UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola d'Enginyeria de Barcelona Est

FINAL DEGREE THESIS

Bachelor's Degree in Biomedical Engineering

PROGNOSIS OF SYMPTOMATIC PATIENTS WITH BRUGADA SYNDROME THROUGH ELECTROCARDIOGRAM BIOMARKERS AND MACHINE LEARNING

Report and Annex

Author:	Luis Tortosa Valero
Director:	Pedro Gomis Roman
Department	ESAII
Co-directors:	Flavio Palmieri
	Elena Arbelo, Hospital Clínic, Universitat de Barcelona
Call:	2023, Januray

Abstract

The Brugada Syndrome (BrS) is a rare but serious cardiovascular disorder that can cause dangerously fast heartbeats and is characterized by a particular set of electrocardiogram (ECG) patterns. It's a very unpredictable condition. Many people don't present symptoms at all, while for others, unfortunately, the first symptom is death.

For high risk patients, having an implantable cardioverter-defibrillator placed is recommended. Unfortunately, that has severe risks associated, like infections and inappropriate shocks, so it's key to identify those high risk patients.

The objective of this project was to develop machine learning based tools that are able to tell symptomatic Brugada Syndrome patients apart from those who are not. Symptomatic patients were considered those who had recovered from cardiac death, suffered an arrhythmogenic syncope or sustained tachycardia.

In order to do so, after an investigation of the state of the art of the relevant subjects, several biomarkers related with Brugada ECG patterns were extracted from 24h ECG recordings of 45 different patients, after having been processed by signal averaging in order to reduce their noise. Those biomarkers, alongside some clinical data, were then separated in different ways in order to train and test different machine learning based automated classifier models.

The performances of those models were very poor. None of them was able to reliably classify BrS patients as desired. Nevertheless, valuable conclusions can be extracted from this first approach to pursue the intended goal further, and useful tools were developed that would allow for a faster processing of the database used.

Resum

La Síndrome de Brugada (BrS) és un trastorn cardiovascular poc comú però greu que pot causar batecs perillosament ràpids i es caracteritza per presentar un conjunt particular de patrons d'electrocardiograma (ECG) als seus pacients. És una condició molt impredictible. Moltes persones no presenten cap símptoma, mentre que per altres, lamentablement, el primer símptoma és la mort.

Per a pacients d'alt risc es recomana col·locar un desfibril·lador cardioversor implantable. Desafortunadament, això té greus riscos associats, com infeccions i descàrregues inadequades, per la qual cosa és clau identificar aquests pacients d'alt risc correctament.

L'objectiu d'aquest projecte ha estat desenvolupar eines basades en aprenentatge automàtic que poguessin diferenciar els pacients amb Síndrome de Brugada simptomàtics dels quals no ho són. Es van considerar pacients simptomàtics aquells que s'havien recuperat de mort cardíaca, van patir un síncope arritmogènic o taquicàrdia sostinguda.

Per fer-ho, després d'una investigació de l'estat de l'art dels temes pertinents, es van extreure diversos biomarcadors relacionats amb els patrons d'ECG de Brugada a partir de registres d'ECG de 24 hores de 45 pacients diferents, després d'haver estat processats per mitjà de promediat de senyal per reduir el soroll.

Aquests biomarcadors, juntament amb algunes dades clíniques, es van separar de diferents maneres per entrenar i provar diferents models de classificació automatitzats basats en aprenentatge automàtic.

Els resultats obtinguts dels models han estat molt pobres. Cap d'ells no ha pogut classificar de manera fiable els pacients amb BrS com es desitjava. Això no obstant, d'aquesta primera aproximació es poden extreure conclusions valuoses per assolir l'objectiu del projecte, i s'han desenvolupat eines útils que poden permetre un processament més ràpid de la base de dades utilitzada.

Resumen

El Síndrome de Brugada (BrS) es un trastorno cardiovascular poco común pero grave que puede causar latidos peligrosamente rápidos y se caracteriza por presentar un conjunto particular de patrones de electrocardiograma (ECG) en sus pacientes. Es una condición muy impredecible. Muchas personas no presentan ningún síntoma, mientras que para otras, lamentablemente, el primer síntoma es la muerte.

Para pacientes de alto riesgo se recomienda la colocación de un desfibrilador cardioversor implantable. Desafortunadamente, eso tiene graves riesgos asociados, como infecciones y descargas inapropiadas, por lo que es clave identificar a esos pacientes de alto riesgo correctamente.

El objetivo de este proyecto era desarrollar herramientas basadas en aprendizaje automático que puedan diferenciar a los pacientes con Síndrome de Brugada sintomáticos de aquellos que no lo son.

Se consideraron pacientes sintomáticos aquellos que se habían recuperado de muerte cardiaca, sufrieron un síncope arritmogénico o una taquicardia sostenida. Para ello, tras una investigación del estado del arte de los temas relevantes, se extrajeron varios biomarcadores relacionados con los patrones de ECG de Brugada a partir de registros de ECG de 24h de 45 pacientes diferentes, después de haber sido procesados mediante promedio de señal para reducir su ruido.

Estos biomarcadores, junto con algunos datos clínicos, se separaron de diferentes maneras para entrenar y probar diferentes modelos de clasificación automatizados basados en aprendizaje automático.

Los resultados de los modelos obtenidos han sido muy pobres. Ninguno de ellos pudo clasificar de manera confiable a los pacientes con BrS como se deseaba. No obstante, de esta primera aproximación se pueden extraer valiosas conclusiones para continuar avanzando hacia el objetivo perseguido, y se desarrollaron herramientas útiles que permitirán un procesamiento más rápido de la base de datos utilizada.

Appreciations

I can't begin this section but by thanking Pedro, the thesis director, first of all for offering me the chance of doing this project in the first place. But most importantly, for how much he has helped me overall, meeting regularly and offering many resources without needing to ask for them, understanding my peculiar work rhythms, accepting my particular interests during the course of the project, and staying by my side during an intense final stage of the project.

I also want to, of course, thank Flavio and Elena, the co-directors, who were able to bring their particular perspectives and expertise into the project, made me feel sure they would answer any doubt they could concerning the broad amount of topics covered here and, well, in the case of Elena also allowed me to do this project itself by being at the front of the serious research behind this humble final degree thesis.

I also want to thank the amazing, often unnoticed and even persecuted work of everyone who tries to make knowledge free and openly available for everyone. People putting effort to make websites like Sci-Hub, endangered or already unavailable ones like Z-library, ones that have been particularly helpful in this project like AnatomyTool or Internet Archive, and even Wikipedia; be up, available for everyone and full of great articles, images and more. Without them, progress would be hindered even more by capitalist dynamics, which make knowledge a commodity whose main purpose is increasing corporate benefits, severely restricting access to it. This project could've definitely not have existed as it is. Their work helps to make people's lives around the world better and should be appreciated as such.

And of course I want to extend these thanks to all of the people who love and support me, friends, family and more. If any of you are reading this, thanks for patiently listening to my fears and anguishes surrounding this project. That's over now, I hope.

As is often said, this document would not be in front of you without your support.

Glossary

ECG: Electrocardiogram

BrS: Brugada Syndrome

ICD: Implantable Cardioverter Defibrillator

ML: Machine Learning

McW: Misclassification Weights

TP: True Positive

TPR: True Positive Rate

ATPR: Asymptomatic True Positive Rate

STPR: Symptomatic True Positive Rate

OVA: Overall Validation Accuracy

ATPRV: Asymptomatic True Positive Rate in Validation

STPRV: Symptomatic True Positive Rate in Validation

OTA: Overall Test Accuracy

STPRV: Symptomatic True Positive Rate in Test

ATPRT: Asymptomatic True Positive Rate in Test

Index

1. Preface

1.1 Origin of the study

Brugada Syndrome (BrS) is a rare but serious condition that can cause unusually fast heartbeats that can be life-threatening in some cases. The long-term outlook for people with BrS is uncertain because the condition is very unpredictable. Many people with Brugada Syndrome don't show any symptoms and don't even realize they have it. Other experience dizziness, sporadic noticeable heartbeats, seizures or blackouts. Unfortunately, in some cases, the first symptom is death.

Aside from avoiding certain habits and medicines, in cases of high risk of developing a dangerously fast heartbeat, specialists recommend having an implantable cardioverter-defibrillator (ICD) placed. They don't prevent a fast heartbeat, but can help stop it and avoid them becoming life threatening. Sadly, these devices are not without risks (infections, inappropriate shocks, etc.), so it's key to identify individuals at high risk of life-threatening arrhythmias.

Affected people with a history of sudden cardiac arrest or fainting are known to have an increased risk for suffering further episodes compared to people with no previous symptoms. However, in people without prior symptoms of BrS, the tools for evaluating the risk of sudden death are suboptimal and may require invasive interventions with associated risks.

This project is part of a collaboration between the cardiology department of the Hospital Clinic, the Centre de Recerca de Investigació Biomèdica (CREB) and the Bioinformatics and Biomedical Signals Laboratory (B2SLab), and has been researching BrS in regards to its diagnosis for various years, among other heart diseases.

The main, long term aim of the collaboration is to use machine learning techniques for the identification of individuals carrying the BrS before symptoms develop, as well as those that have the highest probability of a life-threatening arrhythmia and need the placement of an ICD. This technology could potentially be better than currently available tools and avoid invasive interventions in otherwise healthy individuals.

2. Introduction

It is hypothesized that it is possible to prognosticate symptomatic Brugada Syndrome patients using machine learning tools to process biomarkers extracted from 24h electrocardiograms, possibly alongside clinical data.

2.1 Objectives

This project aims to test the initial hypothesis. That is to predict if any given Brugada Syndrome patient is symptomatic, based on machine learning tools and biomarkers calculated from 24h electrocardiogram (ECG) recordings. Symptomatic patients are considered those patients that have been recovered from cardiac death, suffered an arrhythmogenic syncope or sustained tachycardia.

In order to do so, an investigation of the current state of the art of the different fields of research is conducted. After that, 24h recordings around 45 different patients with Brugada Syndrome, including several symptomatic and asymptomatic ones, are processed and biomarkers extracted from them. These data are then used to train and later test different machine learning models.

On a long term perspective, the goal would be to use a successful model on the described task as a basis to develop other machine learning based models that could predict the development of future symptoms in previously asymptomatic patients.

More specifically, the goals for the project can be divided as:

- Develop an algorithm that allows a quick, automated reading of multiple consecutive electrocardiograms in the chosen software (Matlab), originally incompatible with it, in a desired format.
- Apply noise reduction techniques to eliminate as much of it as possible to the entire 24h ECG signal and while losing as little information as possible.
- Create an algorithm that can apply those techniques to many signals consecutively, streamlining its processing for future uses.
- Develop an algorithm to extract biomarkers related to the Brugada Syndrome from the processed signals.
- Determine an optimal way to combine those biomarkers with clinical data from their respective patients in order use both sets of information in conjunction, as well to arrange the combined information properly for its use on machine learning.

- Train and test multiple machine learning models with the resulting datasets in order to make an extensive but initial evaluation of their ability to classify Brugada Syndrome patients into symptomatic and asymptomatic.
- Analyze the obtained results to understand the functioning of the resulting models. This, alongside the previous one, are the main objectives of the study.

2.2 Scope

This work was developed in the framework of the project "Characterization and risk stratification in Brugada Syndrome through the development and validation of novel artificial intelligence model-based approaches for precision medicine, AI4BrS (Artificial Intelligence FOR precision medicine in Brugada Syndrome)", a project dedicated to the automatic and continuous analysis of electrocardiographic variables in relation to the risk of sudden death in patients with Brugada syndrome using Artificial Intelligence.

The AI4BrS project is mainly carried out at the August Pi i Sunyer Biomedical Research Institute (Consorci Institut d'Investigacions Biomèdiques August Pi i Sunyer) and Cardiology Department of Hospital Clinic (in Barcelona) with Dr. Elena Arbelo-Lainez, a researcher in the arrhythmia unit of the Cardiology Department of the Hospital Clínic, as the main investigator.

The main scope of the project is the assessment of several ECG biomarkers related with the prognosis for sudden death risk stratification in patients with Brugada syndrome. Following that, the main objective of the group is to develop an automatic algorithm that detects electrocardiographic parameters related to worse prognosis in Brugada syndrome.

In this work, the main interest is to assess several machine learning tools, using 24h electrocardiogram recordings to predict patients with Brugada Syndrome at high risks of sudden cardiac death. To carry this out, a database composed of various tens of patients suffering from Brugada Syndrome is already available from Hospital Clinic, including electrocardiogram signals obtained through a Holter recorder placed on symptomatic and asymptomatic patients attended at Hospital Clinic, for a period of 24 hours.

3 Background

3.1 The heart

The heart is a human organ made of muscular tissue that has two main functions: pumping blood from the lungs to the organs of the body, including itself, and collecting blood from them and carrying it into the lungs, in order to keep it oxygenated and free of metabolic waste [1] [2].

It is structurally divided in four chambers, as can be seen in the image bellow (Figure 3.1). The upper ones are called atriums, left and right accordingly, and the lower ones are called ventricles, also left and right accordingly, both in relation to the perspective of the body it's on. The blood vessels, the conducts blood is pumped through, are named according to whether they carry blood into the heart, the veins, or out of it, the arteries.

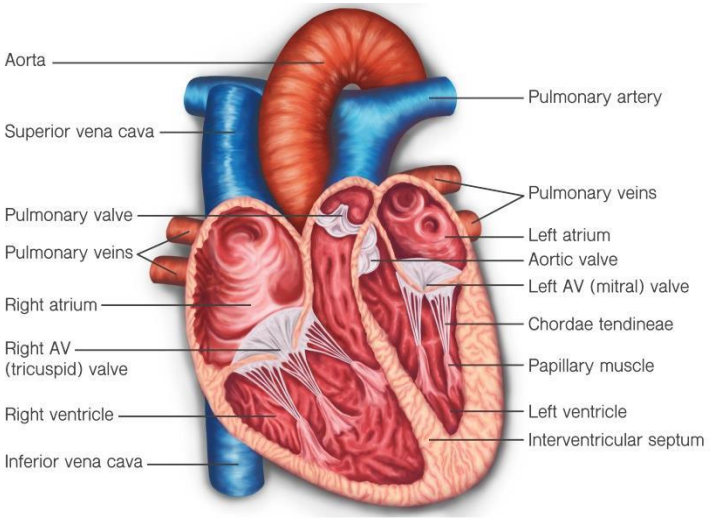


Figure 3.1 Anatomy of the heart with labeled parts. The blood vessels colored in blue go from the organs and towards the lungs, the ones in red from the lungs and towards the organs [3].

The mechanism through which the heart pumps blood is called cardiac cycle, schematized in Figure 3.2. It's divided in two parts, systole and diastole. During the systole, the ventricles contract, increasing blood pressure and ejecting blood from the heart. The diastole occurs when the ventricles relax after that, being filled of blood through the auricles [2].

It is located in the protective thorax, posterior to the sternum and with an oblique position, between the lungs and above the abdomen, with its upper part located below the third rib and its lower tip (apex) located between the fifth and sixth rib [1].

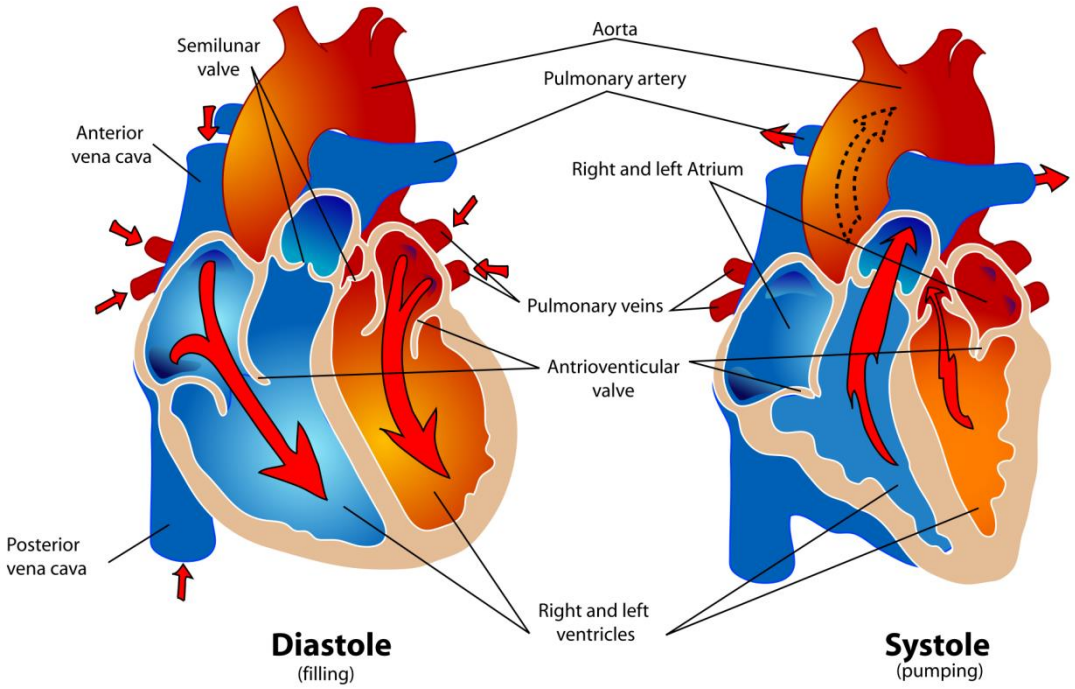


Figure 3.2. Blood flow of the heart during the cardiac cycle. The arrows indicate the blood flow on the heart during diastole and systole [4].

3.2 Electrocardiogram

The electrocardiogram (ECG) is a signal that registers the changes in the electric potential patterns of depolarization and repolarization of cardiac cells, captured by electrodes placed on the surface of the body, in order to represent cardiac activity of the heart [5]. On an ECG signal, the x axis represents time and the y axis voltage, usually in the order of mV.

ECG is considered the non-invasive gold standard to assess cardiac activity. It is extremely useful for cardiac diagnosis because of its simplicity, ease for the patients, low cost, and effectiveness in relation to those positive traits. ECG evaluation is widely considered the best technique for the diagnosis, among others, of cardiac arrhythmias, acute coronary syndromes and chronic myocardial infarction [6]. It is also the main criterion for diagnosis of the Brugada Syndrome [7].

Another fundamental element to consider is the concept of leads. An ECG lead is a graphical description of the aforementioned electrical activity of the heart and is obtained by combining the data proportioned by several electrodes on different parts of the body. There is a standardized 12-

lead ECG that requires the use of 10 electrodes placed on different positions. The leads can be divided into two groups: limb, and chest or precordial leads.

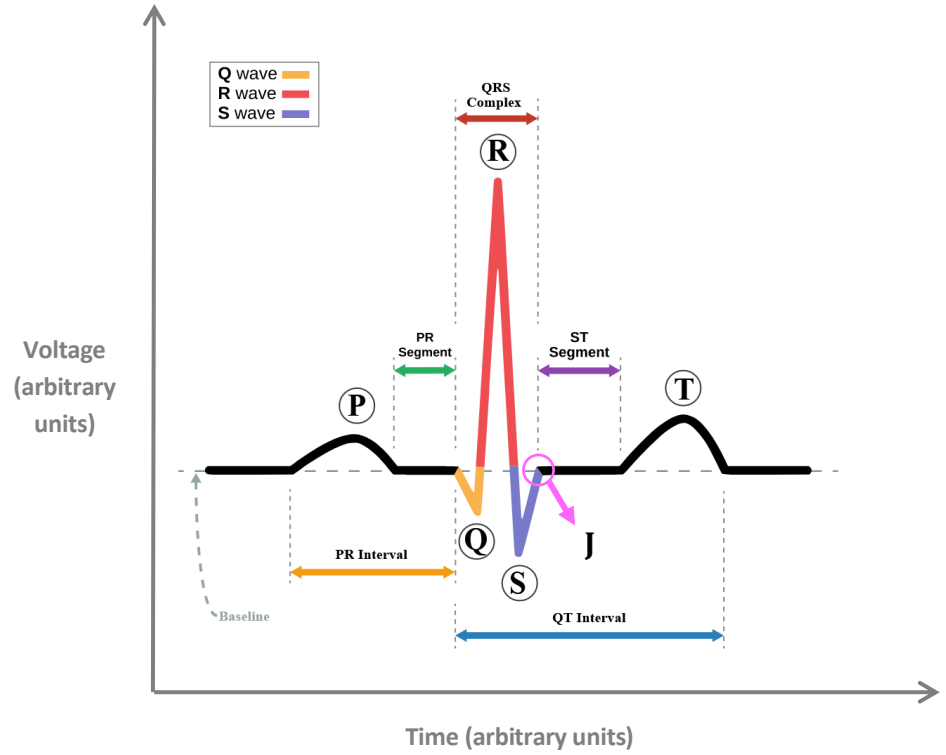


Figure 3.3. Schematic of the parts of a typical ECG, adapted from [8].

The resulting ECG signal of each heartbeat can be divided, as seen in figure 3.3, as follows [9] [10]:

- Baseline: also known as isoelectric line, it can be most easily spotted as the generally and approximately horizontal line that goes between one beat and another and it represents the basal electromagnetic currents that circulate through the heart when no active stimuli is happening.
- P wave: as an academic consensus, it's considered the first peak in the wave, and it appears because of the defibrillation, with its consequent contraction, of the atriums.
- PR segment: it's the part of the baseline that goes after the P wave and up until the next peak. In this period, the auricles are emptied and the transmission of electrical currents through the heart is slowed until the contraction of the ventricles.
- PR interval: it's the interval that goes from the begging of the P wave until the beginning of the QRS complex.
- QRS complex: it occurs because of the contractions of the ventricles, as they pump out the blood in them, and it's composed of three consecutive waves, Q, R and S. The Q wave is negative and usually small, the R wave is positive and can vary widely in height, and the S wave is, again, negative.
- ST segment: it's the part of the isoelectric line that goes from the end of the S wave, also referred to as J-point, and up to the beginning of the T wave.

- QT interval: the part of the signal found between the beginning of the Q wave and the end of the T wave.
- T wave: it's considered the final wave in the ECG signal and is, in most cases, a positive wave. It occurs because of the repolarization of the ventricles' cells.

3.3 Biomarkers

Biomarkers can be defined as quantitative indications of medical state observed from outside the patient that can be measured accurately and reproducibly [11]. They are a very broad subcategory however, being described by researchers in collaboration with the World Health Organization as a term used broadly to refer to close to any measurement (physiological, biochemical at a cellular level or molecular interaction) between a biological system and a potential hazard, be it chemical, physical or biological [12].

Biomarkers are fundamental in biomedical research, this project being an example of just that, but they are often advised to be used with caution and a systematic perspective, as an inadequate use of biomarkers can run the risk of obfuscating the medical realities behind them. Shortcomings of their measurement and factors external to the core of the research must be taken into account to use them properly. It's also recommended to periodically revise their use and effectiveness, as well as to make an effort to always contrast conclusions taken from them with clinical outcomes [11].

The following biomarkers have been used on throughout the present work (the units they are often and have been calculated on in this project are written between parentheses):

- **RR intervals (ms):**

The only biomarker used in the project that was obtained in order to calculate another one, the QT corrected specifically, but not used in the machine learning process.

In and of itself, it is simply the time interval between the R peak of one heartbeat and the next one. In this project, it has been used by averaging every RR interval in a segment of a signal.

- **Average power of the QRS complex or $Pavg_{QRS}$ (mV^2):**

The average power of a continuous signal is calculated as:

$$Pavg_{QRS} = \frac{1}{T}\int_0^T |x_a(t)|^2 dt \qquad \text{(Eq 3.1)}$$

Where x_a(t) is the QRS, beginning at 0 and T being its duration (QRSd). Therefore, the average power of a digitized QRS, where N is the number of QRS samples, is given by:

$$Pavg_{QRS} = \frac{1}{N}\sum_{n=1}^{N}|QRS(n)|^2 \quad \text{(Eq 3.2)}$$

- **Absolute area of the QRS complex or $AbsArea_{QRS}$ (mV·s):**

The QRS area on a continuous signal would be calculated as:

$$Area_{QRS} = \int_{QRSon}^{QRSoff} QRS(t)dt \quad \text{(Eq 3.3)}$$

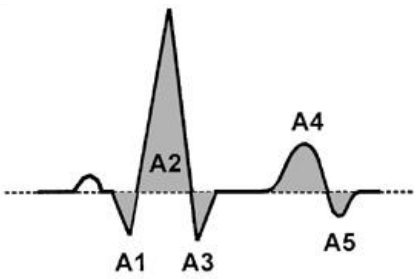
In a digital signal sampled at sampling rate fs, approximating the integral with the rectangular integration method, is

$$Area_{QRS} = \frac{1}{f_s}\sum_{n=1}^{N} QRS(n) \quad \text{(Eq 3.4)}$$

And finally, the absolute value of the QRS area is

$$AreaAbs_{QRS} = \frac{1}{f_s}\sum_{n=1}^{N}|QRS(n)| \quad \text{(Eq 3.5)}$$

The difference between the two is illustratet in the following image:



QRS Area: -A1+A2-A3
Absolute QRS Area: |A1|+|A2|+|A3|

Figure 3.4. Example of calculation of the QRS area and the Absolute QRS area. Adapted from [13].

