEVEN WITH THIRTY-EIGHT CENTS.