As can be seen, simply adding A1, A2 and A3 together would mean, on a typical ECG heartbeat like the one in the image, subtracting the value of A1 and A3. As in this case the goal is to account for the entire area of the QRS segment, the absolute values of the areas will be used instead

- **QT interval or QT (s):**

Time interval between the end of the QRS complex (J-point) and the end of the T wave.

- **Corrected QT or QTc (s):**

Normalized calculation of the QT interval as to account for the heart rate, using the Barzett equation [14], considered adequate when examining an ECG with an average heart rate [15]. The RR interval is divided by 1 second in order to account for the unit adjustment.

$$QT_c = \frac{\mathrm{QT}}{\sqrt{\frac{\mathrm{RR}}{1s}}} \qquad \text{(Eq 3.6)}$$

- **ST_0 or J-point (mV):**

Voltage amplitude of the ECG signal at the J-point, where the ST segment begins. In the case of this project and the digital signals worked with, it was calculated by obtaining the median found between two samples prior and two samples after the J-point.

- **ST_{60} or J-point + 60ms (mV):**

Voltage amplitude of the ECG signal 60 ms after the J-point. In the case of this project and the digital signals worked with, it was calculated by obtaining the median found between two samples prior and two samples after the J-point.

- **ST slope (µV/ms):**

Slope formed by the two previous biomarkers and time (60 being the time between them in ms), as can be seen in Figure 3.5, calculated as:

$$ST_{slope} = \frac{ST_{60} - ST_0}{60} \qquad \text{(Eq 3.7)}$$

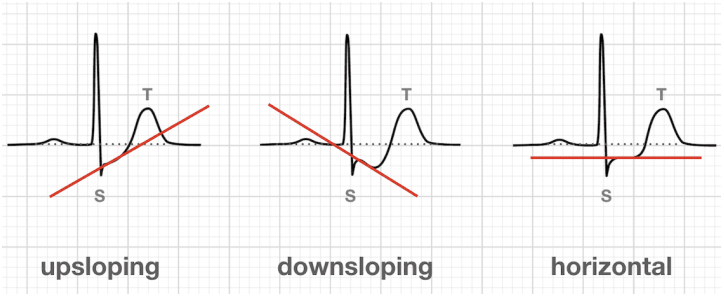


Figure 3.5. Types of slope of the ST segment adapted from [16].

3.4 Brugada Syndrome

The Brugada Syndrome (BrS) is a rare heritable cardiovascular disorder characterized by a peculiar electrocardiogram pattern and associated with a bigger risk of suffering ventricular arrhythmia, cardiac fibrillation and sudden cardiac death [17] [18] [19]. It has traditionally been considered to appear in structurally normal hearts [17] [18], although recent studies seem to be challenging that notion and point to a complex combination and interplay of structural alterations and ion channel dysfunctions [19].

3.4.1 Diagnosis and Brugada morphology

It was first reported by Pedro and Josep Brugada in 1992, on a report concerning 8 patients of very different who had been resuscitated from sudden cardiac death and presented ST-segment elevation in the right precordial leads of their electrocardiograms [20].

Since then, the understanding of the specific Brugada ECG pattern has been widened and deepened quickly and significantly. Current expert consensus recommends diagnosing with Brugada Syndrome in two scenarios.

First, when a spontaneous BrS ECG type 1 pattern is detected (see figure 3.6). Second, when that same pattern is successfully induced by sodium channel blockers (a common diagnose technique for BrS) or fever, but only when a type 2 or type 3 ECG (figure 3.6) appear spontaneously and that is accompanied by a pertinent score the 'Shangai Score System' (see figure 3.8) [21].

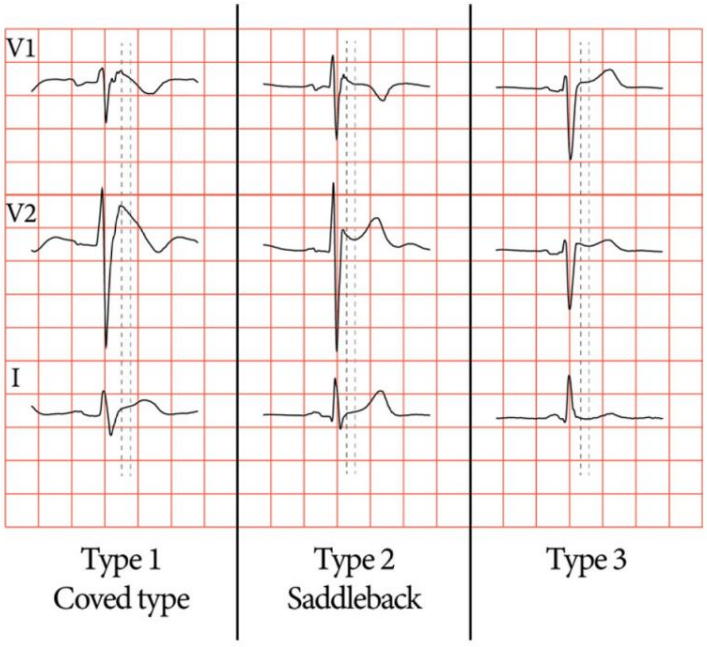


Figure 3.6. Schematics of the 3 Brugada ECG types, detailed below [21].

Type 1 is characterized by a lead V2 J-point (always marked with the first vertical grey line) elevation of 0.2 mV or bigger and a ST_{60} (J-point after 60 ms, second grey line) elevation of the same size. Compared to the standard ECG pattern, as depicted in figure 3.3, the QRS interval is also wider, and the S wave is noticeably wider and deeper in lead I.

Type 2 similarly has a V2 lead J-point, or ST_0, elevation of 0.2 mV or bigger, but a ST_{60} elevation of 0.1mV or more in this case, followed by a positive T wave. The T wave is less wide and deep than the one from type 1. The patient this signal is from developed a type 1 ECG when trying to induce it through specific medicine.

Type 3 also presents a V2 lead ST_0 elevation of 0.2 mV or bigger, but also a ST_{60} elevation of less than 0.1mV, alongside a T wave that is neither deep nor as wide as their counterparts in types 1 and 2. The patient this signal was recorded from also was successfully induced type 1 ECG.

Shanghai Score System	Points
I. ECG (12-Lead/Ambulatory)*	
A. Spontaneous type 1 Brugada ECG pattern at nominal or high leads	3.5
B. Fever-induced type 1 Brugada ECG pattern at nominal or high leads	3
C. Type 2 or 3 Brugada ECG pattern that converts with provocative drug challenge	2
II. Clinical History*	
A. Unexplained cardiac arrest or documented VF/polymorphic VT	3
B. Nocturnal agonal respirations	2
C. Suspected arrhythmic syncope	2
D. Syncope of unclear mechanism/unclear etiology	1
E. Atrial flutter/fibrillation in patients <30 years without alternative etiology	0.5
III. Family History*	
A. First- or second-degree relative with definite BrS	2
B. Suspicious SCD (fever, nocturnal, Brugada aggravating drugs) in a first- or second-degree relative	1
C. Unexplained SCD >45 years in first- or second- degree relative with negative autopsy	0.5
IV. Genetic Test Result*	
A. Probable pathogenic mutation in BrS susceptibility gene	0.5
Score (requires at least 1 ECG finding) ≥3.5: Probable/definite BrS 2–3: Possible BrS <2: Non-diagnostic	

*Only award points once for highest score within this category

Figure 3.7. Shangai Score System for the diagnosis of BrS from type 2 and 3 ECGs [21].

Crucially however, patients presenting any other conditions, detailed in Figure 3.7, that may explain BrS patterns, referred to as phenocopies, should be excluded, requiring further examination [21].

Category	Conditions
ECG abnormalities	Atypical RBBB, Early repolarization
Electrolyte abnormalities	Hyper- and hypokalemia, Hyponatremia
Cardiomyopathy	Ventricular hypertrophy, ARVC/ACM
Infectious	Acute pericarditis/myocarditis
Ischemic	Myocardial ischemia or infarction of right ventricle
Pulmonary	Pulmonary thromboembolism
Drugs	Antiarrhythmics Class IA/1C, Antianginals, Antihistamines, Tricyclic antidepressants, Propofol, Alcohol intoxication, Cannabis
Other	Dissecting aortic aneurysm, Myotonic dystrophy Mechanical compression of RVOT (e.g. pectus excavatum), Hyperthermia, Hypothermia

Figure 3.8. Examples of phenocopies of Brugada Syndrome [21].

It must be noted that the standardized 12-lead ECG precordial electrode collocation used for the detection of the BrS ECG patterns seen in the previous type descriptions, as is usual practice, is different than the usual one (see figure 3.9) [17] [22].

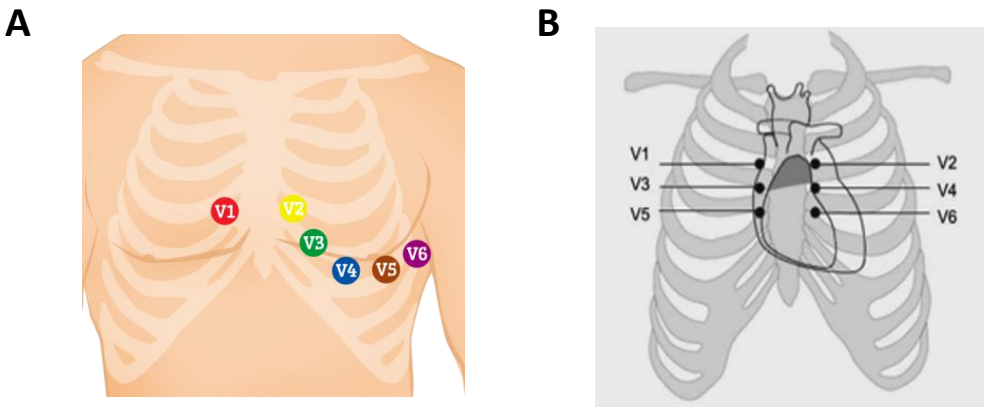


Figure 3.9. A: General standardized precordial electrode collocation for 12 lead ECG [23]. **B:** Standardized precordial electrode collocation for 12 lead ECG for Brugada Syndrome pattern detection [22].

BrS has been studied only for a very short period of time and is a fairly complex disorder, and so is its diagnosis. Ever-changing and with research still challenging many assumptions, further changes on its diagnosis criteria, or relevant findings in general about its diagnosis or prognosis would not come as a surprise.

3.4.2 Management and treatment

And any such finding could be very impactful. Although the data coming from a source previous to changes in the consensus for diagnosis criteria, pretty recent studies cited Brugada Syndrome as a cause for as many as 4 to 12% of all sudden deaths [17]. Estimates on prevalence vary significantly, between 0.12% and 0.8%.

Very importantly however, the management and treatment of BrS are a complicated issue [21]. All patients with BrS are advised to avoid potential triggers for ventricular fibrillation and sudden cardiac death, including excessive alcohol consumption and certain drugs, while developing high body temperatures because of fever should be avoided by taking adequate medicine, which is very relevant in the context of some vaccinations, such as COVID-19 ones. These last two measures require very precise and informed balancing of risks.

There is a unanimous consensus on recent work for advising the use of an implantable cardioverter-defibrillator (ICD) on symptomatic patients with BrS, which are highly effective for SCD prevention, despite discrepancies of the criteria for the symptomatic label [17] [21]. However, the implantation of one of these devices comes with high associated risks, such as potentially lethal device-related complications and a high prevalence of inadequate shocks.

These factors make an effective assertion of the risk associated to each patient a crucial task for the proper management of this disease, as well as the development of strong tools that could aid to determine such risks.

3.4.3 Regional differences in incidence

An aspect left to discuss is the differences in prevalence of BrS in different populations. Its genetic heritability, despite nuances in the genetic transmission of the disease [24], means that the disorder is unevenly distributed around different regions in the world.

A recent estimate, which was non-exhaustive, but fruit of a combination of multiple studies nevertheless [21], showed some of the highest prevalence of the type I Brugada ECG pattern to be in Thailand, with a 0.94%, and Iran, with a 0.359%. The US, Germany and Denmark, among others, had a prevalence of 0.005% or lower, and France, Italy and Greece a somewhat higher one, of 0.02% or bellow.

As most studies covering the topic also point towards [17] [18], the highest prevalence was unevenly found in different points of what could be considered Asia (although it must be noted how big, and

possibly unuseful, of a generalization that is), with Pakistan, Philippines and Japan all having over a 0.15% prevalence, aside from the two previously mentioned countries.

3.4.4 On sex in regards to Brugada Syndrome and its research

Differences in incidence and sudden cardiac rates between males and females are reported in every article that entertains the question and in every region examined [17] [18] [21]. However, contrary to populations of specific regions, where the heritability of BrS directly explains this occurrence, the causes for this discrepancy between sexes are not immediately apparent [17] [21].

Several factors point at sex hormone levels, linked to transmembrane ion current expression, being the fundamental cause for these differences, as prostate cancer treatment that involved decreasing testosterone levels minimized sex-related BrS ECG differences, higher testosterone levels have been detected in men with BrS, and reported differences between males and females are much lower in younger people [17] [25].

A recent study focused on examining current knowledge about BrS in women considered that 'current knowledge supported that sex-related differences not only to sex hormone levels, but possibly also to a complex interplay of gender and age dependent genetic factors that modulate the expression and function of cardiac ion channels' [25]. It, nevertheless, considered further studies necessary to determine the pathophysiological mechanisms underlying such differences. That, however, does not seem to dispute the relevancy of sex hormone levels in the previously mentioned situations, nor does it point to a concrete mechanism for that expression.

A factor that is likely playing a role in the difficulties for the finding of these differences is the shortcomings of the concepts of sex itself, understood as the binary and mutually exclusive classification of humans into males and females at or before birth upon inspection of the babies' genitalia, a classification that assumes an expected development of a certain range of expression of their characteristics of sexual dimorphism, which typically, but not always, occur [26] [27]. It must be brought up that recent literature has often challenged the general understanding of these categories, despite the scope of this project not allowing for a deep dive into those [26] [27] [28].

What the shortcomings of sex as defined earlier definitely cause in regards to the BrS is a very difficult assessment of the risk suffered by anyone who doesn't clearly fit the categories of 'male' or 'female', such as intersex people and trans people on hormone replacement therapy (HRT). The word trans is being used here to refer to anyone whose gender does not align with their assigned gender at birth, assigned alongside sex, which was explained before.

What risk are trans men on HRT comparable to? They have the usual levels of testosterone of a cisgender man (cisgender being someone who is not trans), but their genetic factors at play are most

likely different. This is a common issue, among many, that trans people face regarding healthcare [29]. And in this case, it is clearly an unresolved matter, as the mechanisms that produce sex-related differences are unclear in general, but they are in a particularly precarious condition, because researchers often don't consider their bodies' situation as a possibility to ponder at all.

And being aware the previously exposed information about hormone levels could be very useful for anyone wanting to better understand the risks associated to a particular group of intersex or trans people on HRT.

However, aside from being ultimately inconclusive, it required specifically searching for that specific information among academic texts and navigating various sources. Far from a difficult task for someone used to it, but one that may be inaccessible to people unfamiliar with academic research, a majority of the population, specially accounting for the amount of articles blocked behind paywalls. Less researchers and medical professionals will be aware of that information too.

That inaccessibility occurred both because of the long amount of time that took for the reasons behind sex-related discrepancies to be hypothesized about (coupled with an unfortunate lack of effort to include female patients in studies) [25], and because those differences were usually mentioned focusing only on the resulting differences in populations, with general disinterest to explore their causes, known or not. This could change if more researchers avoided glossing over the reasons behind sex-related differences and gave at least a brief mention of them, as they necessarily have to be, or the shortcomings of the research in the area should be otherwise noted.

On a final note, it must be added that the words 'sex' and 'gender' are often used interchangeably in most, if not all of the previously cited works, as well as other literature consulted, as well as terms such as 'women' and 'female'. However, sex and gender are long since commonly accepted to be separate, even if interconnected, concepts, with gender not referring to biological aspects and instead to social ones [27] [28] [30].

A rigorous definition of gender would need a wider scale discussion that goes beyond the scope of this project, but it is unlikely for Brugada Syndrome patients to have been asked for their self-identified gender at any point of the research or treatment. The studies do, most likely, refer to biological sex in its traditional binary notion, although it may be unclear at some points if social behavior correlated to gender is being considered at some points [25], which makes their reading unclear at times.

None of this goes to say that differences related to sexual dimorphism should not be considered when examining BrS, quite the opposite. Current use of assigned at birth sex in BrS research as a binary category to separate the population into two different, mutually excluding groups, even if

lacking [25] has shown positive results and is clearly preferable to the ignoring of sexual dimorphism when studying the disorder.

However, an approach that considers focusing on the specific physiopathologic mechanisms underlying the differences between people with different sexual dimorphism characteristics, instead of the categories themselves, could benefit future research and would definitely benefit trans and intersex people.

Further examination of this topic is outside the scope of this project, but could definitely be expanded upon.

3.5 Noise reduction by Signal Averaging

Signal averaging is a signal processing technique that can be used to reduce noise in periodic and quasi-periodic signals [31] [32]. Noise which will inevitably appear in some level or another in continuous, digitally stored ECG recordings, mainly from muscle activity and main line interferences.

Continuous, digitally stored ECG recordings can't avoid being contaminated with noise, mainly from chest muscle activity and main line interferences. As the ECG is a quasi-periodic signal, it can be divided into equally sized segments (the different heartbeats) by being aligned in reference to, for example, their R peaks. Thus, transforming the original signal into an ensemble of M different beats, with each beat, $x_m(n)$, described by N samples (see equation 3.8) [31].

$$x_i(n),\ i = 1, \dots, M; n = 1, \dots, N \tag{Eq 3.8}$$

The ensemble is represented by the N x M matrix X

$$\mathbf{X} = [\mathrm{x}_1\, \mathrm{x}_2 \cdots \mathrm{x}_\mathrm{M}\,] \tag{Eq 3.9}$$

Where the m^{th} beat is contained in the column vector

$$\mathrm{x}_\mathrm{m} = \begin{bmatrix} x_m(1) \\ x_m(2) \\ \vdots \\ x_m(N) \end{bmatrix} \tag{Eq 3.10}$$

Although more complex ways of performing the averaging exist, the average signal, $\overline{\mathrm{x}_\mathrm{m}}$, can be, as has been in done in this project, calculated as:

$$\overline{x_m} = \frac{1}{M}\sum_{m=1}^{M} s_m(n) + \frac{1}{M}\sum_{m=1}^{M} \eta_m(n) \quad \text{(Eq 3.11)}$$

Where $s_m(n)$ could be considered the deterministic part of the signal, which does not change from beat to beat, and $\eta_m(n)$ the noise. Assuming random noise, signal averaging will produce a reduction of the root mean square amplitude of the signal, a $\sqrt{M}$ factor. Figure 3.10 depicts schematically the signal averaging process, aligned with the R-peaks of QRS complexes.

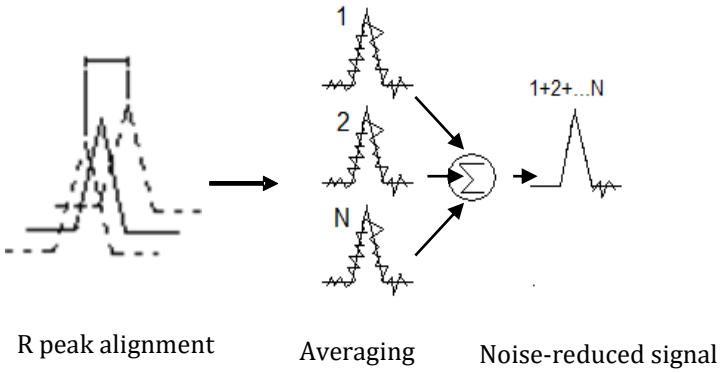


Figure 3.10. Diagram representing the signal averaging of an ECG. R waves are aligned and later summed, creating a noise reduced signal. Adapted from [33].

3.6 Machine Learning and Classification

Machine learning (ML), concept whose introduction is attributed to Alan Turing [34], although initially using the slightly altered but much more stylized 'learning machines' in his 1950 article 'Computing machinery and intelligence' [35]; can be broadly defined as enabling computers to make successful predictions using past experiences [36].

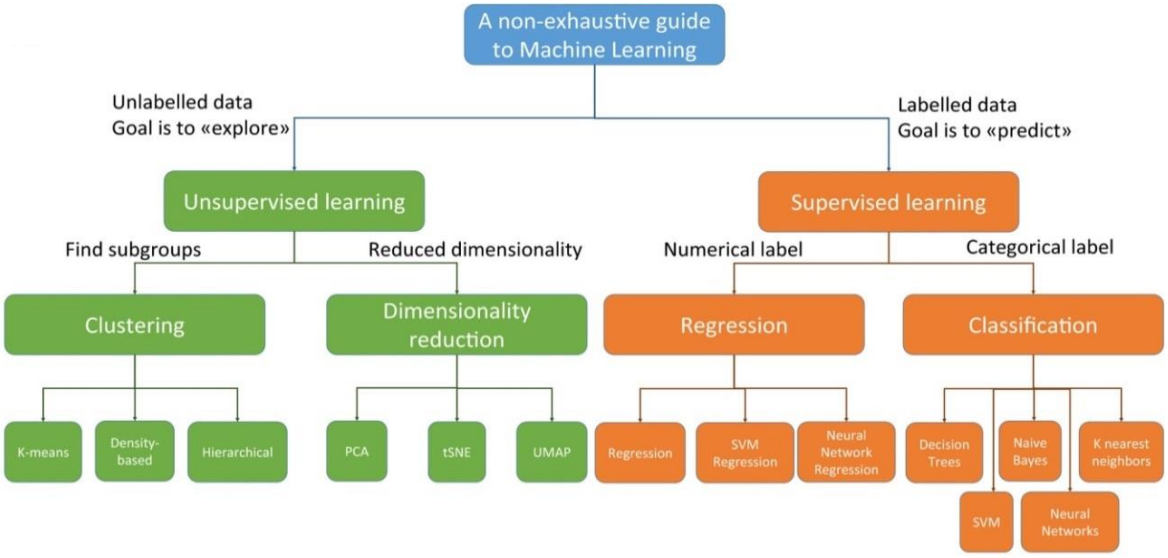


Figure 3.11 Non-exhaustive schematic of the different types of Machine Learning [37].

As can be seen in Figure 3.11, ML can be divided into two broad categories, unsupervised and supervised learning. On the latter, training entries are presented as a set of variables and a category or class linked to each set, and the algorithm develops a function which relates the variables of the entry with a given outcome. Unsupervised learning tools have no prior knowledge of the entry variables, they find common characteristics or occult patterns among them.

Machine learning is a vast field, which has gained a lot of popularity in recent decades but has been discussed since the 1950s, as shown just before. In this section, however, the focus will be on introducing the concepts more directly applied and explored on this project.

3.6.1 Machine learning classification

ML classification, or automated classification from now on, is a technique consisting in the use of machine learning algorithms to assign a specific class label to a group of examples (the so-called training group) in order to later apply that process to new data [34] [38].

Generally, in order to create a ML classifier, a portion of the available data for the creation of the classifier is separated from the rest. That set of data is called the test dataset. The test dataset is used to evaluate the final classifier with data that has not been used for the creation of the model (specific ML classifiers are often referred to as 'models').

The rest of the data, on its turn, has to be divided in order to train a classifier, which can be done in multiple ways. A common approach for large dataset, and the approach taken in this project, was to create two separate datasets as well: the validation data (or validation holdout) and the training data.

For the training itself, the model gives predictions using the variables and following different criteria depending on its type, which are then compared with the classification targets. Depending on the results, the type of model and its specific settings, the parameters of the model will be changed, with the goal of maximizing the accuracy of the predictions.

During that process, the model in training is used to predict the desired categories of the validation dataset, which has not been used in the training prior to that. That provides an unbiased evaluation of the model, with which the model adjusts its parameters further in order to maximize its accuracy.

Finally, once the training is completed, the model should be evaluated with the test dataset, which has remained independent during the entire process. The results for the test, provided that its data is adequately independent from the data used in the training process and it is large and diverse enough, should give an accurate estimate of the performance of the model with any new data processed through it.

3.6.2 Misclassification costs

A concept strongly linked to ML classification in general and this project in particular, is the concept of misclassification costs or misclassification weights (McW) [34] [39].

As explained before, models are adjusted in training according to their performance on predictions, keeping or changing adjustments according to whether they have a positive or negative impact on accuracy.

The calculations that determine said accuracy can be modified by changing its misclassification costs, which means the relative weight of wrong predictions of particular types. For example, it might be desirable to set a higher McW for predicting a sick patient as healthy than for doing the opposite.

As McW have no units, they can be expressed as ratios.

3.6.3 Automated classifier types

Here, some of the main types of machine learning classification, the ones most relevant to the project, will be enumerated briefly [38]:

- Decision Trees.
- Discriminant Analysis Classifiers.
- Logistic Regression Classifiers.
- Naïve Bayes Classifiers.
- Support Vector Machines.
- Nearest Neighbor Classifiers.
- Kernell Approximation Classifiers.
- Ensemble Classifiers.
- Neural Network Classifiers.

4. Materials

Most of the work of this project has been done in an HP Pavilion Laptop 15-cs1xxx, with the ECG data being transferred on a portable solid state drive of 1 TB of capacity, model Samsung Portable SSD T5. The entire dataset comprised dozens of gigabytes of information and was therefore impractical to store in a computer destined for daily use and without a huge storing capacity.

However, this is largely inconsequential to the project itself. The material described in the following sections is that which had a fundamental role in the shaping of the outcome of the project.

4.1 ECG Holter recorder

A Holter consists of the uninterrupted and ambulatory (without the need to stay at the hospital) register of an ECG from a patient, typically for no less than 24h, through a portable electronic device that both records and stores the information, often of homonymous name [40].

4.1.1 SpiderView Digital Holter Recorder

The Holter monitors employed to record the data used in this project were SpiderView® models [41], from the company Livanova [42] and distributed by MicroPort [43].

Figure 4.1 Spider Digital Holter Recorder [41].

4.1.2 Holter recorder placement

As exposed previously on the background section, the standard for 12-lead ECG recording for the examination of Brugada Syndrome ECG patterns is different from the general one (see figure 3.9). The placement of the Holter was approximately similar to figure 4.2.

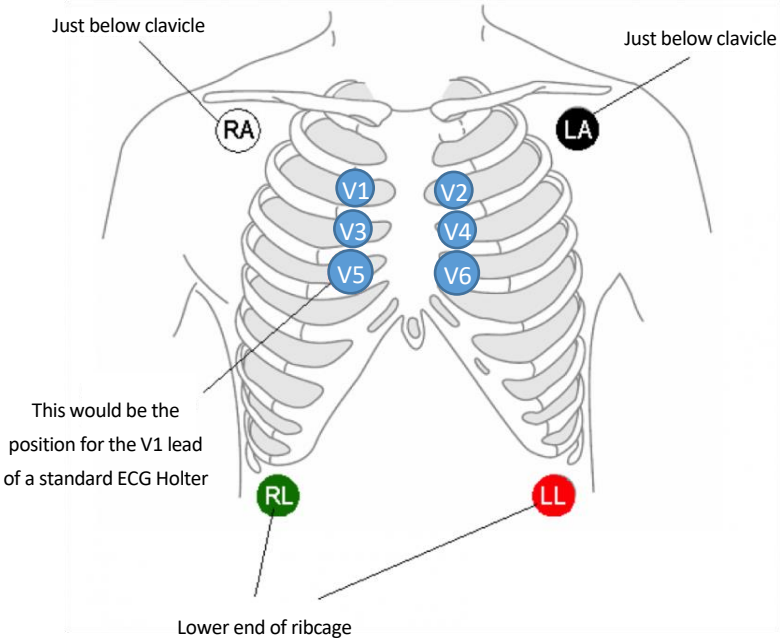


Figure 4.2. Schematic of the placement of the Holter used in the project. Adapted from [44].

4.2 Database

The database used in this project was provided by the Arrhythmia Section of the Cardiology Department of the Hospital Clínic of Barcelona.

ECG signals from BrS patients were obtained using the aforementioned Spiderview Plus 12-lead ECG Holter recorders [41] using a configuration of 2.5 µV of voltage amplitude resolution (16 bits) and sampling frequency of 1000 Hz for each channel. ECG signals from the recorder flashcard were uploaded to Synescope analysis software [41], also from Livanova [42], which allowed exporting the ECG signals in ISHNE format [45] for posterior analysis. The ISHNE files were the ones initially available at the beginning of the project.

Signals of 53 different patients were available in the initial database. However, only 47 of those were found to have their complete, corresponding clinical data available, including their status as symptomatic or asymptomatic, so that was the number of signals that was finally worked with. However, 2 of them were later discarded, so the final number ended up being 45.

The following table, Table 4.1, synthesizes basic information about the patients from the database. As only 45 patients were finally used, the information of the patients whose signals were discarded has not been included.

Table 4.1. *Number of asymptomatic and symptomatic patients on the database, separated between females and males, and with the mean age and age range at the time of recording specified*

Number of patients	Asymptomatic	Symptomatic	Age[1]	Total
Females	8	3	-	11
Males	29	5	-	34
Total	37	8	54 [29-79]	**45**

[1] Mean [Minimum - Maximum]

Symptomatic patients were defined as those having recovered from sudden cardiac death, suffered an arrhythmogenic syncope and/or sustained tachycardia that derived to an appropriate therapy response (electrical shock or anti-tachycardia stimulation) from their implantable cardioverter-defibrillators. That is the definition of the term that will be used throughout the project.

As can be seen, there are more male and asymptomatic patients, as compared to their assigned counterparts. The most common group is, by a large margin, asymptomatic males, followed by asymptomatic females, symptomatic males, and symptomatic females being last, with only 3 patients. A more detailed assessment of these imbalances will be made when detailing the methodology followed.

The ages of the patients at the time of the recording was estimated by the approximate date of creation of the files and the year the patients were diagnosed on, so some margin of error can be expected.

Each signal is stored as a 24 hour file. However, as will be discussed further bellow, the last portions of most recordings are flat 0 signals, leaving usually between 21h and 23h of actual recording.

The total of the biomarkers finally used were the result of almost 932h of ECG recording. The smallest dataset used in the classifier, including training and testing, obtained its data from over 330h of recordings.

Before the recording, each patient agreed to share their data through a document of informed consent that can be seen in the Annex E.

4.3 Matlab

Matlab is a programming and numeric computer platform designed by MathWorks for use in science and engineering primarily [46] [47]. It allows for many different uses with great versatility, in no small way thanks to its toolboxes, which expand the capabilities of the platform; and its apps, which are interactive applications that give additional tools for multiple technical tasks. Some toolboxes include apps that can be accessed on the Matlab workspace once installed.

In this project, programming was done exclusively in Matlab, alongside its signal processing and ECG-kit toolboxes, as well as the Classification Learner App, also from Matlab.

4.3.1 Signal Processing Toolbox

Matlab's Signal Processing Toolbox provides tools for analyzing, filtering and extract characteristics from signals in the form of several functions and apps [48]. Its provider highlights its capacity to design filters, do time-frequency analysis, and apply deep learning and machine learning to signal processing.

4.3.2 ECG-kit

The electrocardiogram kit, or ECG-kit, which will be used from now on, is a free software Matlab Toolbox, described by its creators, Llamedo Soria and Demski, as an application-programming interface, that allows to access and process cardiovascular signals in different formats, including the ISHNE format the signals of the database used in this project are on [49].

The aspects within the processing capabilities that will be more relevant to the project are the detection of the beginning, end and peaks of the different waves in ECGs. The toolbox provides different techniques through which to specifically detect R peaks, from which, after consideration with the director of the project, the Wavelet method [50] was the one used throughout the project.

4.3.3 Classification Learner App

The Classification Learner App from Matlab is a Matlab application that allows training, validating, modifying and testing automated classification models through a specifically designed interface that also facilitates examining the data used, among other things, see figure 4.3 [51].

It can develop classifier models based, as self-described within the app, on the following methods:

- Decision Trees.
- Discriminant Analysis Classifiers.
- Logistic Regression Classifiers.
- Naïve Bayes Classifiers.
- Support Vector Machines.
- Nearest Neighbor Classifiers.
- Kernell Approximation Classifiers.
- Ensemble Classifiers.
- Neural Network Classifiers.

It also allows the changing of misclassification costs and adjustment of several parameters of most of its models, such as the maximum number of splits in decision trees or the number of neighbors in the nearest neighbor classifiers.

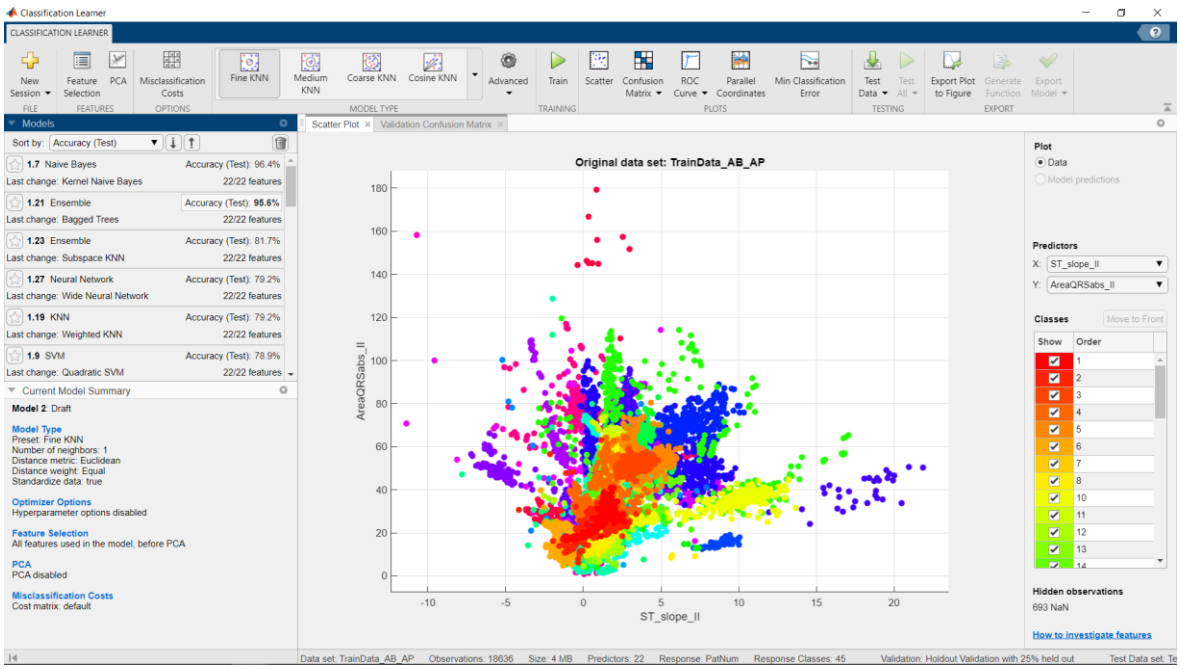


Figure 4.3. Interface of the Classification Learner App from Matlab, depicting a scatter plot of data.

5. Methodology

In this section, the process made from the obtaining of the raw ECG signals handed from Hospital Clínic up until the application of the Classification Learner App using the processed data will be described.

The following schematic briefly summarizes the different steps followed.

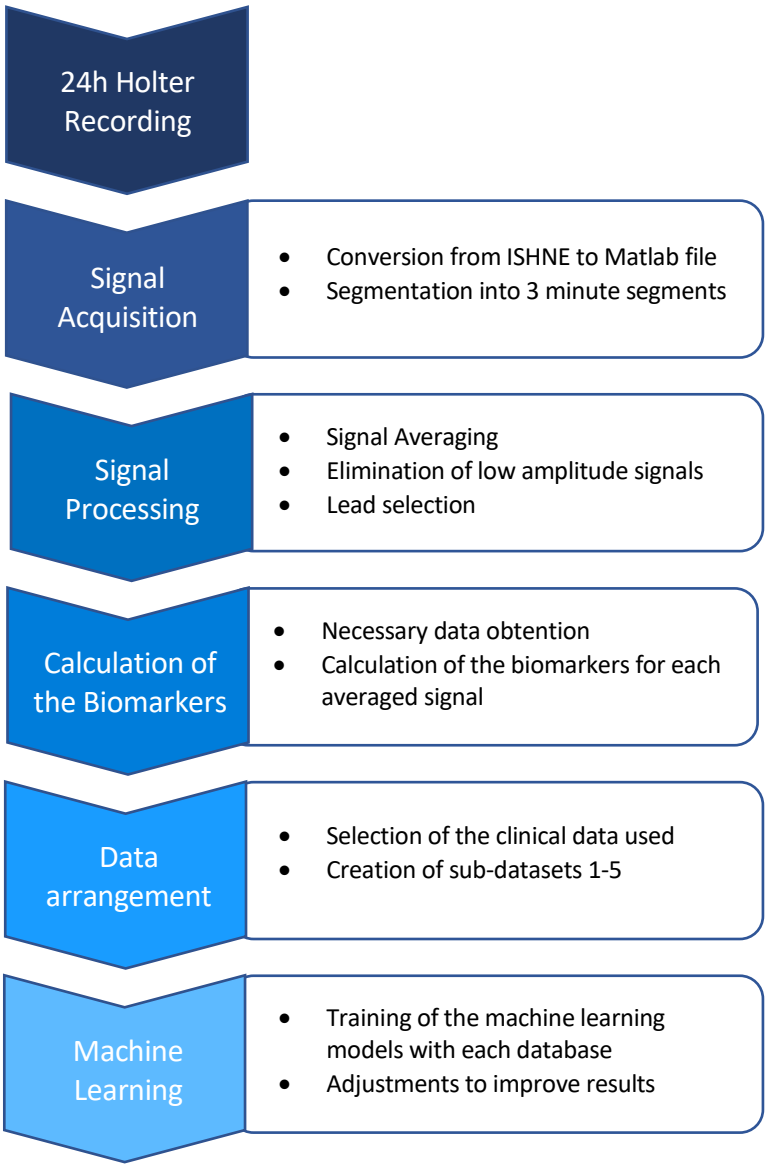


Figure 5.1. Summary of the process developed along the entire project, from the conversion of the raw files to the adjustments of the machine learning models to improve performances.

5.1 Signal Processing

The following section describes the process of conversion of the original ISHNE format file into a Matlab readable format and the noise reduction techniques used, mainly signal averaging.

Some low (under 1 Hz) and high (of 40 Hz and above) frequency filtering could remove important information from the ECG signal, and noise in both of those frequency ranges were quite predominant in the original signals [52]. In order to calculate the biomarkers, however, it is important to have a good signal to noise ratio, which could be obtained through signal averaging while leaving the deterministic signal more unaffected than some alternatives, as described in section 3.5.

Another crucial advantage of using this method is that it can be effectively applied to signals with many different types of noise, something ideal to filter 24h ECG Holter signals recorded under very different circumstances.

The results of the signal averaging process are shown in figure 5.3, page 31.

5.1.1 ISHNE file reading and segmentation

In order to access the 12-lead ECG signals form the patients, originally in ISHNE format [45], which can't be directly worked with through Matlab, the function *ECGwrapper()* from the ECG-kit was used.

This last command also requires an initial and final sample to read and later convert to int16 format, so this function was also used to splice the signal into 3 minute segments that will later be processed separately.

Once the 3 minute segments were obtained, they were first turned into double format, in order to make them easier to manipulate, and then removed their baseline through the *detrend()* function, from the Signal Processing Toolbox and changed its units from nV to mV.

Each file was saved after this processing in a .mat file containing both the signal in double format and a *struct* variable containing the original header, modifying only the number of samples, which would now be 180000 (3 minutes at 1000Hz of sample frequency). The resulting headers had the following structure:

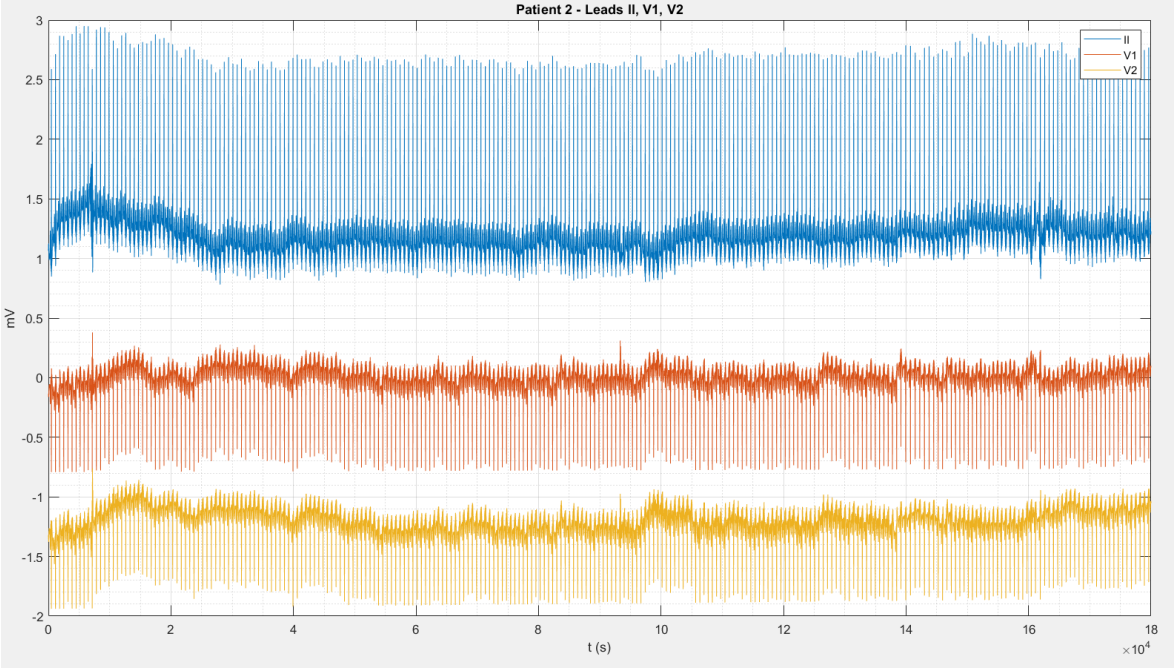
nsamp: 180000
btime: '11:8:59'
bdate: '1/1/2007'
nsig: 12
desc: [12×3 char]

gain: [12×1 double]
freq: 1000
adcres: [12×1 double]
adczero: [12×1 double]
units: [12×2 char]
recname: 'BH0001'

In the header *nsamp* represents the number of samples in the given file, *btime* and *bdate* the time and date of recording, although both of them were typically set to their default mode (1/1/2007). After that, *nsig* gives the number of derivations, 12 in this case, and *desc* their name (I, II, III, aVL, aVR, aVF, V1, V2, V3, V4, V5, V6). Finally, *gain* gives the signal gain ($4x10^{-4}$), *freq* the sampling frequency (1000Hz), *adcres* the resolution of the analog to digital converter (16 bit), *adczero* the signal's zero (0), *units* the original units of the signal (nV), unmodified as they were not going to be used, and *recname* the code assigned to each patient (BH0001 in this case), from which will the patient numbers are given. The only value that changes among files is *recname*, changing from patient to patient.

Each ISHNE file was 24h long, so processing each of them always resulted in 479 .mat files. However, due to Holter battery runtime limitations, almost every signal lasted less than that, resulting in some dozens of files containing zeros or very small numbers (after stabilizing sometimes), which had to later be removed from each patient's dataset.

A



B

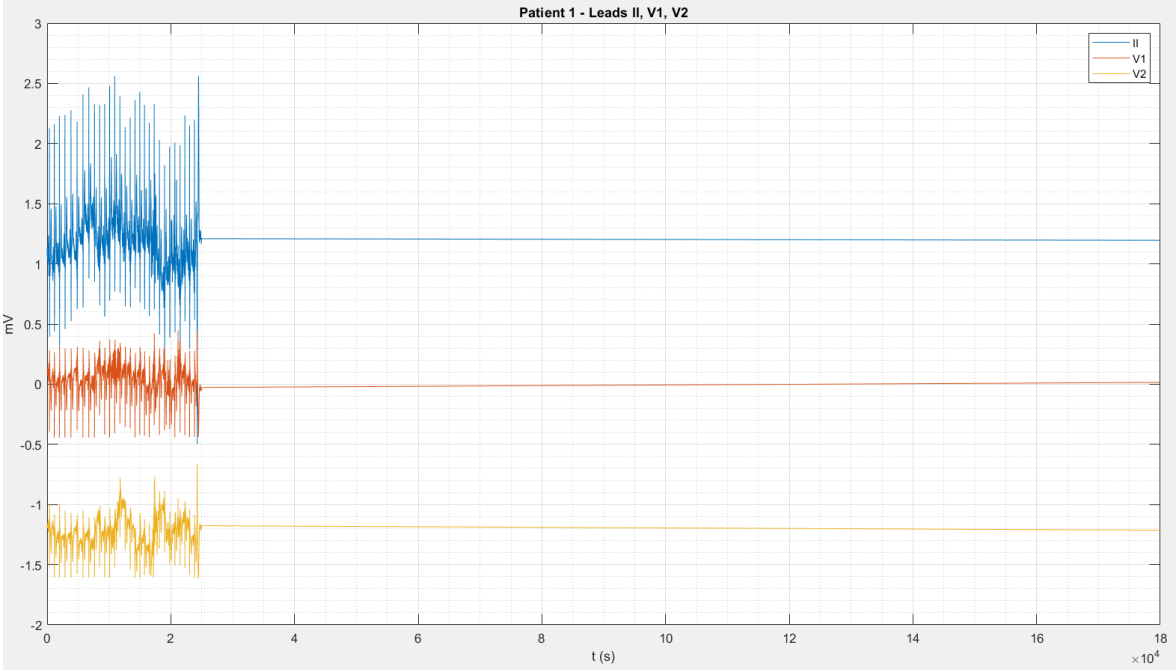


Figure 5.2. Signals were modified to appear cleanly in the figure, as will be the case for all signals displayed. The code used for that can be found in Annex D. **A:** Typical 3 minute segment of the ECG recording. **B:** 3 minute segment that began the empty segments in patient 1.

5.1.2 Signal averaging and lead selection

The biomarkers were calculated from a 60 second signal of a single heartbeat, obtained through the signal averaging of the beats corresponding to each 3 minute segment of the signal.

However, before performing the averaging, and just after reading each signal segment in order to minimize processing time, an *if* clause made it so only signals of an amplitude bigger than 0.5mV in at least one of its leads are processed. This sorted out the empty segments mentioned in the previous section.

Additionally, by setting a relatively higher threshold (the empty segments had often an amplitude of 0, and never anywhere close to 0.5mV), it also leaves out segments which, for one reason or another, had a very small amplitude and carried close to no or very unreliable information. The speculated reason for this occurring is errors in electrode leads connection.

The signals from patients 40 and 41 were finally discarded entirely because they mostly presented those same characteristics through their entire duration, reducing the total number of signals from 47 to 45.

After that, through the ECG-kit function *ECG.ECGtaskHandle.detectors*, the R peaks in the QRS complex were detected in order to use them as a reference point to align the signals for the signal averaging. Different detection methods can be used by calling this function, from which the Wavelet method was finally selected, as it showed a slightly better performance with many of the signals used.

The resulting file contains the position of the R peaks in the 12 leads the Holter originally recorded. However, in order to work with a more manageable dataset for this initial approach and perform the remaining processes faster, it was considered worthwhile to reduce the leads that would be used from this point forward to three: II, V1 and V2 (determined by the Holter placement specified earlier, see figure 4.2, with Brugada specific precordial leads).

These were considered by both the cardiologist and the director of the project to be the most adequate ones because they are among the ones who carry the most relevant information of the Brugada patterns, being commonly used for diagnosis [21].

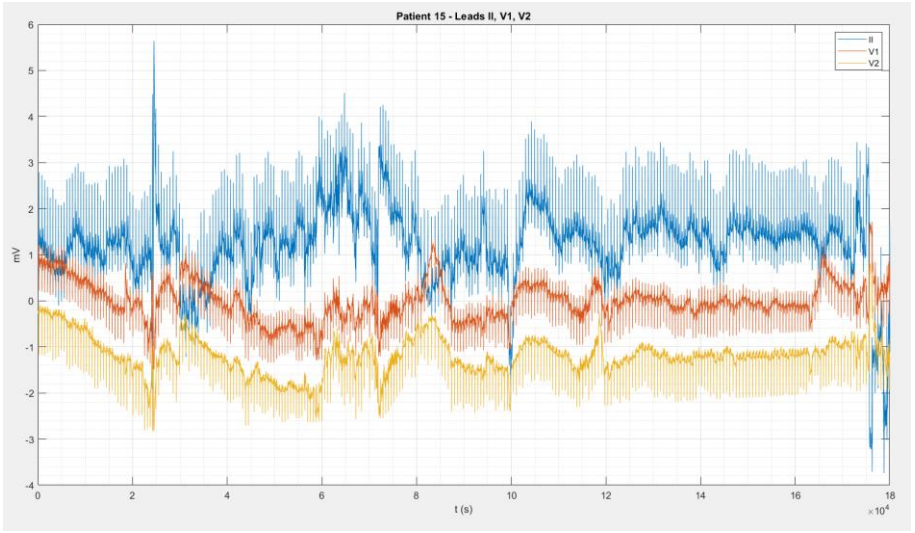
This lead selection was carried out during this part of the averaging, by iterating only through the selected leads on the files.

The selection of the signal fragments to average was then performed by selecting 1 second (1000 samples at 1000 Hz) windows around each R peak (399 samples before it and the remaining 600 after it, as an approximation for its common position), except for the first and the last 5 R peaks.

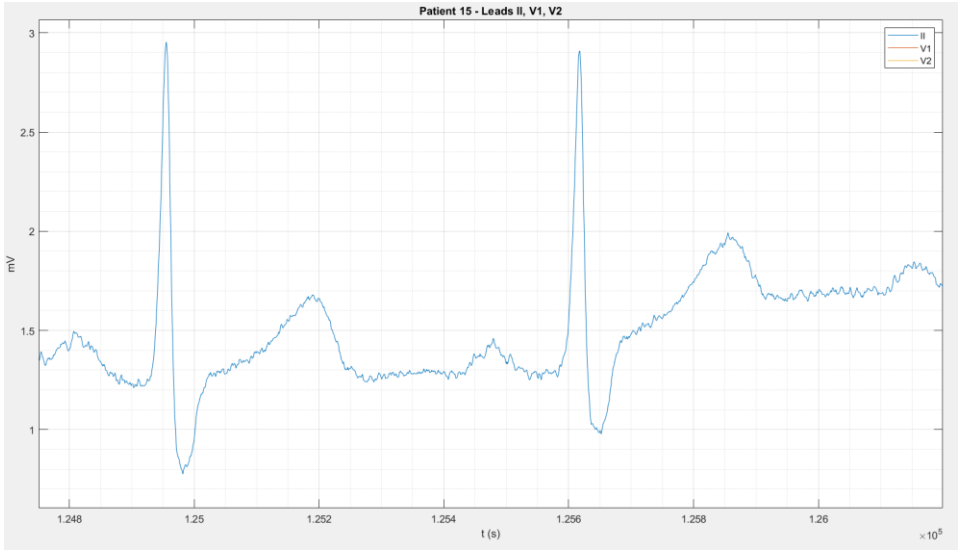
That was decided because, at times, not doing so caused indexing errors, as some R peaks, possibly incorrectly detected sometimes, appeared earlier than the 399^{th} or after the 600^{th} sample of the segment. The noise reduction on most of the segments was satisfying using the remaining 90 beats to perform the averaging regardless.

The resulting signal segments were then averaged with the function *mean()* to create a single 1 second signal. An example of the initial and resulting signals can be seen in figure 5.3.

A



B



C

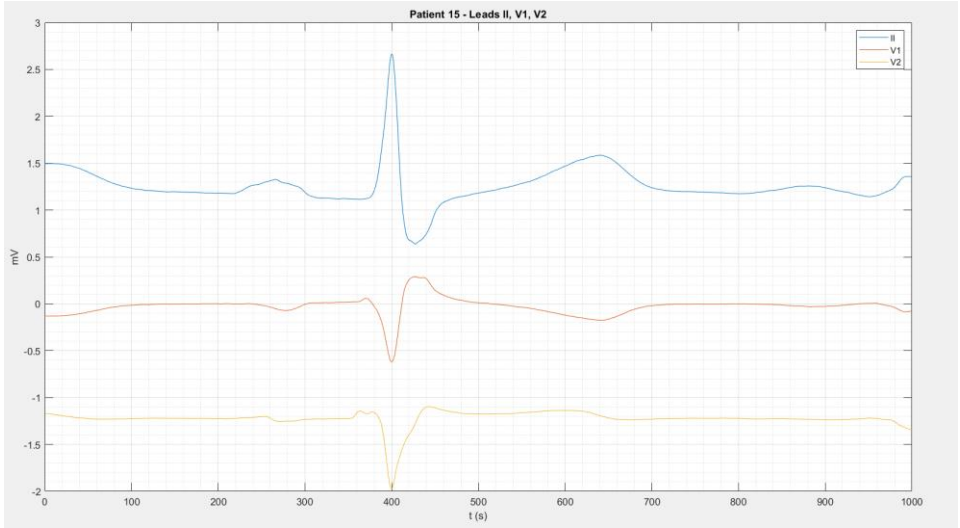


Figure 5.3. A: Original 3 minute ECG segment from patient 15. **B:** close-up of approximately two heartbeats on that same segment, lead II. **C:** Signal resulting from the averaging of the segment.

The resulting files with the averaged signals will be referred to as 'beat files' or just 'beats' from now on, with the latter being used to refer to the information they carry as well, regardless of the format.

5.1.3 Recounting the final files to be worked with

A simple yet essential code was created in order to obtain the number of averaged beat files that resulted from each patient's signal after the processing, the position the first beat from each patient would have in the final dataset table and the total number of averaged beats that would be used. Those resulted in double variables named Nbeats, Pbeats and Tbeats respectively.

It can of course be included in another code file, but during the development of each step of the process it was useful to have them separated, both for better organization, to making it quicker to run the code after modification and to incentivize a more careful assessment of each process and its results.

The number of files for each patient were simply obtained by iterating through each file through a *for* loop for the patient number and a *while* loop for the beat number, and checking, through the function *isfile()*, if a file exists with the naming structure and numbers.

While the *while* loop made the code simpler, it had the side effect of not adding any valid file that came after a 3 minute segment file that could not be previously averaged. However, that was not considered as a big enough flaw to adapt the code to it, as in the only case where that happened, 120 averaged beats from patient 22's signal, the already noisy signal only seemed to get worse after that point (as it is to be expected when a given 3 minute segment doesn't go above the 0.5 mV threshold), and minimizing the weight of noisy signals in the final classification seemed positive either way.

An independent variable counted the total number of beats, which was useful to obtain the position of the first beat of each patient, by simply subtracting the number of beats of each patient to the total number of beats at the end of each patient cycle.

The three variables were saved in a single .mat file named '*BeatsData.mat*' that will later need to be loaded in following codes.

5.1.4 Calculation of the RR intervals

As was shown previously, in order to calculate the corrected QT biomarker it is necessary to know the RR interval of each biomarker, the distance between its R peak and the previous one. As the biomarkers are going to be calculated in a single averaged beat, the RR interval will also have to be calculated from the entire 3 minute averaged segment of the signal, as the RR interval can't be obtained from a single beat.

In order to both have a more organized code, be able to check the results of this particular process more carefully, and iterate the code that calculates the biomarkers, which takes a while to process, faster while writing it; it was decided to have this part of the process to be in a separate file.

However, the process to calculate the RR intervals themselves is fairly simple. For each patient and each segment of their signals, R peaks are detected through the same ECG-kit command that was used in the averaging, *ECG.ECGtaskHandle.detectors*, through the Wavelet method only in this case, and the time for each peak extracted.

Then, the *diff()* Matlab command is used to obtain the intervals from the previous positions, the intervals are divided by the sampling frequency to convert them into seconds, and the average of all the intervals is calculated in order to obtain a single value per segment.

The final version of this program include lines that check if *detection* files in a given folder for each beat file already exist, be it because of previous interrupted iterations or different processes that used the same command, in order to avoid unnecessary repetitions of processing, if possible.

The final results were saved in a .mat file containing a double variable, named '*RRmean.mat*'.

5.2 Calculation of the biomarkers

Before starting to calculate the biomarkers, both the *BeatsData.mat* and *RRmean.mat* files need to be loaded, and both the patient numbers that will be processed and the leads their signals have properly determined.

Also, in later iterations of the program, an *if* clause was added that will load previous versions of the final resulting file if it already exists in order to be able to stop the relatively long processing halfway through and continue it from where it had stopped, with minimal adaptations further in the code. If it does not, it creates a zeros matrix of the final size the table will have.

Once that is done, a loop is started that will comprise most of the following process, that goes both through each patient's folder, containing all of their beat files, and through each beat file itself. To properly iterate the later, the length of the loop is obtained from the number of files, saved in each according value from each patient on the *Nbeats* double variable, contained in the *BeatsData* file created previously.

The final result from this script will be a single Matlab file in table format that will contain the biomarkers from each patient, ordered by patient number, which will be included in the first column. In order to set the position of the biomarkers of each beat right, their corresponding *Pbeats* is used, also from *BeatsData* which indicates the position of the first beat of each patient within the table.

An additional *if* clause checks on whether each of the rows of the table has already been filled or not, continuing the process only if hasn't. This is, again, to make the process quicker (as most of the processing time is spent on the delineation, which comes afterwards), which is only necessary and possible because of the *if* clause that may load a previously started but not completed table that was included before.

If possible, the delineation process would go next, right after loading the files. The delineation refers to the task *ECG_delineation* on the *ECG-kit* that calculates, among other data, the beginning and end of the different segments of each beat captured in the ECG signal, which will be used to calculate the biomarkers.

However, because of the programming of the function from the ECG-kit, the delineation can only begin after one second of the signal, something meant to avoid errors when detecting cut beats or other issues at the beginning of the recording, but which is only an obstacle in this case, where the signal records a single, one second, processed beat.

In order to avoid this, 1000 samples (1 second, as the sampling frequency is 1000 Hz) of the same value as the first one in the original sample are added at the beginning of the signal. The modified signal is saved in a new file.

A final if clause of this kind checks whether or not a delineation file for the two second beat file has already been created in the file they are two be saved in, and loads it in case it has. As this is way faster than the delineation process and is repeated thousands of times, it may save up quite some time.

It was especially useful while creating the code, as changes in desired presentation or plain mistakes made the final resulting file unusable, but the delineation files are always going to be the same as long as the signal stays the same too. This may be useful again if anyone tries to adapt the code to their specific needs or improve its efficiency in the future and several different attempts are needed.

After the delineation is performed or loaded, the beginning and end of the QRS segment and the beginning and end of the T wave are loaded for each lead, and the proper calculation of the biomarkers begins.

The biomarkers (in bold) are calculated as follows (omitting checks to avoid errors in case the delineation values are NaN, Not a Number), based on the formulas described in section 3.3 (Eq. 1-7):

```
qrs = signal(QRSon:QRSoff,k);  % QRS segment values are separated
QRSl = length(qrs);    % QRS segment length

Pavg = 1/QRSl*sum(qrs.^2); %  Average potency of QRS (mV^2)
AreaQRSabs = 1/fs*sum(abs(qrs))*1000; % Absolute QRS area (µV·s)


qt = Toff - QRSon +1;
qt = qt/fs;  % QT in seconds

% The particular RR value is selected for each beat
RRi = RRmean(i,Npat);

QTc = qt/RRi.^(1/2);  % QTc_Bazett vector
QTc = QTc*fs;     % Turned to miliseconds (ms)

ST_0 = median(signal(QRSoff+(-2:2),k)); % mV
ST_60 = median(signal(QRSoff+60+(-2:2),k)); % mV
ST_slope = 1000*(ST_60 - ST_0)/60; % µV/ms
```

This process is performed for every lead of each beat file, and repeated for all beats and patients. If the delineation function was unable to obtain any of the required data to calculate each biomarker, its result will be a NaN value.

The biomarker values are saved in a single file, with each line containing all the biomarkers for each beat, and each column containing a biomarker, except for the first one, with the patient number on it, which will be necessary on later processing. The three biomarkers of the same type but a different lead are grouped together, and the biomarkers go in the same order as presented before.

Once this process is over, the only step left is to save an additional file in excel format with its first row containing the different names of the biomarkers, added after transforming the variable containing the biomarkers into a cell, in the format seen in the code above.

5.3 Machine Learning

After having completed the calculation of the biomarkers, all data that was planned to be used on the classifier was finally available. As the biomarkers can be considered as a filtered condensation of all the original ECGs, the whole of them alongside their according clinical data will still be referred as 'dataset', 'main dataset' or 'database' if not specified in context that they refer to other sets of data.

The Classification Learner was always trained with data from this same pool, but the efficacy of the classifiers will change a lot depending on how the training and test datasets are selected from the main dataset, as well as whether or not all the data is used and how that selection is done.

The following sections describe the criteria used to separate each training and test dataset, as well as outline the criteria for which clinical data was finally used.

The Matlab Classification Learner App was considered the best available tool for the implementation of machine learning to this dataset within this project for different reasons. Although a secondary one, as other programming platforms could have been used, it being a Matlab app was a considerable positive, as it kept the process simple and allowed the application of machine learning through already familiar tools.

Mainly however, it was selected because it is a quite intuitive and adequate for users without much experience in the field, and allows for an extensive approach to training models, allowing to easily train several models at once that work very different principles. That is especially useful on a first approach to a particular goal for a machine learning model, such as this one.

Despite unsupervised machine learning being a completely valid and possibly promising venue through which approaching this problem, it was decided that supervised machine learning was a better fit for a first approach to the task at hand with the established tools.

The code used to add the clinical data, separate the dataset in different train and test datasets, as well as to separate carry out the process described on section 5.3.5, was fairly rudimentary. For the former, it iterated through the beats main file and added the clinical data desired

For the two latter ones, it simply iterated on a loop through the main database, already including the clinical data, and added the beats with the desired characteristics to the according files, following the criteria detailed on the following sections. Each arrangement of the data was created through a different Matlab file, but their code was nearly identical, as can be seen in the Annex C.

5.3.1 Use of clinical data

Using clinical data from the patients as a part of the machine learning process was a part of the conception of the project since its beginning, but various approaches were considered regarding its particular use through the course of the project.

The first aspect on this regard to be firmly decided was to not present the data from different beats to the classifier as belonging to particular patients, meaning that different beats were, in the 'eyes' of the classifier, independent from each other.

Directly including the patient numerical identifier as a parameter to analyze in the classification would not be of much use in and of itself, as models would not only correlate them to a single value (that links the beats to the patient), they would associate certain numerical values to being asymptomatic or symptomatic, as correlation would be absolute, and that would be absolutely counterproductive when testing the model with untrained patients.

Grouping beats from the same patient before feeding them to the classifier, which could be done in many different ways, is an interesting method to explore, as it could allow to identify cardiac patterns displayed over an extended period of time. However, this approach has its own downsides, as the number of samples in which to find whose patterns would be smaller, and it would require additional processing and analysis, and likely control and knowledge over the machine learning models, which time and scope limitations would not have allowed to do optimally.

The Brugada Syndrome, as stated previously, is diagnosed over patterns on specific waves, patterns captured in the selected biomarkers. The classifier, by processing individual beats, tried to identify particularities on the patterns, represented by biomarkers, of the symptomatic patients in particular. A remarkable benefit of this approach is that it maximizes the number of items the models were trained and tested on.

Another fundamental decision was whether to use clinical data about patient's pathologies and genetic data as parameters to be added to each beat for the classifier or not.

The final consensus was not to use them at all, as more or less apparent or direct interconnections may exist between those pathologies or genetic factors and the symptoms linked to the 'symptomatic' category, which could derive into the models relying on them not as risk factors but as direct predictors, diverting attention for the much preferable direct clinical analysis of the biomarkers. Also, not all of these data are or may be accessible for all patients on the current or future databases used, which would diminish the model's utility.

It was also considered to only use sex and age as additional clinical data. However, age was also considered to be a poor addition to this dataset as, despite being almost always available without the need for any test, it is to be expected for symptoms of Brugada Syndrome to have appeared more often in older people. Not only because of worsening health conditions in general with an older age, but also because more time has passed in which symptoms could have manifested.

That could increase accuracy, yet that increase would not be based on underlying patterns in the beats but in a statistical tendency. It could be considered a valuable addition to a model that already accurately identifies the proposed patterns that remained the core driving factor behind the classification, but that should be first achieved in a model proper.

Not including it even on such a model would also be a reasonable proposition if the goal were to perform an analysis of the beats only based on clinical information, and the person operating the model could later evaluate the additional risk themselves if desired. It could also be assessed that, despite people in a group being more unlikely to suffer symptoms, avoiding worst case scenarios is a priority, so efforts should be made to avoid that imbalance lessen their prognosis rates.

Finally, it was decided to only use sex as additional clinical data for the classifier, aside from the symptomatic or asymptomatic labels of course. Despite the nuance around the label, as exposed in the Background section, there are common differences in ECG patterns between sexes that could probably aid the classification of the beats, and it is often an easily available data that could be probably be fed to a model on further iterations or tests.

From now on, for the sake of convenience, the terms ‘symptomatic beat’ and ‘asymptomatic beat’ will be used when referring to the set of biomarkers extracted from each averaged beat, alongside the sex of the patient, if such patient is symptomatic or asymptomatic respectively, despite this categories not being strictly applicable to heartbeats themselves, be they product of signal

5.3.2 Creation of different train and test datasets

In order to be able, not only to create an automated classifier, but to also improve its performance, different test and train datasets must be created in order to understand its functioning, correct its biases, consider different scenarios, etc.

Describing the process of creating these and discussing the results obtained from the models required using some key words very often, so additionally to the some abbreviations being used in further sections, for the sake of convenience, the different distributions of training and testing data drawn and selected from the main dataset in the following sections are referred as ‘sub-datasets’.

Each of them has a number assigned in order to distinguish them easily. When a training dataset or a test dataset is mentioned, it is referring to the specific test or training set of data that are part of a specific sub-dataset, although it will be opted for alternative formulations as often as possible in order to avoid confusion.

On the next sections, the characteristics of each sub-dataset will be described alongside the considerations had when creating it, in order to properly contextualize the results shown in the following sections.

Table 5.1 displays a summary of the information presented in those sections. The patient identifiers used for each sub-dataset are shown, in parenthesis, in order to give more precise information about the data used for the tests and its possible biases.

Table 5.1. *Summary of the main characteristics of the sub-datasets*

	Sub-dataset 1	Sub-dataset 2	Sub-dataset 3	Sub-dataset 4	Sub-dataset 5 [Large Test | Small Test]
Separate patients for the test	No	Yes	Yes	Yes	Yes
Patients separated for testing [Total (Patient Identifier)]	-	4 (2, 10, 23, 33)	4 (2, 10, 23, 33)	2 (3, 5)	12|2 (1, 3, 5, 6, 14, 18, 29, 32, 35, 36, 37, 42)|(3, 6)
Beat Trimming	No	No	Yes	Yes	No
Manually selected signals	No	No	No	No	Yes

5.3.3 Sub-dataset 1

The sub-dataset 1 was composed by the totality of the beats in the dataset whose clinical data was available, with seven beats for each of the 45 patients, located always on the same positions (which were away from the very first and last ones, as they tended to be slightly noisier), being selected to conform the test dataset.

It was intended as a first approach to the classifier, using as much data as possible from as many patients as possible and with the least possible filtering of it. Even if the way of selecting this test dataset meant that it wouldn't be too reliable, it would give potentially useful and unique information about the behavior of the models with this general dataset.

As with the rest of the sub-datasets, the following two tables underneath show some general data about the train and test data from the dataset.

Table 5.2. *Patient data on sub-dataset 1 training and test*

Number of patients	Asymptomatic, Training	Symptomatic Training	Total Training	Asymptomatic Test	Symptomatic Test	Total Test
Females	8	3	11	8	3	11
Males	29	5	34	29	5	34
Total	37	8	**45**	37	8	**45**

Table 5.3. *Number and percentages of beats on training and test of sub-dataset 1*

	Total beats	Asymptomatic	% Asymptomatic	Symptomatic	% Symptomatic
Training Data	18639	15278	82%	3361	18%
Test Data	360	296	82%	64	18%

5.3.4 Sub-dataset 2

The test of the sub-dataset 1 allowed examining how well it could classify beats it had not been trained with from patients it had been trained with. If the model wasn't able to do that, it could hardly perform any further tasks correctly. However, that is not the real use scenario the model is aiming for. Its purpose is to be able to correctly classify heartbeats from patients whose ECG it has never been trained on before.

Therefore, the second sub-dataset is meant to check that. The pool of beats it pulls from is the exact same but, when creating it, instead of selecting certain beats from each patient, 4 different patients (2 symptomatic and 2 asymptomatic) were selected among the ones with less noisy and more typical signals, and their beats were not included in the training, being left to be used only for the test after the training is completed.

This is expected, with almost certainty, to reduce the accuracy of the model, as less beats ()))) are used in the training and their diversity is reduced. Further than that, the ones being removed are checked to be among the less noisy ones, which should further impact the model's performance.

Nevertheless, it only makes sense for the most reliable possible beats to be used in the testing, as to make sure the information deduced from there, being the test the most important part of the evaluation of a model, is as trustworthy as possible.

Table 5.4. *Patient data on sub-dataset 2 training and test*

Number of patients	Asymptomatic, Training	Symptomatic Training	Total Training	Asymptomatic Test	Symptomatic Test	Total Test
Females	7	2	9	1	1	2
Males	28	4	32	1	1	2
Total	35	6	**43**	2	2	**4**

Table 5.5. *Number and percentages of beats on training and test of sub-dataset 2*

	Total beats	Asymptomatic	% Asymptomatic	Symptomatic	% Symptomatic
Training Data	17273	14708	85%	2565	15%
Test Data	1723	863	50%	860	50%

5.3.5 Sub-dataset 3

Sub-dataset 3 aims to evaluate the models' performance when trained with a cleaner dataset, composed by beats that more closely resemble the typical ECG pattern and have less noise left. That, however, can only be achieved at the expense of the sub-database size, the reason why those less reliable beats were used in the first place.

The chosen method to do so was to eliminate every beat that included one or more biomarkers within a certain range on the total of that biomarker's values.

It was considered to discard the beats one of whose biomarkers' value belonged in the upper and lower 5% of their specific type of biomarker within the database. However, this left the database with fewer than 5000 beats.

Contrary to what was believed previously to the first attempt at performing this filtering of the data, the extreme biomarkers did not, in general, belong to the same beats. Many of the noisier beats had average signal width after the averaging, therefore not being outliers despite carrying little useful information.

Therefore, the same process was performed but removing beats with biomarkers whose values belonged in the upper and lower 1% of their specific type of biomarker. After this process, almost 11250 beats remained, which is still a severe decrease, but far from the previous one.

The processed described has been referred as 'trimming' in this project.

These resulting beats were distributed as the following tables indicate. The number of patients in total, test and validation remains the same, but a table describing those was added as well for the sake of convenience.

Additional statistics on this trimmed dataset, used both in sub-dataset 3 and the following sub-dataset 4, can be found in the Annex F.

Table 5.6. *Patient data on sub-dataset 3 training and test*

Number of patients	Asymptomatic, Training	Symptomatic Training	Total Training	Asymptomatic Test	Symptomatic Test	Total Test
Females	7	2	9	1	1	2
Males	28	4	32	1	1	2
Total	35	6	**41**	2	2	**4**

Table 5.7. *Number and percentages of beats on training and test of sub-dataset 3*

	Total beats	Asymptomatic	% Asymptomatic	Symptomatic	% Symptomatic
Training Data	10095	8898	88%	1197	12%
Test Data	1150	540	47%	610	53%

5.3.6 Sub-dataset 4

After two sub-datasets whose tests were made up of all the beats of 4 specific patients, each with specific characteristics, the sub-dataset 4 allowed to carefully assess the performance of the model when leaving all beats from 2 completely different patients to be used in the testing, and training the model with the rest of the data, while keeping on using only the beats that went through the cut in the previous section.

Doing so, additional information was obtained that allowed contrasting the performance when leaving out of the train dataset the 4 particular patients of sub-datasets 2 and 3, with the performance when 2 entirely different patients are left out and tested with.

The model itself was of course different, as the training dataset was not the same, but it nevertheless provided relevant information in order to determine whether or not the behavior of models trained with the previous sub-datasets was a product of the particularities of the patients selected to be separated.

A smaller test has downsides when trying to analyze this aspect, as a good fit would only allow to extract conclusions about these two particular patient's data. However, a poor performance with this sub-dataset would be a solid indicator that the model performs poorly overall.

Additionally, having only two patients reserved for the test, one symptomatic and one asymptomatic to check both results, and both males and not on the extreme ends of the ages of the patients, in order to reduce variables and test the model with the more common cases; allows for a very clear and direct analysis of the results. And, of course, the more data the model is trained on, the more accurate it can get to be.

Table 5.8. *Patient data on sub-dataset 4 training and test*

Number of patients	Asymptomatic, Training	Symptomatic Training	Total Training	Asymptomatic Test	Symptomatic Test	Total Test
Females	8	3	11	0	0	0
Males	28	4	32	1	1	2
Total	36	7	**43**	1	1	**2**

Table 5.9. *Number and percentages of beats on training and test of sub-dataset 4*

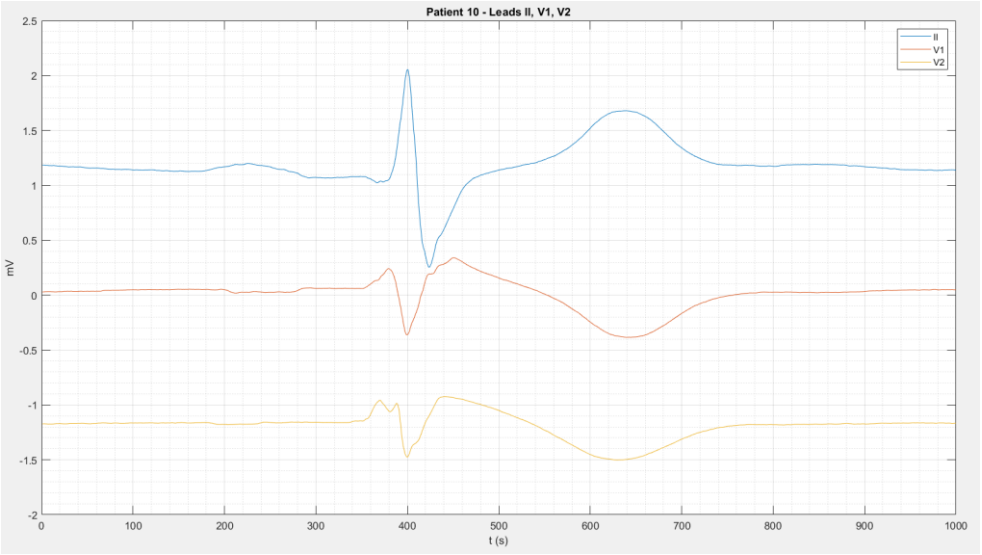
	Total beats	Asymptomatic	% Asymptomatic	Symptomatic	% Symptomatic
Training Data	10557	9133	87%	1424	13%
Test Data	687	304	44%	383	56%

5.3.7 Sub-dataset 5

This fifth and final sub-dataset aimed to be composed, both in training and testing, exclusively by beats from patients whose signals were generally stable, clear and therefore assumed to be more reliable, regardless of the size of the training or testing datasets.

After semi random inspection of multiple segments of every patient recorded at different moments in time, 4 out of the 8 symptomatic patients were evaluated to have remarkably clear signals with easily distinguishable segments, as well as a complete and reliable clinical dataset. Figure 5.4 exemplifies a beat that fulfills those characteristics and one that doesn't.

A



B

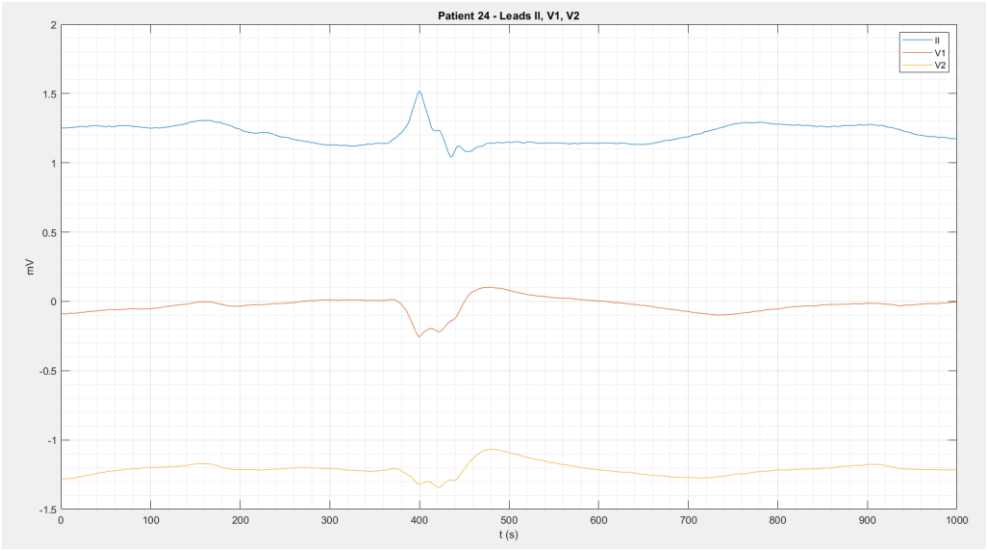


Figure 5.4. A: Example of a beat from a patient used in this sub-dataset. **B:** Example of a beat from a patient who was decided not to be used in this dataset.

In consequence, 12 additional asymptomatic patients were selected using the same criteria in order to complete the sub-dataset. More asymptomatic patients could have been added, with only slightly less clean signals, but there were already many more asymptomatic than symptomatic patients in the sub-dataset, and with the latter being the most relevant of the two, there were no further additions.

It was considered that at least 2 symptomatic patients had to be separated for testing so it could have a minimum diversity on the symptomatic side. Considering their correct classification as a priority, as well as how every model up until now had been trained on a much larger amount of asymptomatic than symptomatic beats, it was decided to attempt to train the models on a sub-dataset that was balanced in that proportion, finally selecting only two additional patients, both

asymptomatic in this case. The rest were used for testing, being the largest test dataset used up until that point.

Two of the patients' beats used in the test dataset, one patient being symptomatic and the other asymptomatic, were selected in order to test the resulting models with data that could be quickly and directly analyzed. This test dataset will be referred from now on as 'small test' and the previous, bigger one as 'large test', for the sake of simplicity.

This smaller scale dataset was not expected to return reliably accurate and precise predictions for a variety of patients, as its training dataset is simply too small to account for many scenarios. However, it was considered that valuable information could be extracted from its performance.

Table 5.10. *Patient data on sub-dataset 5 training and test*

Number of patients	Asymp, Training	Symp, Training	Total Training	Asymp, Large Test	Symp, Large Test	Total Large Test	Asymp, Small Test	Symp, Small Test	Total Small Test
Females	1	1	2	2	0	2	1	0	1
Males	1	1	2	8	2	10	0	1	1
Total	2	2	**4**	10	2	**12**	1	1	**2**

Table 5.11. *Number and percentages of beats on training and tests of sub-dataset 5*

	Total beats	Asymptomatic	% Asymptomatic	Symptomatic	% Symptomatic
Training Data	1724	864	50%	860	50%
Large Test Data	5121	4300	84%	821	16%
Small Test Data	853	433	51%	420	49%

5.4 Brief statistics on the different sub-datasets

Although statistical analysis of the obtained biomarkers is not a part of this project, the mean, median and percentiles 5 and 95 of the biomarkers were calculated from the different sub-datasets.

Sub-datasets 1 and 2, as well as 3 and 4, have each pair only a single group of data they pool from (all the beats and beats after trimming respectively). They are simply distributed differently in training and test sets of data, as explained previously. Therefore, their statistics are the same and were presented together.

The main goal of presenting these statistical data is to allow the reader a broad comprehension of the value ranges of the data used, if so desired, so they will not be discussed here.

5.4.1 Sub-datasets 1 and 2

Table 5.12. *Statistical data from average power and absolute value of the QRS area on sub-datasets 1 and 2*

	P_{avg} (mV^2)	P_{avg} (mV^2)	P_{avg} (mV^2)	$AbsArea_{QRS}$ II (mV·s)	$AbsArea_{QRS}$ V1 (mV·s)	$AbsArea_{QRS}$ V2 (mV·s)
Mean	0.285	0.076	0.066	36.697	21.603	20.351
Median	0.375	0.090	0.096	37.302	22.773	22.254
Perct. 0.05	0.009	0.015	0.009	10.21	9.524	7.165
Perct. 0.95	1.168	0.231	0.284	72.76	45.18	50.86

Number of patients: N=45. Number of beats: B=18999

Table 5.13. *Statistical data from QT, QTc and ST deviation at J-point (ST_0) on sub-datasets 1 and 2*

	QT II (s)	QT V1 (s)	QT V2 (s)	QTc II (ms)	QTc V1 (ms)	QTc V2 (ms)	ST_0 II (mV)	ST_0 V1 (mV)	ST_0 V2 (mV)
Mean	0.374	0.379	0.382	414.2	424.7	422.5	0.003	0.069	0.374
Median	0.337	0.329	0.324	386.2	380.0	367.3	-0.036	0.086	0.337
Perct. 0.95	0.328	0.333	0.316	372.9	375.5	370.0	-0.336	-0.045	0.328
Perct. 0.05	0.481	0.479	0.482	567.6	554.4	546.4	0.123	0.239	0.481

Number of patients: N=45. Number of beats: B=18999

Table 5.14. *Statistical data from ST deviation at J-point + 60 ms (ST_{60}) and ST slope from J-point to J-point + 60 ms on sub-datasets 1 and 2*

	ST_{60} II (mV)	ST_{60} V1 (mV)	ST_{60} V2 (mV)	ST_{slope} II (μV/ms)	ST_{slope} V1 (μV/ms)	ST_{slope} V2 (μV/ms)
Mean	0.071	0.007	0.035	1.553	-1.147	-0.813
Median	0.074	0.004	0.040	1.832	-1.353	-1.179
Perct. 0.05	-0.096	-0.143	-0.093	-1.145	-3.820	-5.357
Perct. 0.95	0.239	0.131	0.185	6.661	0.598	0.935

Number of patients: N=45. Number of beats: B=18999

5.4.2 Sub-datasets 3 and 4

Table 5.15. *Statistical data from average power and absolute value of the QRS area on sub-datasets 3 and 4*

	P_{avg} (mV^2)	P_{avg} (mV^2)	P_{avg} (mV^2)	$AbsArea_{QRS}$ II (mV·s)	$AbsArea_{QRS}$ V1 (mV·s)	$AbsArea_{QRS}$ V2 (mV·s)
Mean	0.412	0.097	0.089	38.952	23.806	22.114
Median	0.315	0.084	0.067	38.367	23.028	21.021
Perct. 0.05	0.021	0.019	0.010	12.968	10.672	7.987
Perct. 0.95	1.143	0.232	0.246	65.892	42.904	42.097

Number of patients: N=45. Number of beats: B=11244

Table 5.16. *Statistical data from QT, QTc and ST deviation at J-point (ST_0) on sub-datasets 3 and 4*

	QT II (s)	QT V1 (s)	QT V2 (s)	QTc II (ms)	QTc V1 (ms)	QTc V2 (ms)	ST_0 II (mV)	ST_0 V1 (mV)	ST_0 V2 (mV)
Mean	0.374	0.382	0.382	420.1	428.9	429.1	-0.035	0.080	0.112
Median	0.369	0.377	0.382	411.6	422.6	421.3	0.003	0.066	0.074
Perct. 0.05	0.330	0.336	0.324	373.6	377.9	374.5	-0.262	-0.038	-0.032
Perct. 0.95	0.441	0.455	0.444	484.9	499.9	511.2	0.099	0.230	0.405

Number of patients: N=45. Number of beats: B=11244

Table 5.17. *Statistical data from ST deviation at J-point + 60 ms (ST_60) and ST slope from J-point to J-point + 60 ms on sub-datasets 3 and 4*

	ST_{60} II (mV)	ST_{60} V1 (mV)	ST_{60} V2 (mV)	ST_{slope} II (µV/ms)	ST_{slope} V1 (µV/ms)	ST_{slope} V2 (µV/ms)
Mean	0.087	-0.007	0.032	2.032	-1.457	-1.344
Median	0.079	-0.004	0.025	1.673	-1.343	-0.891
Perct. 0.05	-0.016	-0.131	-0.086	0.122	-3.607	-5.287
Perct. 0.95	0.220	0.095	0.170	5.381	0.232	0.658

Number of patients: N=45. Number of beats: B=11244

5.4.3 Sub-dataset 5

Table 5.18. *Statistical data from average power and absolute value of the QRS area on sub-dataset 5*

	P_{avg} (mV^2)	P_{avg} (mV^2)	P_{avg} (mV^2)	$AbsArea_{QRS}$ II (mV·s)	$AbsArea_{QRS}$ V1 (mV·s)	$AbsArea_{QRS}$ V2 (mV·s)
Mean	0.554	0.120	0.097	40.61	24.54	21.00
Median	0.401	0.099	0.075	37.30	23.58	20.76
Perct. 0.05	0.102	0.028	0.008	16.19	9.04	5.997
Perct. 0.95	1.335	0.263	0.261	66.91	46.05	41.50

Number of patients: N=16. Number of beats: B=6844

Table 5.19. *Statistical data from QT, QTc and ST deviation at J-point (ST_0) on sub-datasets 5*

	QT II (s)	QT V1 (s)	QT V2 (s)	QTc II (ms)	QTc V1 (ms)	QTc V2 (ms)	ST_0 II (mV)	ST_0 V1 (mV)	ST_0 V2 (mV)
Mean	0.325	0.318	0.322	360.9	355.9	359.7	-0.083	0.064	0.099
Median	0.367	0.375	0.377	405.6	415.5	414.3	-0.056	0.055	0.057
Perct. 0.05	0.332	0.336	0.330	376.5	375.9	375.7	-0.500	-0.048	-0.036
Perct. 0.95	0.461	0.485	0.439	492.0	553.9	489.0	0.116	0.192	0.356

Number of patients: N=16. Number of beats: B=6844

Table 5.20. *Statistical data from ST deviation at J-point + 60 ms (ST_60) and ST slope from J-point to J-point + 60 ms on sub-datasets 5*

	ST_{60} II (mV)	ST_{60} V1 (mV)	ST_{60} V2 (mV)	ST_{slope} II (µV/ms)	ST_{slope} V1 (µV/ms)	ST_{slope} V2 (µV/ms)
Mean	0.047	0.002	0.041	2.171	-1.043	-0.958
Median	0.058	0.001	0.031	1.847	-1.014	-0.619
Perct. 0.05	-0.172	-0.140	-0.085	-1.598	-3.459	-5.220
Perct. 0.95	0.212	0.140	0.195	7.914	0.727	0.991

Number of patients: N=16. Number of beats: B=6844

6. Results

The following results are extracted from the accuracies depicted in the confusion matrixes the Classification Learner App presents its success and failure rates on, as can be seen in figure 6.1.

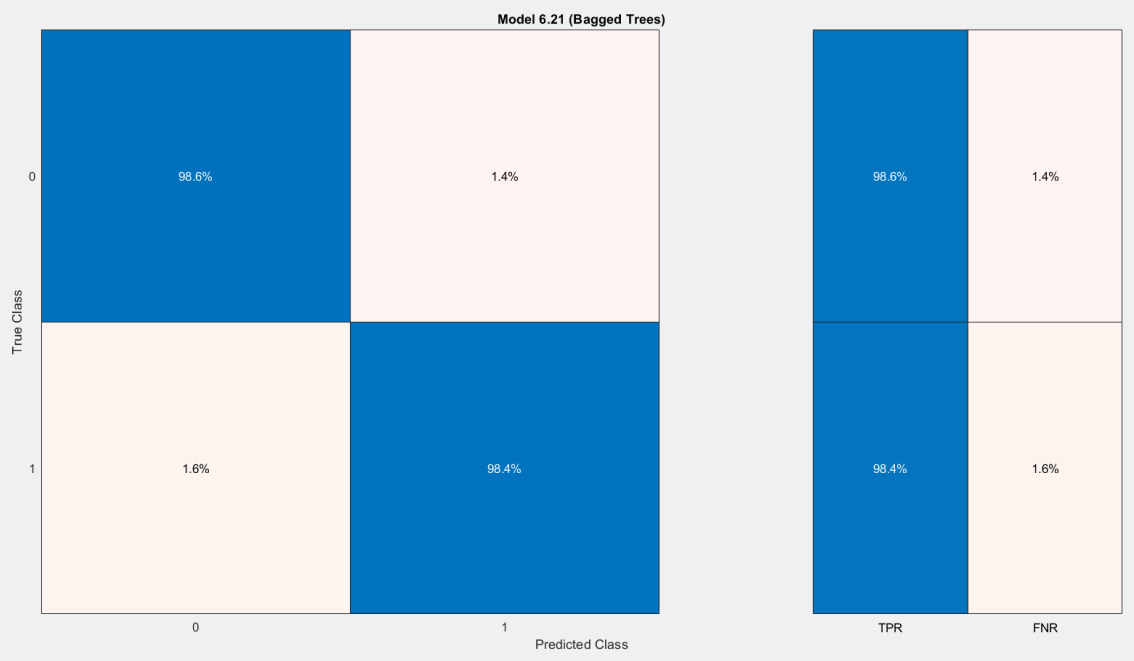


Figure 6.1. Confusion matrix example from the Classification Learner App

As the data is being classified into two possible categories, the confusion matrix (in the left table) is 2 by 2. As the legend indicates, vertical axis indicates the true class and the horizontal one the predicted class. Asymptomatic patients were assigned a 0 and symptomatic patients a 1 to indicate that same status.

Therefore, the top left value of that same table is the true positive rate for the asymptomatic beats, the top right value is the false negative rate for the asymptomatic beats, the bottom left value is the false negative rate for the symptomatic beats and the bottom right value is the true positive rate for the symptomatic beats. True positive rates and false negative rates are grouped in the left and right columns respectively in the table in the right, an alternative and potentially easier to read presentation.

The overall accuracies, not featured in the previous table, but easily available in the interface of the classifier app, see figure 4.3, are a result in percentage of the division of the totality of the correct guesses by the totality amount of data, be it on the training or test, depending on what it's referring

to. True and false positive rates are calculated following the same principle, applied to the particular aspect they represent.

In order to synthesize the information presented hereunder, only the overall accuracy of each model and both true positive rates will be displayed, as false negative rates would practically give redundant information and having the overall accuracy as well as both of the other true positive rates allows to approximately deduce the approximate weight of each type of beat in a given dataset without needing to search for that information.

One might notice by the previous paragraphs that the following sections are going to include a lot of repetition around terms such as true positive rate, overall accuracy, etc. In order to make reading (and writing) more agile, from this point on and most of the times, true positive rate was abbreviated as TPR (TP will also sometimes be used to refer simply to 'true positive'), ATPR when referred to the TPR of asymptomatic beats and STPR when referred to the same for symptomatic beats.

An additional, and final, letter is added to those two last acronyms in order to refer to each of them in the context of a validation or in a test, resulting in ATPRV and STPRV on one side, and ATPRT and STPRT in the other.

Finally, OVA refers to the overall validation accuracy, OTA refers to the overall test accuracy, and McW to misclassification weights. Although not an acronym, the process of training and testing machine learning models for each sub-dataset will be sometimes referred as iteration from now on.

The results of these processes can be reproduced approximately, however slight discrepancies in performance are most likely going to appear, as the random generated number based validation holdout selects different beats on different iterations.

The validation data were composed by a 25% of the training data in all iterations of the training, and was selected automatically by the Classification Learner App.

If not stated otherwise, the models' settings were the default ones.

The Classification Learner App allows training and testing models of all its available types with each dataset uploaded consecutively and through a single command. In order to maximize the final accuracies obtained, that was always done initially with each train and test dataset, although further modifications were often performed exclusively to the best performing models.

It is set as a priority in the project to put additional emphasis on trying to maximize the ATPR on the models. That is in order to, on an a faraway future application scenario, not miss on symptomatic patients over the possibility of a false positive for symptomatic leading to additional tests that would,

most likely, not lead to significant negative consequences to their health without additional reassurance of the necessity of such procedures.

McW were, in consequence, often modified to improve STPR, most often successfully. It is a usual course to take, especially when using training databases that over represent a part of its population, as in this case.

However, increasing the McW for symptomatic beats further than a certain point, usually 7:1 or 8:1, made performances consistently worse no matter how much the ratio was increased, often in all characteristics examined, although sometimes the STPR increased somewhat, while the ATPR decreased several times more. Whenever that's the case, in order to keep this section concise, the worse performing model's accuracies won't be shown, and it will mostly be commented on passingly (if so), as their behavior was very similar and is considered to be outlined here.

At times, the best performing model with a given set of misclassification costs is taken and its misclassification costs changed only to that one, assuming that, if the McW were changed to all models, this best performing one would remain the best performing. This is far from a certainty, but on subsequent attempts where McW were changed to all models in order to maximize accuracy, it has been observed to be a pretty reasonable assumption, happening most times and with small differences when that happens. That's especially true if the modified model is a Bagged Tree model, which was consistently among the models that improved the most when its McW were modified.

The models presented in the results will be the best performing ones in one of the characteristics considered as most important: overall validation accuracy (OVA), overall test accuracy (OTA), symptomatic true positive rate in validation (STPRV), symptomatic true positive rate in test (STPRT) or, occasionally, serve to illustrate a particular point about the models or performances.

The models were assigned numeric codes according to the sub section they are presented on, and were often referred as in the following format: 'MX.Y'. With M referring to model, and sometimes serving to omit the word, 'X' being the number of the sub section, and Y an additional number assigned in order of appearance.

6.1 Sub-dataset 1

Every accuracy rate turned out around the 98% on their best performing models, as can be seen in the table below.

Table 6.1. *Best overall accuracy and STPR for validation and test on sub-dataset 1 (right sub-columns) alongside their corresponding models (left sub-columns)*

	Validation		Test	
Best Overall Accuracy	97.9%	M1.2	98.0%	M1.4
Best True Positive Rate (Symptomatic)	98.6%	M1.3	98.4%	M1.3, M1.4

Table 6.2. *Model number, type, McW, OVA and OTA of each model discussed from sub-dataset 2*

Model Num.	Model Type	Misclassification Weights (Sympt.:Asympt.)	Validation Accuracy	Test Accuracy
1.1	Bagged Trees	1:1	95.6%	96.1%
1.2	Bagged Trees	4:1	**97.9%**	98.1%
1.3	Bagged Trees	5:1	97.7%	**98.6%**
1.4	Bagged Trees	6:1	96.9%	98.1%

The initial best performing model had very high, even if improvable OVA. However, observing only the overall accuracies gives only a limited view of the models. This model had lacking aspects that, as will be often the case when examining the rest of the models, can be better understood when examining ATPRs and STPRs, displayed in the following table.

Table 6.3. *Model number, ATPRV, STPRV, ATPRT and STPRT of each model discussed from sub-dataset 1*

Model Num.	TP Rate Asymptomatic Validation	TP Rate Symptomatic Validation	TP Rate Asymptomatic Test	TP Rate Symptomatic Test
1.1	99.9%	76.5%	100%	78.1%
1.2	99.4%	90.8%	99.3%	92.2%
1.3	97.8%	97.1%	98.6%	**98.4%**
1.4	96.7%	**98.0%**	98.0%	**98.4%**

As can be seen in this table, despite the high overall accuracy of M1.1, its STPR is significantly lower than its ATPR (76.5% to 99.9%). The comparatively high overall accuracy occurs due to the much larger number of symptomatic beats, under 20% of the total. In order to change that, in accordance to the objectives set for the classifier, its misclassification weights (McW) were modified.

Model 1.2, with McW of 4:1 (these will always be favoring symptomatic patients in the project), obtained the best OVA, and its STPRV was significantly higher, but the difference between STPR and ATPR was still quite large.

Model 1.3, with 5:1 McW, had slightly worse OVA, but a much better STPRV (97.1% compared to 90.8%). Model 1.4, with 6:1 McW, achieved an even STPRV than that, hitting the 98% mark, although of course that meant a drop in both ATPRV and OVA. Increasing McW to 7:1 or higher started to show diminishing returns.7:1, for example, resulted in a 97.7% STPRV, while still dropping on OVA.

As these values are pretty high already, possible further improvements by adjusting other parameters in the model (such as the maximum number of splits or the number of learners for these models) were delayed until further tests and the models were trained with differently arranged datasets, in order to understand them more.

Test results were very similar to validation results for each model. Model 1.3 had the best overall test accuracy, and M1.3 and M1.4 had the same STPRT, leaving the former as the clear best performing one with the test. Similarly, models that performed worse than these on validation also performed worse when being tested.

6.2 Sub-dataset 2

Table 6.4. *Best overall accuracy and STPR for validation and test on sub-dataset 2 (right sub-columns) alongside their corresponding models (left sub-columns)*

	Validation		Test	
Best Overall Accuracy	98.3%	M2.2	54.3%	M2.4
Best True Positive Rate (Symptomatic)	97.7%	M2.4	17.9%	M2.4

Accuracies on this dataset remain, on the best performing models, very high for the validation, over 95% both overall and on STPR. These results contrast with the test accuracies, with a maximum of 54.3% and 17.9% respectively.

Table 6.5. *Model number, type, McW, OVA and OTA of each model discussed from sub-dataset 2*

Model Num.	Model Type	Misclassification Weights (Sympt.:Asympt.)	Validation Accuracy	Test Accuracy
2.1	Subspace KNN	1:1	94.9%	51.0%
2.2	Bagged Trees	5:1	98.3%	45.5%
2.3	Bagged Trees	6:1	98.2%	45.3%
2.4	Bilayered Neural Network	-	77.1%	54.3%
2.5	RUS Boosted Trees	1:1	94.2%	52.8%
2.6	RUS Boosted Trees	6:1	93.2%	53.2%
2.7	Coarse KNN	1:1	88.5%	21.0%

Table 6.6. *Model number, ATPRV, STPRV, ATPRT and STPRT of each model discussed from sub-dataset 2*

Model Num.	TP Rate Asymptomatic Validation	TP Rate Symptomatic Validation	TP Rate Asymptomatic Test	TP Rate Symptomatic Test
2.1	99.6%	68.0%	98.6%	3.1%
2.2	99.3%	92.8%	90.5%	0.3%
2.3	98.6%	96.3%	90.3%	0.2%
2.4	73.5%	97.7%	90.6%	17.9%
2.5	96.6%	80.0%	88.2%	17.2%
2.6	95.5%	80.2%	88.8%	17.4%
2.7	94.3%	55.4%	31.4%	10.6%

Examining models individually, it can be seen that, in this case, the best performing model on overall accuracy on validation with 1:1 McW is of KNN type (M1.1), Subspace KNN specifically. It has an accuracy quite similar to its homonymous in the previous section, and a similarly lower STPRV.

Bagged Tree models were, again, the best performing ones when McW were modified. M2.2 has slightly a better OVA and M2.3 has a substantially higher STPRV.

Test results for these three models, however, were much lower. Their OTA were between 45 and 51%, but their STPRT did not go above 4%, not even above 1% in the case of M2.2 and M2.3.

The best performing model on the test, both overall and on STPR, was M2.4, a Bilayered Neural Network, with a 54.3% and a 17.9% respectively. These accuracies are still very low, so parameters on the models were modified in order to try to improve its performance, which could also benefit its STPRV, 73.5%, the reason why despite its 97.7% STPRV (the highest with this dataset) its OVA was among the lower ones, 77.1%.

However, every change in setting decreased the model's performance significantly. McW can't be changed in this type of models but, for example, when the model was trained with its first layer size reduced from 10 to 9, which could somewhat minimize overfitting, the STPRT dropped from 17.9% to 9.1%.

Other models were modified in order to try to maximize OTA and STPRT, but no significant results were achieved. M2.5, a RUS Boosted Trees model, one of the overall higher performing models on test before these changes and the best performing one among decision trees, could only have its OTA increased from 52.8% to 53.2% and it STPRT from 17.2% to 17.4% (see M2.6). Those improvements were achieved, again, through changing its McW. Changes oriented to reduce overfitting had diminishing returns.

M2.7 had the lowest OTA, with a 21.0%. It must be noted that, despite its OVA being 88.5% (lower, but not too far away from the best performing ones), its STPRV was 55.4%.

6.3 Sub-dataset 3

Table 6.7. *Best overall accuracy and STPR for validation and test on sub-dataset 3 (right sub-columns) alongside their corresponding models (left sub-columns)*

	Validation		Test	
Best Overall Accuracy	99.6%	M3.2	53.7%	M3.3
Best True Positive Rate (Symptomatic)	98.3%	M3.2	29.5%	M3.4

Compared to the last iteration, the highest accuracies for validation slightly increase, while the highest OTA stays practically equal and STPRT, despite increasing significantly, remains very low.

Table 6.8. *Model number, type, McW, OVA and OTA of each model discussed from sub-dataset 3*

Model Num.	Model Type	Misclassification Weights (Sympt.:Asympt.)	Validation Accuracy	Test Accuracy
3.1	Bagged Trees	1:1	99.5%	45.1%
3.2	Bagged Trees	7:1	99.6%	41.9%
3.3	Kernel Naive Bayes	9:1	90.8%	53.7%
3.4	Gaussian Naive Bayes	6:1	76.3%	47.3%
3.5	Cubic SVM	6:1	99.0%	14.8%

Table 6.9. *Model number, ATPRV, STPRV, ATPRT and STPRT of each model discussed from sub-dataset 3*

Model Num.	TP Rate Asymptomatic Validation	TP Rate Symptomatic Validation	TP Rate Asymptomatic Test	TP Rate Symptomatic Test
3.1	99.8%	97.3%	96.1%	0.0%
3.2	99.8%	98.3%	89.2%	0.2%
3.3	91.4%	86.6%	87.0%	24.3%
3.4	77.2%	69.9%	67.3%	29.5%
3.5	99.2%	97.7%	20.8%	9.5%

In this iteration, Bagged Trees were again the best performing models on the validation, as M3.1 and M3.2 show. The adjustment of McW only brought forth a 1% improvement in STPRV and a 0.1% improvement in OVA, but the original rates were already close to the highest in all models.

The OTA and STPRT of these models are similar to the best performing models on validation of the previous dataset, with an even slightly lower STPRT, being 0.0% and 0.3% on M3.1 and M3.2 respectively.

After changing McW in order to maximize this parameter, the classifier with the best OTA, M3.3, was a Kernel Naive Bayes model and had a 53.7% OTA, a smaller value than the best from the previous dataset only by a 0.6%.

Its STPRT is quite higher than the best ones in the previous iteration though, with a 24.3% against a 17.9%, both being very low still. When trying to maximize this parameter, the best performing model created was M3.4, a Gaussian Naive Bayes classifier with modified McW and a 29.5% STPRT. It may be worth pointing out that its OVA is by far the lowest among the selected models though, a 76.3%.

The model that achieved the lowest OTA was a Cubic SVM one, with only a 14.8%, and yet a 99.0% OVA. STPRT was higher than the ones from M3.1 and M3.2, with a 9.5%.

6.4 Sub-dataset 4

Table 6.10. *Best overall accuracy and STPR for validation and test on sub-dataset 4 (right sub-columns) alongside their corresponding models (left sub-columns)*

	Validation		Test	
Best Overall Accuracy	99.5%	M4.1	54.7%	M4.2
Best True Positive Rate (Symptomatic)	98.7%	M4.1	20.6%	M4.2

All the best rates are close to the same as in the previous sub-dataset, except for the STPRT, which is around a 9% lower.

Table 6.11. *Model number, type, McW, OVA and OTA of each model discussed from sub-dataset 4*

Model Num.	Model Type	Misclassification Weights (Sympt.:Asympt.)	Validation Accuracy	Test Accuracy
4.1	Fine KNN	1:1	**99.5%**	26.6%
4.2	Subspace KNN	1:1	97.6%	**54.7%**
4.3	Fine Gaussian SVM	1:1	98.5%	44.3%
4.4	Wide Neural Network	1:1	99.4%	1.9%

Table 6.12. *Model number, ATPRV, STPRV, ATPRT and STPRT of each model discussed from sub-dataset 4*

Model Num.	TP Rate Asymptomatic Validation	TP Rate Symptomatic Validation	TP Rate Asymptomatic Test	TP Rate Symptomatic Test
4.1	99.7%	**98.7%**	60.2%	0.0%
4.2	99.9%	83.3%	97.7%	**20.6%**
4.3	100.0%	88.6%	100.0%	0.0%
4.4	99.6%	98.1%	4.3%	0.0%

The model with the best OVA (99.5%) and STPRV (98.7%) was the same one in this iteration, M4.1, a Fine KNN type with 1:1 McW. McW were modified in this and the rest of the models in order to check if improvements on their, already pretty high, validation performances could be made, but no improvement was achieved.

A particularity of this iteration is that it's the only one where none of the best performing models are Bagged Trees, and in fact none of the models are decision trees. Nevertheless, it should be noted that Bagged Trees models were among the best performing on validation when weights were changed, as in the rest of the iterations, just not the absolute best.

The best performances on both OVT and STPRT, 54.7% and 20.6% respectively, were also achieved in the same model, M4.2; also a KNN classifier, but using the Subspace KNN method in this case.

M4.1 had a 26.6% OTA, the lowest among top-performing models on validation, with an also low 60.2% ATPRT compared to those, and a 0.0% STPRT, the only test performance that was not unusual.

Models 4.3 and 4.4 display the most extreme performances found among models with a 0.0% STPRT, whose ATPRT went from M4.3's 100% to M4.4's 4.3%. M4.4, therefore, guessed almost the exact opposite to the desired prediction.

Despite these results on the test, M4.3 has a 98.5% OVA and a 88.6% STPRV, and M4.4 has a 99.4% OVA and a 98.1% STPRV, close to this iteration's best one (98.7%).

6.5 Sub-dataset 5

Table 6.13. *Best overall accuracy and STPR for validation, small and large tests on sub-dataset 5 (right sub-columns) alongside their corresponding models (left sub-columns)*

	Validation		Large Test		Small Test	
Best Overall Accuracy	100%	M5.1	75.5%	M5.2	100%	M5.1
Best True Positive Rate (Symptomatic)	100%	M5.1	58.7%	M5.3	100%	M5.1, M5.3

Despite not being much higher than in previous iterations, as those were close to perfect as well, validation accuracies, both overall and STPR, reach 100% in this iteration, although it

must be noted that the validation holdout for this sub-dataset is smaller than in previous ones.

The small test also hits 100% accuracies for both of those rates, but the best performances for the large test remain far from those numbers, despite 75.5% being the highest any OTA has been on sub-datasets that use test datasets composed by patient's beats that have not been used in training (and the same happening with the 58.7% STPRT).

Table 6.14. *Model number, type, McW, OVA, large test overall accuracy and small test overall accuracy of each model discussed from sub-dataset 5*

Model Num.	Model Type	Misclassification Weights (S.:As.)	Validation Accuracy	Large test Accuracy	Small test Accuracy
5.1	Bagged Trees	1:1	**100%**	52.8%	**100%**
5.2	Linear Discriminant	7:1	88.1%	**75.5%**	83.7%
5.3	Coarse Tree	1:1	98.4%	40.3%	76.2%

Table 6.15. *Model number, ATPRV, STPRV, ATPR and STPR for the large test and ATPR and STPR for the small test of each model discussed from sub-dataset 5*

Model Num.	TP Rate Asympt. Validation	TP Rate Sympt. Validation	TP Rate Asympt. Large Test	TP Rate Sympt. Large Test	TP Rate Asympt. Small Test	TP Rate Sympt. Small Test
5.1	100%	**100%**	53.0%	51.3%	100%	**100%**
5.2	99.5%	76.7%	82.6%	38.5%	97.9%	69.0%
5.3	98.6%	98.1%	24.8%	**58.7%**	53.1%	**100%**

Model 5.1, a Bagged Trees classifier, achieved a 100% OVA, STPRV, OTA and STPR for the small test, all without requiring any modifications. Its McW were not changed from the by default 1:1. Additionally, these results were achieved consistently on repeated trainings.

However, once again, the large test showed much worse results, with an overall accuracy of 52.8% and a STPR of 51.3%.

It should be noted, nonetheless, that the small test was roughly equal in size to the sub-dataset 4 one, in fact it shared one of its two patients with it, and yet its best OTA was 54.7%.

As with many previous models, it was attempted to modify the numbers of learners in M5.1 in order to see if the performance improved as a consequence. The attempt, however, was not successful, as severe decreases were required to provoke minor differences in the model, and performance mostly tended to worsen overall.

Other models, most of them classification trees, achieved similarly outstanding performances in the same parameters as M5.1, but also similarly lower ones in the rest.

After changing parameters in multiple models of different types, a Linear Discriminant model (M5.2) achieved the best overall accuracy on the large test, a 75.5%. Its OVA, though, was significantly lower than that of M5.1, 88.1%, and its STPR was only 38.5% on this same test.

In that regard, the best STPR was achieved by a Coarse Tree classifier (M5.3), with a 58.7%.

It's also the first instance since the test is performed with patients completely outside of the train dataset where a model has a higher tendency to guess a beat as pertaining to a symptomatic than an asymptomatic patient, with the small test having a 100% STPR and a 53.1% ATPR and an equal amount of beats (although other models in this iteration behaved similarly, M5.3 is just an example).

7. Discussion

7.1 General Results

Overall, results were consistently strong on validations: the best OVA were always between 98.5% and 100% on each sub-dataset, and the best STPRV between 97.7% and 100%.

Crucially though, results were mostly terrible for tests composed of beats from patients the entirety of whose beats were used in the test only, those of sub-datasets 2-5.

The best OTA were between 54.7% and 53.7% in sub-datasets 2-4, barely performing better than throwing a coin to do the guesses would. STPRT were the truly appalling results however, with the best falling between 17.9% and 29.5%.

Only models from sub-dataset 5 showed results above that, M5.1 in particular obtained an OVA for the small test of a 100%. The other two models, M5.2 and M5.3, appeared to have performed much better than usual as well, with OVA of 83.7% and 76.2%.

However, their results on the large test of their sub-dataset, much more reliable than the small one due to its size, were around the same as the rest: 52.8%, 75.5% and 40.3%, in numerical order. M5.2 is the best test performance overall, yet that must be put in context with its 38.5% STPR in this same test, obtaining a better overall result only due to its tendency to classify beats more often as symptomatic and the overrepresentation of those in the test dataset.

Only sub-dataset 1 is left to mention in regards to test performance. Its test was made of seven beats selected from each of the 45 patients in the same positions, with the rest of the beats being used to train the dataset. Its performance was very good on the test, and that information can be useful to analyze the model's behavior on different scenarios, but it does little to influence the general conclusions of the general efficacy of a model oriented to be able to make predictions on data that the model has not been trained on.

From all of this, it can be concluded that the models created were not able to reliably classify ECGs as belonging to a symptomatic or asymptomatic Brugada Syndrome patient with the resources here used and described.

7.2 Analysis of the classification criteria

Despite the models not being able to classify beats as symptomatic and asymptomatic reliably, results seem to indicate that models, at least the best performing ones on validation, and possibly others too; are finding specific patterns in specific patient's beats (and maybe groups of them at

times), common for some of the validation symptomatic or asymptomatic beats, and classifying the test beats according to which of those patterns they more closely adhere to.

These patterns, however, would not be common to all, or even most, symptomatic or asymptomatic patients, nor directly correlated with these groups in any meaningful way, hence the overall accuracies of tests being usually somewhere close to 50%. Again, they seem to be mostly patterns found in particular patients beats, be they common or from specific segments of the signal.

The higher ATPR in tests would also fit this explanation. Training a model with more data of one specific feature is likely to make it skew towards guessing that feature more often, and this would be the explanation for that in this case, both for the effect and its proportions.

In a model trained with more asymptomatic patients, if a common pattern (or similar amount of sets of patterns) is not found for symptomatic patients on one side and asymptomatic patients on another, and instead the models find patterns common on individual patients or small groups of them that are unrelated to the two explored conditions, it is to be expected for more beats to be classified as asymptomatic.

Such a model, given those previous assumptions, has a larger set of patterns associated with the asymptomatic label, so beats will be therefore more likely to resemble one of them.

Again, this is a common tendency on machine learning when training datasets are unbalanced, but it is important to analyze how such tendencies occur in each instance in order to create better performing ones in the future and understand their less apparent mechanisms, even when they achieve the proposed goals.

Results from models trained and tested on sub-dataset 1 are also coherent with this assessment. Their high performance would be explained by the fact that, in this case, the patterns found on particular patient's beats are not being applied on beats from new patients, independent to the model up to that point, and with their own individual particularities unrelated to BrS symptoms; but to beats whose patterns the model has already examined.

They, of course, more closely resemble patterns pertaining to the patients they're from, and therefore get correctly classified as symptomatic or asymptomatic, just as their patients were on the training.

Finally, the results of the sub-dataset 5 models also point to the specified behavior of the model. The 100% accuracy of the small test is an outlier, as no model had achieved such a strong accuracy before, even though the one from sub-dataset 4 was even smaller (687 beats versus 853), and the explanation for that doesn't seem to be that it's correctly identifying patterns

However, that would make perfect sense if the model was classifying beats according to patient specific patterns. This model was trained only with 4 patients, selected for being 2 symptomatic and 2 asymptomatic and having very clear signals.

It is not unlikely for the beats of the two patients used in the test to, by chance, be more similar to one or more patients who just also happened to be symptomatic or asymptomatic.

That was not the case on sub-dataset 4 despite testing only with 2 patients as well, though. Patients were not classified uniformly there. A possible explanation for that is simply that, having been trained with 43 different patients, there were many more patient or section specific patterns for them to be considered similar to, thus not having any instance where accuracy happened to be really high and mostly evening out to 50% instead.

There was model 4.4 though, which did have an impressively low test accuracy while having a high validation accuracy, which could mean that the exact opposite to model 5.1 could have happened. Neural networks did not performed well most often however, so its results are not too reliable.

It must be addressed, to conclude this section, why only the best performing models on validation have been meaningfully discussed. The best performing models on both validation and tests, alongside a few, often outlier ones, were showcased in the results in order to give an accurate and minimally broad perspective on the results of the training of the dozens of available models.

However, when training dozens of models that classify data into a binary category under a wrong criteria that gives usually around a 50% accuracy on tests, and a much lower one to one of its options because of its bias, it is likely for some model that is simply not performing the task properly at any level at all to perform better than others that at least do on validation just because of sheer probability.

7.3 Attempts at avoiding this issue

Trying to improve the performance of the models of the test had no positive. Good validation results alongside poor test results often happen because of overfitting, which is similar to the hypothesized reason for the poor results on the test, although not knowing for certain the attempted classification is possible gives additional nuance to the matter.

The same classification tendencies, however, did not seem to change when measures to avoid overfitting were taken, as reducing the number of learners or splits in tree models. Performances generally worsened alongside those decreases, but rates stayed pretty similar. Changes on opposite directions were attempted briefly, as they would likely be counterproductive, to similar effect.

These changes in settings in order to prove the models were far from exhaustive, however. This will be expanded on when discussing future work but, despite not estimating this to be the major driving factor for the ultimate ineffectiveness of the models, but further investigation could benefit significantly from an approach geared towards a deeper control of the features of few models estimated to be able to give good results when given sufficient and adequate data.

That rings especially true when considering taking measures against overfitting while creating models that account for the many differences that necessarily appear between different groups of people

within both the symptomatic and asymptomatic populations, even on most optimal conditions of signal registering.

7.4 Data quality and quantity

There is a tension between having a big amount of data and that data being of the best possible quality, under criteria determined in each instance, at least most often. A balance must be found between the two. However, it is hard to evaluate that when the proposed goal has not been achieved or meaningfully approached, when concerning results.

Performances seemed to be slightly better on sub-datasets with more refined data, and models of more types reached high validation accuracies, but that can hardly be considered a reliable source of information.

That said, in case the beats were indeed classified following patient specific patterns, minimizing differences in their signals arisen from non-physiological factors should likely be a priority (noise, differences in behavior during the recording, used equipment…), as well as accounting for those that are unavoidable, such as population differences (regarding many factors, from which conditions they might have suffered previously to the how frequently they exercise).

7.5 Limitations

There are several limitations to be taken into account when examining this project. First, is the use of only 3 of the 12 leads of the ECG recordings, chosen because of time restrictions added to the potential additional difficulties of processing and manipulating 4 times as much information.

Also, there is the proposed used by the team in Hospital Clinic of the entire 24 hour runtime of the signals. That should, in principle, not be much of an impediment, as a big number of information is usually positive for machine learning. And, while the approach is reasonable and worth exploring, and no alternatives to this approach would've probably been taken regardless of that indication, that meant limiting the control over the balance between quantity and quality of information, as it implied avoiding the elimination of noisy segments of the signal.

As briefly stated right before, there are of course time constraints. Machine learning takes time, both for preparing the data it processes and for iterating its models.

Finally, and crucially, the main limitation is, as usual with machine learning, the amount and quality of the data used. From the initial 53 patients, only 45 could be used because of several reasons, and among those, around 20 were later informally classified later as being suboptimal recordings. Some studies have situated realistic number of patients necessary for ECG machine learning applications far from the forties and in the order of thousands [53]. The amount of data was initially conceived by the amount of time available. That seemed indeed significant, in the upper hundreds of hours. However, if the amount of data is considered by patients, the amount is much smaller.

Despite the previous mention of thousands of patients, an extreme estimate, it's likely that the main goal would be achievable with many less patients. A recent study, published during the course of this project, has seemingly been able to determine if patients were going to suffer (in a controlled period of time, which was later checked) or had suffered potentially fatal arrhythmias in the past with substantial performance [54].

They did so through deep learning techniques, specifically a convolutional neural network, concluding that subtle patterns, undetectable by humans, had been detected by the AI to do so. It did so for that single event, drawing from 2053 much smaller (10 s) ECG recordings, but supervised by professionals on a clinical, controlled environment and from 157 different patients.

Keeping in mind that the most fundamental aspect for the data quantity seems to be the number of patients, the previous study helps to put into perspective the limitations of this project and its difficulty.

However, it also seems to point, with quite certainty, towards the possibility of the detection of specific patterns for particular BrS related symptoms, as has been the goal in this project despite their grouping into a single 'symptomatic' category; as well as to give many indications on how to go forward with such a task.

8. Environmental study

This project has a minimal environmental impact, as only a set of computers, mainly the author's, the hardware the data was stored on, and a set of Holter devices, if those are even to be accounted for, as the data was generally long since collected and used for multiple different purposes.

It can be noted that the use of hardware to store the used data, despite implemented for the sake of convenience, meant a lesser energy use than cloud storing would have, and would be advisable if that was a priority.

In order to provide some calculations on the matter, the carbon emissions associated to the energy consumption during this project will be roughly estimated.

Laptop average wattage (W)	Total hours worked + hours of processing of data (h)	Emission equivalent of energy consumption in Spain in April of 2022 (g CO_{2eq}/kWh)	TOTAL (kg CO_{2eq})
65	640	259 g CO_{2eq}/kWh	10.77

The result is around 10,8kg of CO_2 equivalent. To put that into perspective, a private jet emits around 2 metric tons of CO_2 an hour [55].

Aside from that, any rigorous environmental analysis, however, should consider the place of its specific subject on the larger whole of the environment, how that influences it and, more importantly, the state of the former and the utility of the present analysis in relation to the state of the totality examined.

That is, of course, a monumental task that is not possible to do so under the scope of a section of a final degree thesis, pushing the analysis to a simpler, likely more isolated plane, likely concerning the project in hand alone, which renders it unfortunately pointless.

Isolated differences are not going to cause meaningful changes concerning the curate ecological crisis, or mostly any matter at all really. This is an issue that must be acted upon with a systematical approach, as many experts on the field have been claiming for a long time [56]. Without changes in the current economic structure, any small change is mostly inconsequential.

Therefore, efforts would be better spent discussing those matters and how to achieve them. Hereunder is a brief listing of some measures that experts that share this vision estimate would seriously begin to tackle the situation and prevent the terrible consequences that may arise due to climate collapse and the economic crisis that will entail [56]:

- Annulation of current debts.
- Radical reform of the financial system in order to separate it from private enterprise, charging interests and an economic management oriented towards growth.
- Redefinition of money to more localized forms.
- Changing the nature of states into more localized and truly democratic structures.
- Definition of localized, specific plans for the adjustment to the worsening of conditions, while collaboration between localities is kept.
- Orienting economic and social policies to the preservation of basic services.

Conclusions

The initial hypothesis, stating that it would be possible to prognosticate symptomatic patients of Brugada Syndrome through machine learning trained with biomarkers extracted from electrocardiograms and clinical data, was not demonstrated.

When tested with beats from patients not present in the training group, the best performing ML models developed seemed to classify the test beats according to their adherence to patterns specific to individual patients from the training phase, but not common to all (or most) symptomatic or asymptomatic patients.

In other words, tests beats were likely classified depending on which training signals they resembled the most in aspects unrelated to symptomatic or asymptomatic ECG patterns.

In any case, results were very poor without exception when trained with beats from patients not used on the training, an implied part of their established goal.

The fact that the ML models seemed to be able to distinguish specific patterns in individual patients indicates that the ECG processing was effective and that the selected biomarkers contained useful and detailed clinical information that allows classifying their beats consistently.

Algorithms that allowed a successive processing of multiple files for all of those aspects were successfully developed, and would be easy to modify for different, future applications.

This, however, does not imply that proving the initial hypothesis is an impossibility, and in fact some cited research seems to point towards a likely feasibility. Many aspects could be altered in the steps here taken to process the ECG, extract data from them and then apply the machine learning that could allow for that premise to be realized, and some suggestions of possible alterations will be made hereunder.

Future Work

Future exploration of the general premise and goals of this project could probably benefit substantially from the groundwork laid here, especially if the basic structure of the processing of the signals and biomarker extraction is kept.

However, it is fundamental to keep an open mind when considering data treatment and processing in machine learning. In that line, a very important consideration for future development is the use of

entire signals instead of biomarkers to train the models, as that could potentially allow for the detection of differences not considered or registered by the comparatively more limited and biased information that biomarkers can provide. The length of those signals would also be a key aspect to consider when applying that approach, as different longer term patterns may appear and be relevant for the classification.

That would probably also imply the use of different machine learning tools, but that could also be hugely beneficial. Applying machine learning with a very in-depth control of the inner workings of the models, if in big part not done here mainly due to scope limitations, could be a rewarding path forward to avoid the specific issues arisen in this initial approach. Deep learning in particular could be promising if used with clear understanding of its inner workings, as positive results have been shown in cited works and neural networks could potentially be useful to avoid focusing on patient-specific patterns for classification.

Perhaps the biggest concern going forward, however, is the number of available patients, and the quantity and quality of the data. As mentioned previously, similar studies place an optimal number of patients for machine learning tasks, such as this one, in the higher hundreds at least and with . Also, crucially, have used ECG recordings obtained in similar conditions and very controlled environments, despite the consequent decrease in the amount of data of each single patient.

That is, of course, the hardest aspect to change in this project, but it's still worth pointing out. On one side, as there are inherent limitations to the machine learning capabilities that can be developed from each set of data. But also because an aspect that can be more easily changed is the approach to those data. Evidence seems to point towards high quality, noiseless and controlled recordings being a priority over their quantity of data for each individual patient. That can be achieved on the recording stage, but can also be strived towards on the processing of the data.

A final aspect to point towards in regards to future development is the reexamination of the core premise, at least in its consequent application in programming and machine learning. The 'symptomatic' label was attributed in this project to every patient having suffered one or more of three different conditions, and classification was attempted to directly tell them apart, as a group, from patients who had not showed any of those symptoms at all.

However, it might be more effective, and results have been obtained using this approach in other studies, to focus machine learning algorithms on finding patterns from specific pathologies. That could require a longer developing time, and it would mean an additional shrinking of the available data for each symptomatic pattern explored, but it is worth considering nevertheless, and a successful development of models that can classify each symptom in particular would allow achieving the proposed goal as well.

Economic analysis

Labor

In total, 4 different workers participated in this project: 3 biomedical engineers and 1 cardiologist. The author, for simplification, will be assumed to be a fully-fledged engineer, with an according title, but without any previous experience.

The estimate of the hours of work has been estimated with input from the parts workers themselves, including preparing files, meetings, text and code revisions, etc.

Their hour salaries were estimated from their role in the project and the information extracted from []. It must be noted that this calculation assumes average salaries for the according positions on the project, which would likely not be the case, as the presence of a cardiologist in an engineering project like this one would likely happen only as a consultancy, with a much higher pay per hour. That won't be taken into account, but that means this will be a low estimate for the labor budget.

	Profession	Hours (h)	Salary (€/h)	TOTAL (€)
Author	Biomedical Engineer	600	13.20	7 920
Supervisor	Biomedical Engineer	80	19.20	1 536
Co-Supervisor	Biomedical Engineer	40	14.60	584
Co-Supervisor	Cardiologist	40	32.10	1 284
TOTAL	-	760		4 688€

Equipment

Next, the equipment costs are estimated. This is a strongly rounded up estimate as, for example, the laptop would never be used exclusively for this project, and a less expensive model would have been roughly as effective for this project in particular.

Item	Model/Details	Price (€)
Laptop	HP Pavilion Laptop 15-cs1xxx	780
Portable Solid Disk	Samsung Portable SSD T5	140
Matlab license	Matlab annual comercial license	860
TOTAL	-	1 780 €

The cost of the energy used would be around 10€, so it is considered negligible.

Total

Adding those two results, we obtain:

$$\boldsymbol{TOTAL\ BUDGET} = 4688€ + 1780€ = \mathbf{6468}€$$

Reference

[1] Weinhaus, A.J., Roberts, K.P. (2005). Anatomy of the Human Heart. In: Iaizzo, P.A. (eds) Handbook of Cardiac Anatomy, Physiology, and Devices. Humana Press. https://doi.org/10.1007/978-1-59259-835-9_4

[2] F. Shaffer, R. McCraty, and C. L. Zerr, "A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability," Frontiers in Psychology, vol. 5, 2014, Accessed: Dec. 27, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01040

[3] "Dundee - Drawing Cross-Section of Heart - English labels" by Shereen Kadir, © University of Dundee School of Medicine, license: CC BY-NC-ND. AnatomyTOOL. https://anatomytool.org/content/dundee-drawing-cross-section-heart-english-labels (accessed Dec. 26, 2022).

[4] "LadyofHats - Drawing Bloodflow of the heart during diastole and systole - English labels" by Mariana Ruiz Villarreal (Wikimedia: LadyofHats). Public Domain | AnatomyTOOL." https://anatomytool.org/content/ladyofhats-drawing-bloodflow-heart-during-diastole-and-systole-english-labels (accessed Dec. 26, 2022).

[5] R. J. Martis, U. R. Acharya, y H. Adeli, «Current methods in electrocardiogram characterization», Computers in Biology and Medicine, vol. 48, pp. 133-149, may 2014, doi: 10.1016/j.compbiomed.2014.02.012.

[6] A. Bayés de Luna, M. Fiol-Sala, y E. Antman, The 12 Lead ECG in ST Elevation Myocardial Infarction: a practical approach for clinicians. USA: Blackwell Futura, 2007.

[7] B. Benito, R. Brugada, J. Brugada, and P. Brugada, "Brugada Syndrome," Progress in Cardiovascular Diseases, vol. 51, no. 1, pp. 1–22, Jul. 2008, doi: 10.1016/j.pcad.2008.05.002.

[8] C. by Agateller, Schematic diagram of normal sinus rhythm for a human heart as seen on ECG (with English labels). 2007. Accessed: Jan. 02, 2023. [Online]. Available: https://commons.wikimedia.org/wiki/File:SinusRhythmLabels.svg

[9] Azcona, Luis. El electrocardiograma. Libro de la salud cardiovascular del Hospital Clínico San Carlos y la fundación BBVA. Directores: López Farré A, Macaya Miguel C. Bilbao: Fundación BBVA, 2009, p. 49-56.

[10] M. Sampson, “Understanding the ECG. Part 8: myocardial ischaemia and infarction (part A),” British Journal of Cardiac Nursing, vol. 11, pp. 290–298, Jun. 2016, doi: 10.12968/bjca.2016.11.6.290.

[11] K. Strimbu and J. A. Tavel, “What are Biomarkers?,” Curr Opin HIV AIDS, vol. 5, no. 6, pp. 463–466, Nov. 2010, doi: 10.1097/COH.0b013e32833ed177.

[12] World Health Organization and I. P. on C. Safety, “Biomarkers and risk assessment : concepts and principles,” World Health Organization, 1993. Accessed: Dec. 10, 2022. [Online]. Available: https://apps.who.int/iris/handle/10665/39037

[13] L. Oikarinen et al., “Electrocardiographic assessment of left ventricular hypertrophy with time - Voltage QRS and QRST-wave areas,” Journal of human hypertension, vol. 18, pp. 33–40, Feb. 2004, doi: 10.1038/sj.jhh.1001631.

[14] E. Burns, M. Cadogan, and E. B. and M. Cadogan, “Myocardial Ischaemia,” Life in the Fast Lane • LITFL, Aug. 01, 2020. https://litfl.com/myocardial-ischaemia-ecg-library/ (accessed Jan. 04, 2023).

[15] J. Molnar, J. S. Weiss, and J. E. Rosenthal, “The missing second: what is the correct unit for the Bazett corrected QT interval?,” Am J Cardiol, vol. 75, no. 7, pp. 537–538, Mar. 1995, doi: 10.1016/s0002-9149(99)80603-1.

[16] P. G. Postema and A. A. M. Wilde, “The measurement of the QT interval,” Curr Cardiol Rev, vol. 10, no. 3, pp. 287–294, Aug. 2014, doi: 10.2174/1573403x10666140514103612.

[17] J. Brugada, O. Campuzano, E. Arbelo, G. Sarquella-Brugada, and R. Brugada, “Present Status of Brugada Syndrome: JACC State-of-the-Art Review,” J Am Coll Cardiol, vol. 72, no. 9, pp. 1046–1059, Aug. 2018, doi: 10.1016/j.jacc.2018.06.037.

[18] X.-Q. Quan, S. Li, R. Liu, K. Zheng, X.-F. Wu, and Q. Tang, “A meta-analytic review of prevalence for Brugada ECG patterns and the risk for death,” Medicine (Baltimore), vol. 95, no. 50, p. e5643, Dec. 2016, doi: 10.1097/MD.0000000000005643.

[19] A. Oliva et al., “Structural Heart Alterations in Brugada Syndrome: Is it Really a Channelopathy? A Systematic Review,” Journal of Clinical Medicine, vol. 11, no. 15, Art. no. 15, Jan. 2022, doi: 10.3390/jcm11154406.

[20] P. Brugada and J. Brugada, “Right bundle branch block, persistent ST segment elevation and sudden cardiac death: A distinct clinical and electrocardiographic syndrome: A multicenter report,” Journal of the American College of Cardiology, vol. 20, no. 6, pp. 1391–1396, Nov. 1992, doi: 10.1016/0735-1097(92)90253-J.

[21] E. M. J. Marsman, P. G. Postema, and C. A. Remme, "Brugada syndrome: update and future perspectives," Heart, vol. 108, no. 9, pp. 668–675, May 2022, doi: 10.1136/heartjnl-2020-318258.

[22] B. Gray et al., "Twelve-lead ambulatory electrocardiographic monitoring in Brugada syndrome: Potential diagnostic and prognostic implications," Heart Rhythm, vol. 14, no. 6, pp. 866–874, Jun. 2017, doi: 10.1016/j.hrthm.2017.02.026.

[23] C. and Sensors, "12 Lead ECG Placement Guide | Cables & Sensors," Cables and Sensors. https://www.cablesandsensors.com/pages/12-lead-ecg-placement-guide-with-illustrations (accessed Jan. 06, 2023).

[24] M. Cerrone, S. Costa, and M. Delmar, "The Genetics of Brugada Syndrome," Annual Review of Genomics and Human Genetics, vol. 23, no. 1, pp. 255–274, 2022, doi: 10.1146/annurev-genom-112921-011200.

[25] E. Martínez-Barrios et al., "Brugada Syndrome in Women: What Do We Know After 30 Years?," Front Cardiovasc Med, vol. 9, p. 874992, 2022, doi: 10.3389/fcvm.2022.874992.

[26] W. G. Reiner, "Gender Identity and Sex Assignment: A Reappraisal for the 21st Century," in Pediatric Gender Assignment: A Critical Reappraisal, S. A. Zderic, D. A. Canning, M. C. Carr, and H. McC. Snyder, Eds. Boston, MA: Springer US, 2002, pp. 175–197. doi: 10.1007/978-1-4615-0621-8_11.

[27] C. G. Costello, "Beyond Binary Sex and Gender Ideology," in The Oxford Handbook of the Sociology of Body and Embodiment, N. Boero and K. Mason, Eds. Oxford University Press, 2020, p. 0. doi: 10.1093/oxfordhb/9780190842475.013.14.

[28] Sexing the Body. 2019. Accessed: Jan. 16, 2023. [Online]. Available: https://www.basicbooks.com/titles/anne-fausto-sterling/sexing-the-body/9781541672895/

[29] S. N. Dubin, I. T. Nolan, C. G. Streed, R. E. Greene, A. E. Radix, and S. D. Morrison, "Transgender health care: improving medical students' and residents' training and awareness," Adv Med Educ Pract, vol. 9, pp. 377–391, May 2018, doi: 10.2147/AMEP.S147183.

[30] J. P. Butler and J. Butler, Gender Trouble: Feminism and the Subversion of Identity. Routledge, 1990.

[31] L. Sörnmo and P. Laguna. Bioelectrical Signal Processing in Cardiac and Neurological Applications. Academic Press, 2005.

[32] S. Noorzadeh, M. Niknazar, B. Rivet, J. Fontecave-Jallon, P.-Y. Gumery, and C. Jutten, "Modeling quasi-periodic signals by a non-parametric model: application on fetal ECG extraction," Annu Int Conf IEEE Eng Med Biol Soc, vol. 2014, pp. 1889–1892, 2014, doi: 10.1109/EMBC.2014.6943979.

[33] Berbari, Edward and Steinberg, Jonathan. A Practical Guide to the Use of the High-resolution Electrocardiogram. Futura Publishing Company, 2000.

[34] D. D. Valle, Artificial Intelligence-A Modern Approach (3rd Edition). Accessed: Dec. 26, 2023. [Online]. Available: https://www.academia.edu/45007883/Artificial_Intelligence_A_Modern_Approach_3rd_Edition_

[35] A. M. TURING, "I.—COMPUTING MACHINERY AND INTELLIGENCE," Mind, vol. LIX, no. 236, pp. 433–460, Oct. 1950, doi: 10.1093/mind/LIX.236.433.

[36] Y. Baştanlar and M. Özuysal, "Introduction to Machine Learning," in miRNomics: MicroRNA Biology and Computational Analysis, M. Yousef and J. Allmer, Eds. Totowa, NJ: Humana Press, 2014, pp. 105–128. doi: 10.1007/978-1-62703-748-8_7.

[37] S. Badillo et al., "An Introduction to Machine Learning," Clinical Pharmacology & Therapeutics, vol. 107, no. 4, pp. 871–885, 2020, doi: 10.1002/cpt.1796.

[38] "Train models to classify data using supervised machine learning - MATLAB." https://www.mathworks.com/help/stats/classificationlearner-app.html (accessed Jan. 07, 2023).

[39] "Misclassification Costs in Classification Learner App - MATLAB & Simulink." https://www.mathworks.com/help/stats/misclassification-costs-in-classification-learner-app.html (accessed Jan. 06, 2023).

[40] "Holter electrocardiografico. Diagnóstico. Clínica Universidad de Navarra." https://www.cun.es/enfermedades-tratamientos/pruebas-diagnosticas/holter (accessed Jan. 10, 2023).

[41] SpiderView®. MicroPort®., "SORIN CRM LA456 SpiderView Holter Recorder User Manual," UserManual.wiki. https://usermanual.wiki/SORIN-CRM/LA456 (accessed Jan. 11, 2023).

[42] "LivaNova." https://www.livanova.com/Home (accessed Jan. 11, 2023).

[43] "Home page," Microport. https://microport.com/ (accessed Jan. 11, 2023).

[44] "Patient Setup Tips for Holter Monitors," Physicians Ancillary Systems. https://physiciansancillarysystems.com/patient-setup/ (accessed Jan. 13, 2023).

[45] Badilini, Fabio and ISHNE Standard Output Format Task Force. The ISHNE Holter standard output file format. Annals of Noninvasive Electrocardiology, 1998, vol. 3, no 3, p. 263-266. http://www.amps-llc.com/uploads/2017-12-11/the_ishne_format.pdf

[46] "MathWorks - Makers of MATLAB and Simulink." https://www.mathworks.com/ (accessed Dec. 28, 2022).

[47] "MATLAB - MathWorks." https://www.mathworks.com/products/matlab.html (accessed Dec. 28, 2022).

[48] "Signal Processing Toolbox." https://www.mathworks.com/products/signal.html (accessed Dec. 28, 2022).

[49] A. J. Demski and M. Llamedo Soria, "ecg-kit: a Matlab Toolbox for Cardiovascular Signal Processing," JORS, vol. 4, no. 1, p. 8, Apr. 2016, doi: 10.5334/jors.86.

[50] A. Afroz and A. Abdullah, "Detection of ECG R-Waves using Wavelet Transform," European Journal of Scientific Research, vol. 145, pp. 181–187, May 2017.

[51] "Classification Learner App - MATLAB & Simulink." https://www.mathworks.com/help/stats/classification-learner-app.html (accessed Jan. 09, 2023).

[52] S. Luo and P. Johnston, "A review of electrocardiogram filtering," J Electrocardiol, vol. 43, no. 6, pp. 486–496, 2010, doi: 10.1016/j.jelectrocard.2010.07.007.

[53] S. Aziz, S. Ahmed, and M.-S. Alouini, "ECG-based machine-learning algorithms for heartbeat classification," Sci Rep, vol. 11, no. 1, Art. no. 1, Sep. 2021, doi: 10.1038/s41598-021-97118-5.

[54] T. Nakamura, T. Aiba, W. Shimizu, T. Furukawa, and T. Sasano, "Prediction of the Presence of Ventricular Fibrillation From a Brugada Electrocardiogram Using Artificial Intelligence," Circulation Journal, vol. advpub, p. CJ-22-0496, 2022, doi: 10.1253/circj.CJ-22-0496.

[55] A. Chiu, "Celebrities use private jets excessively. It's a climate nightmare.," Washington Post, Aug. 03, 2022. Accessed: Jan. 14, 2023. [Online]. Available: https://www.washingtonpost.com/climate-environment/2022/08/02/taylor-swift-kylie-jenner-private-jet-emissions/

[56] A. Turiel, Petrocalipsis : Crisis energética global y cómo (no) la vamos a solucionar. Editorial Alfabeto, 2020. Accessed: Jan. 14, 2023. [Online]. Available: https://digital.csic.es/handle/10261/220715

[57] “Salario en España - Salario Medio,” Talent.com. https://es.talent.com/salary (accessed Jan. 13, 2023).

Annex A. Signal processing

Matlab code used to convert, segment and average the ECG recordings.

A1. 'm01_Process_ISHNE_ecg_kit_Holter2segment_3min.m'

```
% Process the ISHNE files with the holter ECG registers, saving them
% in 3 minute segments as .mat files
% The free software ecg-kit is used:
% http://marianux.github.io/ecg-kit/
% Install and execute: InstallECGkit.m
% Most recent master version in: https://github.com/marianux/ecg-kit

% Author P.Gomis and L.Tortosa(2022)

% The file (patient, etc.) names should be modified to fit each case

for pat=[1,10] %adapt to needs

cd(sprintf('D:\\Pat_BH%d',pat))

ECG =
ECGwrapper('recording_name',sprintf('D:\\BrugadaSignals\\BH000%d',pat));

% Total samples of the recording
Nsamp=ECG.ECG_header.nsamp;
fs = ECG.ECG_header.freq; % Sampling frequency

% Number of 3 minute segments
Nsegm = round(Nsamp/(fs*60*3));
% The first segment will begin after 60 seconds
ini = 60001;

for i = 0:Nsegm-2
    segm = i+1;
    fin = ini+3*60*fs-1;
    signal = ECG.read_signal(ini,fin);  % signals in
    % Output in ADC units and int16 format
    signal = double(signal);   % signals in int16 --> float
    % Convert to Physical units (by default nV)
    gain = ECG.ECG_header.gain; % gain=4e-4 -> 1/2500 (Resolution = 2500
nV)
    baseline = ECG.ECG_header.adczero;
    [N,leads]=size(signal);

    for ii=1:leads
        signal(:,ii) = (signal(:,ii)-baseline(ii))/gain(ii);
    end

    % Units in nV (nanoVolts) --> mV (*1e-6)
    signal = signal.*1e-6;   % units in mV
    signal = detrend(signal);  % remove offset and working with all leads
    header = ECG.ECG_header;
```

```
    header.nsamp = N;
    eval(['save ECG_' num2str(pat) '_' num2str(segm) ' signal header'])
end

end
```

A2. 'm02_Process_Signal_Average_beats.m'

```
%%% Additional processing and signal averaging
%%% of each segment of each patient

% Author P.Gomis and L.Tortosa(2022)


N = 1000;  %Number of samples of the resulting files

patients = [1:8,10:39,42:46]; %Patient numbers, required because of the
file names used

for Npat = 1:length(patients)

    pat = patients(Npat);

    cd(sprintf('D:\\Pat_BH%d',pat))

for i = 1:479

    numE=num2str(i);
    ecgName = sprintf('ECG_%d_%d.mat',pat,i);
    ECG = ECGwrapper('recording_name',ecgName);

    header = ECG.ECG_header; %complete header saved

    % Save the file FileName_QRS_detection.mat
    fs = ECG.ECG_header.freq;
    s = ECG.read_signal(1,ECG.ECG_header.nsamp);
    leads = {'I','II','III','aVR','aVL','aVF', 'V1', 'V2', 'V3',
'V4','V5','V6'};

    % Parts without information are removed

    s2 = s(:,2);
    sV1 = s(:,7);
    sV2 = s(:,8);

    M2 = max(s2);
    m2 = min(s2);
    MV1 = max(sV1);
    mV1 = min(sV1);
    MV2 = max(sV2);
    mV2 = min(sV2);

    dif2 = abs(M2-m2);
    difV1 = abs(MV1-mV1);
    difV2 = abs(MV2-mV2);

    if dif2 > 0.5 || difV1 > 0.5 || difV2 > 0.5
```

```
    % Detect QRS: R wave peak
    ECG.ECGtaskHandle = 'QRS_detection';
    ECG.ECGtaskHandle.detectors = {'wavedet', 'pantom'}; % R peak
detection methods
    ECG.Run;

    ecgNameDect = sprintf('D:\\ECG_%d_%d_QRS_detection.mat',pat,i);

    load(ecgNameDect)
    Nb = zeros(1,3);
    Nb2 = zeros(1,3); %additional signal using pantomkins, finally unused
    signal = zeros(fs,3);
    signal2 = zeros(fs,3);

    for ii = 1:3
        lst = [2,7,8];
        k = lst(ii);
        eval(['tw = wavedet_' leads{k} '.time;'])
        eval(['tp = pantom_' leads{k} '.time;'])
        Nb(ii)=length(tw);
        Nb2(ii)=length(tp);
        nint = fs;    % A window of 1 second is chosen, even if that
includes part of the following beat
        ECGens=zeros(Nb(ii)-1,nint);
        ECGens2=zeros(Nb2(ii)-1,nint);

        %Signal averaging itself

        for n=5:Nb(ii)-5
            ECGens(n,:)=s(tw(n)-399:tw(n)+nint-400,k);
        end

        for n=5:Nb2(ii)-5
            ECGens2(n,:)=s(tp(n)-399:tp(n)+nint-400,k);
        end
        signal(:,ii) = mean(ECGens);
        signal2(:,ii) = mean(ECGens2);
    end

    header.nsamp=N;
    header.nsig=3;
    header.desc=['II';'V1';'V2'];
    ecgNameOUT = sprintf('ECG_%d_%d_beat',pat,i);
    save(ecgNameOUT, 'signal', 'signal2', 'header', 'Nb','Nb2')
    delete(ecgNameDect)

    else

    end
end

end
```

Annex B. Biomarker Calculation

Matlab code used, directly or indirectly, to calculate the biomarkers.

B.1 'm03_Calc_BeatsData.m'

```
% Number of beats for each patient, position and
% total beats are calculated,
% necessary for the rest of the code.

% Author L.Tortosa (2022)

patients = [1:8,10:39,42:46,48,50]; %adapt to needs

m=1;
Nbeats = zeros(length(patients),1);
Pbeats = zeros(length(patients),1);
Tbeats = 0;

for Npat = 1:length(patients)

    pat = patients(Npat);

    cd(sprintf('D:\\Pat_BH%d',pat))

    while (isfile(sprintf('ECG_%d_%d_beat.mat',pat,m))) == 1

    m = m+1;
    Tbeats = Tbeats+1;

    end

    Tbeats = Tbeats-1;
    Nbeats(Npat) = m-2;

    if Npat>1

    Pbeats(Npat) = Tbeats-Nbeats(Npat);

    else

    Pbeats(Npat) = 1;

    end

    m = 1;
end

cd('D:\\')

save BeatsData Nbeats Pbeats Tbeats
```

B.2 'm04_Calc_RRmean'

```
%%% Biomarker necessary to calculate the corrected QT

% Author L.Tortosa (2022)

patients = [1:8,10:39,42:46,48,50]; %adapt to needs

leads = {'II', 'V1', 'V2'};

cd('D:')

load('D:\\BeatsData.mat')

RRmean = zeros(max(Nbeats),length(patients));


for Npat = 1:length(patients)

    for i = 1:Nbeats(Npat)

    pat = patients(Npat);

    cd(sprintf('D:\\Pat_BH%d',pat))

    ecgName = sprintf('ECG_%d_%d.mat',pat,i);
    load(ecgName)

    ECG = ECGwrapper('recording_name',ecgName);

    ecgDet = sprintf('D:\\ECG_%d_%d_QRS_detection.mat',pat,i);
    ecgDet2 =
sprintf('D:\\Detections\\ECG_%d_%d_QRS_detection.mat',pat,i);

    if isfile(ecgDet) == 1

        load(ecgDet)

    elseif isfile(ecgDet2) == 1

        load(ecgDet2)

    else

        ECG.ECGtaskHandle = 'QRS_detection';
        ECG.ECGtaskHandle.detectors = {'wavedet'};
        ECG.Run;

        load(ecgDet)
        movefile(ecgDet,'D:\\Detections');

    end

    tw = wavedet_II.time;
    fs = ECG.ECG_header.freq;

    RR = diff(tw);
    RR = RR/fs;
```

```
    RRmean(i,Npat) = mean(RR);

    end
end


% This part of the code removes the two last values of the RRmean table,
% which usually means one with a NaN value (as the algorithm finds
nothing
% and stops) and one with unreliable data before that, as a fraction of
it
% will be an empty signal


RRmean((max(Nbeats)+3),:) = 0;

 for Npat = 1:length(patients)

     for beat = 3:(Nbeats(Npat)+3)

     if RRmean(beat,Npat) == 0

         RRmean(beat-1,Npat) = 0;
         RRmean(beat-2,Npat) = 0;

     else
     end
     end
 end

 RRmean = RRmean(1:max(Nbeats),:);
```

B.3 'm05_Calc_Biomark.m'

```
%%% Code for calculating the biomarkers. Loading the files resulting from
%%% Calc_BeatsData and Calc_RRmean is required

% Author L.Tortosa (2022)


patients = [1:8,10:39,42:46,48,50];

leads = {'II', 'V1', 'V2'};

load('D:\\BeatsData.mat')

load('D:\\RRmean.mat')

if isfile('D:\\Excels\\Allbeats_nleads_wPatnum.mat') == 1 || H < TBeats

    load('D:\\Excels\\Allbeats_nleads_wPatnum.mat')
    Allbeats_nleads_wPatnum(Tbeats,:) = 0;

elseif isfile('D:\\Excels\\Allbeats_nleads_wPatnum.mat') == 0
```

```
    Allbeats_nleads_wPatnum = zeros(Tbeats,1+(7*3));

else
    load('D:\\Excels\\Allbeats_nleads_wPatnum.mat')

end

for Npat = 1:length(patients)

    pat = patients(Npat);
    Nbeat = Nbeats(Npat);

  for i = 1:Nbeat

    nrow = Pbeats(Npat,1);

    if Allbeats_nleads_wPatnum(nrow+i,1) == 0

    ecgName = sprintf('ECG_%d_%d_beat.mat',pat,i);
    cd(sprintf('D:\\Pat_BH%d',pat))
    load(ecgName)

    % Add 1000 previous samples of the initial value
    % Necessary as the delineation function starts after second 1

    s1 = signal(1,:); %Only the signal averaged using the wavelet method
is used
    signal = [s1.*ones(1000,3);signal];
    header.nsamp = 2000;

    %New file with the 2s signal
    ecgName2 = sprintf('ECG_%d_%d_beat2.mat',pat,i);
    save(ecgName2, 'signal', 'header')
    ECG = ECGwrapper('recording_name',ecgName2);
    fs = ECG.ECG_header.freq;
    s = ECG.read_signal(1,ECG.ECG_header.nsamp); %E?

    % Delineation process: calculation of beginning and end of waves and
    % peaks

    if
isfile(sprintf('D:\\Delineations\\ECG_%d_%d_beat2_ECG_delineation.mat',pa
t,i)) == 1


load(sprintf('D:\\Delineations\\ECG_%d_%d_beat2_ECG_delineation.mat',pat,
i))

    else

    ECG.ECGtaskHandle = 'ECG_delineation';
    ECG.Run;

    ecgNameD = sprintf('D:\\ECG_%d_%d_beat2_ECG_delineation.mat',pat,i);
    %ecgNameD = sprintf('ECG_%d_%d_beat_ECG_delineation.mat',nP,i);

    load(ecgNameD)
```

```
movefile(sprintf('D:\\ECG_%d_%d_beat2_ECG_delineation.mat',pat,i),'D:\\De
lineations')

    end

    w = wavedet;

    % Calculation of the biomarkers

    for k = 1:length(leads)
        if ~isempty(eval(['w.' leads{k} '.QRSon']))
            eval(['QRSon = w.' leads{k} '.QRSon(end);'])  % Beginning QRS
        else
            QRSon = NaN;
        end
        if ~isempty(eval(['w.' leads{k} '.QRSoff']))
            eval(['QRSoff = w.' leads{k} '.QRSoff(end);'])  % End QRS
        else
            QRSoff = NaN;
        end

         if ~isempty(eval(['w.' leads{k} '.T']))
            eval(['T = w.' leads{k} '.T(end);'])  % Beginning T
        else
            T = NaN;
        end
        if ~isempty(eval(['w.' leads{k} '.Toff']))
            eval(['Toff = w.' leads{k} '.Toff(end);'])  % End T
        else
            Toff = NaN;
        end

        %%%QRS [Pavg, AreaQRSabs]
        if ~isnan(QRSon) && ~isnan(QRSoff)
            qrs = signal(QRSon:QRSoff,k);  % qrs signal values are
separated
            QRSl = length(qrs);    % QRS length

            Pavg = 1/QRSl*sum(qrs.^2); %  Average potency of QRS (mV^2)
            AreaQRSabs = 1/fs*sum(abs(qrs))*1000; % Absolute QRS
area(uV·s)

        else
            QRSl = NaN;
            Pavg = NaN;
            AreaQRSabs = NaN;
        end


        Allbeats_nleads_wPatnum(nrow+i,1) = pat;

        if k == 1
            Allbeats_nleads_wPatnum(nrow+i,2) = Pavg;
            Allbeats_nleads_wPatnum(nrow+i,5) = AreaQRSabs;

        elseif k == 2
            Allbeats_nleads_wPatnum(nrow+i,3) = Pavg;
            Allbeats_nleads_wPatnum(nrow+i,6) = AreaQRSabs;
```

```
        else
            Allbeats_nleads_wPatnum(nrow+i,4) = Pavg;
            Allbeats_nleads_wPatnum(nrow+i,7) = AreaQRSabs;

        end

        %%% QT [QT, QTc]
        % QT: goes from QRSon until Toff
        qt = Toff - QRSon +1;
        eval(['QT.' leads{k} '(Npat,3*i-2)  = Toff - QRSon +1;']) % QT in
ms

        % Calculation of the baseline

        if ~isnan(QRSon) && ~isnan(QRSoff)
            segPR = median(signal(QRSon-30:QRSon, k));
            signal(:,k) = signal(:,k) - segPR;  % se pone la linea
IsoElect en 0
        else
        end

        qt = qt/fs;  % QT in seconds

        % The particular RR value is selected for each beat
        RRi = RRmean(i,Npat);

        if ~isnan(qt)
            QTc = qt/RRi.^(1/2);  % QTc_Barzett vector
            QTc = QTc*fs;     % Turned to miliseconds (ms)
        else
            QTc = NaN;
        end

        if k == 1
            Allbeats_nleads_wPatnum(nrow+i,8) = qt;
            Allbeats_nleads_wPatnum(nrow+i,11) = QTc;

        elseif k == 2
            Allbeats_nleads_wPatnum(nrow+i,9) = qt;
            Allbeats_nleads_wPatnum(nrow+i,12) = QTc;

        else
            Allbeats_nleads_wPatnum(nrow+i,10) = qt;
            Allbeats_nleads_wPatnum(nrow+i,13) = QTc;

        end


        %%% ST = [ST_0, ST_60, ST_slope]
        % ST_0: ST level at 0 ms (J point)
        if  ~isnan(QRSoff)
            ST_0 = median(signal(QRSoff+(-2:2),k));
            ST_60 = median(signal(QRSoff+60+(-2:2),k));
            ST_slope = 1000*(ST_60 - ST_0)/60; % uV/ms
        else
            ST_0 = NaN;
            ST_60 = NaN;
            ST_slope = NaN;
```

```
        end

        if k == 1
            Allbeats_nleads_wPatnum(nrow+i,14) = ST_0;
            Allbeats_nleads_wPatnum(nrow+i,17) = ST_60;
            Allbeats_nleads_wPatnum(nrow+i,20) = ST_slope;

        elseif k == 2
            Allbeats_nleads_wPatnum(nrow+i,15) = ST_0;
            Allbeats_nleads_wPatnum(nrow+i,18) = ST_60;
            Allbeats_nleads_wPatnum(nrow+i,21) = ST_slope;

        else
            Allbeats_nleads_wPatnum(nrow+i,16) = ST_0;
            Allbeats_nleads_wPatnum(nrow+i,19) = ST_60;
            Allbeats_nleads_wPatnum(nrow+i,22) = ST_slope;

        end
    end



    else
    end

  end


end

cd('D:\\')

save Biomarkers_Raw Allbeats_nleads_wPatnum

cd('D:\\Excels')

save Biomarkers_Raw Allbeats_nleads_wPatnum

F = num2cell(Allbeats_nleads_wPatnum);

F(1,:) = {'PatNum','Pavg_II','Pavg_V1','Pavg_V2',...
    'AreaQRSabs_II','AreaQRSabs_V1','AreaQRSabs_V2',...
    'qt_II','qt_V1','qt_V2','QTc_II','QTc_V1','QTc_V2',...
    'ST_0_II','ST_0_V1','ST_0_V2','ST_60_II','ST_60_V1',...
    'ST_60_V2','ST_slope_II','ST_slope_V1','ST_slope_V2'};

writecell(F, 'Biomarkers_Raw.xlsx')
```

Annex C. Data selection and distribution

C.1 'm06_Optional_Trim_data.m'

```
%%%Necessary to run before 'Add_clinical_data.m'

% Author L.Tortosa (2022)


T = readtable('D:\\Excels\\Biomarkers_Raw.xlsx'); %adapt name/directory

H = height(T);
W = width(T);

pcs1_1 = prctile(T.Pavg_II,[1,50,99]);
pcs1_2 = prctile(T.Pavg_V1,[1,50,99]);
pcs1_3 = prctile(T.Pavg_V2,[1,50,99]);

pcs2_1 = prctile(T.AreaQRSabs_II,[1,50,99]);
pcs2_2 = prctile(T.AreaQRSabs_V1,[1,50,99]);
pcs2_3 = prctile(T.AreaQRSabs_V2,[1,50,99]);

pcs3_1 = prctile(T.qt_II,[1,50,99]);
pcs3_2 = prctile(T.qt_V1,[1,50,99]);
pcs3_3 = prctile(T.qt_V2,[1,50,99]);

pcs4_1 = prctile(T.QTc_II,[1,50,99]);
pcs4_2 = prctile(T.QTc_V1,[1,50,99]);
pcs4_3 = prctile(T.QTc_V2,[1,50,99]);

pcs5_1 = prctile(T.ST_0_II,[1,50,99]);
pcs5_2 = prctile(T.ST_0_V1,[1,50,99]);
pcs5_3 = prctile(T.ST_0_V2,[1,50,99]);

pcs6_1 = prctile(T.ST_60_II,[1,50,99]);
pcs6_2 = prctile(T.ST_60_V1,[1,50,99]);
pcs6_3 = prctile(T.ST_60_V2,[1,50,99]);

pcs7_1 = prctile(T.ST_slope_II,[1,50,99]);
pcs7_2 = prctile(T.ST_slope_V1,[1,50,99]);
pcs7_3 = prctile(T.ST_slope_V2,[1,50,99]);

TPr = zeros(round(0.5*H),W);
Ta = table2array(T);

for i = 1:H

    count = 0;

    for j = 1:7

        for lead = 1:3

            count = count+1;
```

```
            eval(sprintf('mini = pcs%d_%d(1);',j,lead));
            eval(sprintf('maxi = pcs%d_%d(3);',j,lead));

            if mini < T{i,count+1} && T{i,count+1} < maxi

            else
                T{i,count+1} = 999;
                Ta(i,:) = 999;
            end

        end
    end

end

TPc = 1;

for ii = 1:H

    if Ta(ii,1) == 999

    else
        TPr(TPc,:) = T{ii,:};
        TPc = TPc+1;
    end

end

Addinglables = zeros(TPc,width(TPr));
Addinglables(2:TPc,:) = TPr;

Biomarkers_Pr = num2cell(Addinglables);

Biomarkers_Pr(1,:) = {'PatNum','Pavg_II','Pavg_V1','Pavg_V2',...
    'AreaQRSabs_II','AreaQRSabs_V1','AreaQRSabs_V2',...
    'qt_II','qt_V1','qt_V2','QTc_II','QTc_V1','QTc_V2',...
    'ST_0_II','ST_0_V1','ST_0_V2','ST_60_II','ST_60_V1',...
    'ST_60_V2','ST_slope_II','ST_slope_V1','ST_slope_V2'};

cd('D:\\')

writecell(Biomarkers_Pr, 'Biomarkers_Pr.xlsx');

save Biomarkers_Pr Biomarkers_Pr

length_Bio_Pr = TPc;
save length_Bio_Pr length_Bio_Pr
```

C.2 'm07_Add_clinical_data.m'

```
% Author L.Tortosa (2022)

% Additional clinical data from the excel can be selected
% by uncommenting specific commands

% Only the ones finally used in the project are uncommented
```

```
patients = [1:8,10:39,42:46,48,50];

T = readtable('D:\\Biomarkers_Pr.xlsx');
CData = readtable('D:\\Clinical_data_final.xlsx'); %adapt to new formats

L = height(CData);

H = height(T);

T.Symp = NaN(H,1);
T.Sex = NaN(H,1);
T.Dia_age = NaN(H,1);
% T.Proband = NaN(H,1);
% T.FH_SCD = NaN(H,1);
% T.FH_SCD_bef55 = NaN(H,1);
% T.Sy_b_Dg_wo_vagal_wo_oth = NaN(H,1);
% T.Sync_bef_Dg = NaN(H,1);
% T.SD_bef_Dg = NaN(H,1);
% T.Symp_bef_Dg_DIC_wo_vagal = NaN(H,1);
% T.sponECGtype1_all = NaN(H,1);
% T.sponECG_fixedtype1_all = NaN(H,1);
% T.EPS_inducib_dic = NaN(H,1);
% T.EPS_HV = NaN(H,1);
% T.Genotype_SCN5A = NaN(H,1);

for beat = 1:H

    PatNum = T.PatNum(beat);

    for j = 1:L

    CDPatNum = CData.Pat_Num(j);

    if PatNum == CDPatNum

    T.Symp(beat) = CData.Symp(j);
    T.Sex(beat) = CData.Sex(j);
    T.Dia_age(beat) = CData.Diagnosis_age(j);
%     T.Proband(beat) = CData.Proband(j);
%     T.FH_SCD(beat) = CData.FH_SCD(j);
%     T.FH_SCD_bef55(beat) = CData.FH_SCD_bef_55(j);
%     T.Sy_b_Dg_wo_vagal_wo_oth(beat) =
CData.Symptoms_bef_Dg_wo_vagal_wo_other(j);
%     T.Sync_bef_Dg(beat) = CData.Syncope_bef_Dg(j);
%     T.SD_bef_Dg(beat) = CData.SD_bef_Dg(j);
%     T.Symp_bef_Dg_DIC_wo_vagal(beat) =
CData.Symptoms_bef_Dg_DIC_wo_vagal(j);
%     T.sponECGtype1_all(beat) = CData.sponECGtype1_all(j);
%     T.sponECG_fixedtype1_all(beat) = CData.sponECG_fixedtype1_all(j);
%     T.EPS_inducib_dic(beat) = CData.EPS_inducib_dic(j);
%     T.EPS_HV(beat) = CData.EPS_HV(j);
%     T.Genotype_SCN5A(beat) = CData.Genotype_SCN5A(j);

    else

    end
    end
```

```
end

% T = movevars(T,"Genotype_SCN5A",'After',"PatNum");
% T = movevars(T,"EPS_HV",'After',"PatNum");
% T = movevars(T,"EPS_inducib_dic",'After',"PatNum");
% T = movevars(T,"sponECG_fixedtype1_all",'After',"PatNum");
% T = movevars(T,"sponECGtype1_all",'After',"PatNum");
% T = movevars(T,"Symp_bef_Dg_DIC_wo_vagal",'After',"PatNum");
% T = movevars(T,"SD_bef_Dg",'After',"PatNum");
% T = movevars(T,"Sync_bef_Dg",'After',"PatNum");
% T = movevars(T,"Sy_b_Dg_wo_vagal_wo_oth",'After',"PatNum");
% T = movevars(T,"FH_SCD_bef55",'After',"PatNum");
% T = movevars(T,"FH_SCD",'After',"PatNum");
% T = movevars(T,"Proband",'After',"PatNum");
T = movevars(T,"Dia_age",'After',"PatNum");
T = movevars(T,"Sex",'After',"PatNum");
T = movevars(T,"Symp",'After',"PatNum");

save Tv4 T
writetable(T, 'Tv4.xlsx')
```

C.3 'm08_Select_test_data.m'

```
% Author L.Tortosa (2022)

cd('D:\\')
T = readtable('Tv3.xlsx'); %revise name/directory
Ta = table2array(T);

load('D:\\BeatsData.mat')

patients = [1:8,10:39,42:46,48,50];

H = height(T);
W = width(T);

TestData = zeros((length(patients)*8)+1,W);
TrainData = zeros(H-(height(TestData)+1),W);
TrainData(2,:) = Ta(1,:);
PTestD = zeros(height(TestData),1);

Testc = 2;

for pat = 1:length(patients)

    Pos = Pbeats(pat)+25;
    TestData(Testc,:) = Ta(Pos,:);
    PTestD(Testc-1) = Pos;
    Testc = Testc+1;

    Pos = Pbeats(pat)+50;
    TestData(Testc,:) = Ta(Pos,:);
    PTestD(Testc-1) = Pos;
    Testc = Testc+1;
```

```
    Pos = Pbeats(pat)+75;
    TestData(Testc,:) = Ta(Pos,:);
    PTestD(Testc-1) = Pos;
    Testc = Testc+1;

    Pos = Pbeats(pat)+100;
    TestData(Testc,:) = Ta(Pos,:);
    PTestD(Testc-1) = Pos;
    Testc = Testc+1;

    Pos = Pbeats(pat)+125;
    TestData(Testc,:) = Ta(Pos,:);
    PTestD(Testc-1) = Pos;
    Testc = Testc+1;

    Pos = Pbeats(pat)+150;
    TestData(Testc,:) = Ta(Pos,:);
    PTestD(Testc-1) = Pos;
    Testc = Testc+1;

    Pos = Pbeats(pat)+175;
    TestData(Testc,:) = Ta(Pos,:);
    PTestD(Testc-1) = Pos;
    Testc = Testc+1;

    Pos = Pbeats(pat)+200;
    TestData(Testc,:) = Ta(Pos,:);
    PTestD(Testc-1) = Pos;
    Testc = Testc+1;

end

Trainc = 2;
Pcount = 1;

for i = 1:H

   if i == PTestD(Pcount)
       Pcount = Pcount + 1;

   else
       TrainData(Trainc,:) = Ta(i,:);
       Trainc = Trainc+1;

   end
end

FTe = num2cell(TestData);
FTr = num2cell(TrainData);

FTe(1,:) = T.Properties.VariableNames;
FTr(1,:) = T.Properties.VariableNames;

writecell(FTe, 'TestData.xlsx');
writecell(FTr, 'TrainData.xlsx');

save TestData FTe
save TrainData FTr
```

C.4 'm09_Select_test_data_2.m'

```
% Author L.Tortosa (2022)

cd('D:\\')
T = readtable('Tv3.xlsx'); %revise name/directory
Ta = table2array(T);

load('D:\\BeatsData.mat')

patients = [1:8,10:39,42:46,48,50];

H = height(T);
W = width(T);

%Numbers are adapted to account for the missing patient 9
%This part requires having performed Calc_BeatsData on the dataset
before,
%but it can be removed with ease in order to not need it (see m10)

TestData2 = zeros((Nbeats(2)+Nbeats(9)+Nbeats(22)+Nbeats(32)),W);

TestData_Pat2 = zeros(Nbeats(2),W);
TestData_Pat10 = zeros(Nbeats(9),W);
TestData_Pat23 = zeros(Nbeats(22),W);
TestData_Pat33 = zeros(Nbeats(32),W);

TrainData2 = zeros(H-(height(TestData2)+1),W);

Testc = 2;
Pat2c = 2;
Pat10c = 2;
Pat23c = 2;
Pat33c = 2;
Trainc = 2;

for beat = 1:H

   if T.PatNum(beat) == 2

       TestData2(Testc,:) = Ta(beat,:);
       Testc = Testc + 1;

       TestData_Pat2(Pat2c,:) = Ta(beat,:);
       Pat2c = Pat2c + 1;

   elseif T.PatNum(beat) == 10

       TestData2(Testc,:) = Ta(beat,:);
       Testc = Testc + 1;

       TestData_Pat10(Pat10c,:) = Ta(beat,:);
       Pat10c = Pat10c + 1;

   elseif T.PatNum(beat) == 23
```

```
        TestData2(Testc,:) = Ta(beat,:);
        Testc = Testc + 1;

        TestData_Pat23(Pat23c,:) = Ta(beat,:);
        Pat23c = Pat23c + 1;

    elseif T.PatNum(beat) == 33

        TestData2(Testc,:) = Ta(beat,:);
        Testc = Testc + 1;

        TestData_Pat33(Pat33c,:) = Ta(beat,:);
        Pat33c = Pat33c + 1;

    else

        TrainData2(Trainc,:) = Ta(beat,:);
        Trainc = Trainc + 1;

    end
end

FTe2 = num2cell(TestData2);
FTeP2 = num2cell(TestData2);
FTeP10 = num2cell(TestData2);
FTeP23 = num2cell(TestData2);
FTeP33 = num2cell(TestData2);
FTr2 = num2cell(TrainData2);

FTe2(1,:) = T.Properties.VariableNames;
FTeP2(1,:) = T.Properties.VariableNames;
FTeP10(1,:) = T.Properties.VariableNames;
FTeP23(1,:) = T.Properties.VariableNames;
FTeP33(1,:) = T.Properties.VariableNames;
FTr2(1,:) = T.Properties.VariableNames;

writecell(FTe2, 'TestData2.xlsx');
writecell(FTeP2, 'TestData2P2.xlsx');
writecell(FTeP10, 'TestData2P10.xlsx');
writecell(FTeP23, 'TestData2P23.xlsx');
writecell(FTeP33, 'TestData2P33.xlsx');
writecell(FTr2, 'TrainData2.xlsx');

save TestData2 FTe2
save TestData2P2 FTeP2
save TestData2P10 FTeP10
save TestData2P23 FTeP23
save TestData2P33 FTeP33
save TrainData2 FTr2
```

C.5 'm10_Select_test_data_3.m'

```
% Author L.Tortosa (2022)

cd('D:\\')
T = readtable('Tv4.xlsx'); %revise name/directory
Ta = table2array(T);

patients = [1:8,10:39,42:46,48,50];

H = height(T);
W = width(T);


TestData3_Pat2 = zeros(10,W);
TestData3_Pat10 = zeros(10,W);
TestData3_Pat23 = zeros(10,W);
TestData3_Pat33 = zeros(10,W);

TestData3 = zeros(50,W);
TrainData3 = zeros(100,W);

Testc = 2;
Pat2c = 2;
Pat10c = 2;
Pat23c = 2;
Pat33c = 2;
Trainc = 2;

for beat = 1:H

   if T.PatNum(beat) == 2

       TestData3(Testc,:) = Ta(beat,:);
       Testc = Testc + 1;

       TestData3_Pat2(Pat2c,:) = Ta(beat,:);
       Pat2c = Pat2c + 1;

   elseif T.PatNum(beat) == 10

       TestData3(Testc,:) = Ta(beat,:);
       Testc = Testc + 1;

       TestData3_Pat10(Pat10c,:) = Ta(beat,:);
       Pat10c = Pat10c + 1;

   elseif T.PatNum(beat) == 23

       TestData3(Testc,:) = Ta(beat,:);
       Testc = Testc + 1;

       TestData3_Pat23(Pat23c,:) = Ta(beat,:);
       Pat23c = Pat23c + 1;

   elseif T.PatNum(beat) == 33

       TestData3(Testc,:) = Ta(beat,:);
       Testc = Testc + 1;
```

```
        TestData3_Pat33(Pat33c,:) = Ta(beat,:);
        Pat33c = Pat33c + 1;

    else

        TrainData3(Trainc,:) = Ta(beat,:);
        Trainc = Trainc + 1;

    end
end

FTe3 = num2cell(TestData3);
FTe3P2 = num2cell(TestData3);
FTe3P10 = num2cell(TestData3);
FTe3P23 = num2cell(TestData3);
FTe3P33 = num2cell(TestData3);
FTr3 = num2cell(TrainData3);

FTe3(1,:) = T.Properties.VariableNames;
FTe3P2(1,:) = T.Properties.VariableNames;
FTe3P10(1,:) = T.Properties.VariableNames;
FTe3P23(1,:) = T.Properties.VariableNames;
FTe3P33(1,:) = T.Properties.VariableNames;
FTr3(1,:) = T.Properties.VariableNames;

writecell(FTe3, 'TestData3.xlsx');
writecell(FTe3P2, 'TestData3P2.xlsx');
writecell(FTe3P10, 'TestData3P10.xlsx');
writecell(FTe3P23, 'TestData3P23.xlsx');
writecell(FTe3P33, 'TestData3P33.xlsx');
writecell(FTr3, 'TrainData3.xlsx');

save TestData3 FTe3
save TestData3P2 FTe3P2
save TestData3P10 FTe3P10
save TestData3P23 FTe3P23
save TestData3P33 FTe3P33
save TrainData3 FTr3
```

C.6 ‘m11_Select_test_data_4.m’

```
% Author L.Tortosa (2022)

cd('D:\\')
T = readtable('Tv4.xlsx'); %revise name/directory
Ta = table2array(T);

patients = [1:8,10:39,42:46,48,50];

H = height(T);
W = width(T);
```

```
TestData5_Pat3 = zeros(10,W);
TestData5_Pat5 = zeros(10,W);

TestData5 = zeros(50,W);
TrainData5 = zeros(100,W);

Testc = 2;
Pat3c = 2;
Pat5c = 2;
Trainc = 2;

for beat = 1:H

   if T.PatNum(beat) == 3

       TestData5(Testc,:) = Ta(beat,:);
       Testc = Testc + 1;

       TestData5_Pat3(Pat3c,:) = Ta(beat,:);
       Pat3c = Pat3c + 1;

   elseif T.PatNum(beat) == 5

       TestData5(Testc,:) = Ta(beat,:);
       Testc = Testc + 1;

       TestData5_Pat5(Pat5c,:) = Ta(beat,:);
       Pat5c = Pat5c + 1;

   else

       TrainData5(Trainc,:) = Ta(beat,:);
       Trainc = Trainc + 1;

   end
end

FTe5 = num2cell(TestData5);
FTe5P3 = num2cell(TestData5);
FTe5P5 = num2cell(TestData5);
FTr5 = num2cell(TrainData5);

FTe5(1,:) = T.Properties.VariableNames;
FTe5P3(1,:) = T.Properties.VariableNames;
FTe5P5(1,:) = T.Properties.VariableNames;
FTr5(1,:) = T.Properties.VariableNames;

writecell(FTe5, 'TestData5.xlsx');
writecell(FTe5P3, 'TestData5P3.xlsx');
writecell(FTe5P5, 'TestData5P5.xlsx');
writecell(FTr5, 'TrainData5.xlsx');

save TestData5 FTe5
save TestData5P2 FTe5P3
save TestData5P10 FTe5P5
save TrainData5 FTr5
```

C.7 'm12_Select_test_data_5.m'

```
% Author L.Tortosa (2022)

cd('D:\\')
T = readtable('Tv3.xlsx'); %revise name/directory
Ta = table2array(T);

patients = [1:8,10:39,42:46,48,50];

H = height(T);
W = width(T);

TestData6_s = zeros(500,W);
TestData6_l = zeros(2500,W);
TrainData6 = zeros(800,W);

Testc_s = 2;
Testc_l = 2;
Trainc = 2;

for beat = 1:H

   if T.PatNum(beat) == 3 || T.PatNum(beat) == 6

       TestData6_s(Testc_s,:) = Ta(beat,:);
       Testc_s = Testc_s + 1;

       TestData6_l(Testc_l,:) = Ta(beat,:);
       Testc_l = Testc_l + 1;

   elseif T.PatNum(beat) == 1 || T.PatNum(beat) == 5 || T.PatNum(beat) ==
14 || T.PatNum(beat) == 18 ...
       || T.PatNum(beat) ==29 || T.PatNum(beat) == 32 || T.PatNum(beat)
== 35 || T.PatNum(beat) == 36 ...
       || T.PatNum(beat) == 37 || T.PatNum(beat) == 42

       TestData6_l(Testc_l,:) = Ta(beat,:);
       Testc_l = Testc_l + 1;

   elseif T.PatNum(beat) == 2

       TrainData6(Trainc,:) = Ta(beat,:);
       Trainc = Trainc + 1;

   elseif T.PatNum(beat) == 10

       TrainData6(Trainc,:) = Ta(beat,:);
       Trainc = Trainc + 1;

   elseif T.PatNum(beat) == 23

       TrainData6(Trainc,:) = Ta(beat,:);
       Trainc = Trainc + 1;

   elseif T.PatNum(beat) == 33

       TrainData6(Trainc,:) = Ta(beat,:);
       Trainc = Trainc + 1;
```

```
    else

    end
end

FTe6s = num2cell(TestData6_s);
FTe6l = num2cell(TestData6_l);
FTr6 = num2cell(TrainData6);

FTe6s(1,:) = T.Properties.VariableNames;
FTe6l(1,:) = T.Properties.VariableNames;
FTr6(1,:) = T.Properties.VariableNames;

writecell(FTe6s, 'TestData6s.xlsx');
writecell(FTe6l, 'TestData6l.xlsx');
writecell(FTr6, 'TrainData6.xlsx');

save TestData6s FTe6s
save TestData6l FTe6l
save TrainData6 FTr6
```

Annex D. 'view_signal.m'

```
% Author P.Gomis and L.Tortosa (2022)

%%% For 1 second signals, beats

figure(1)
t=(1:1000);
plot(t,signal(:,1)+1.2,t,signal(:,2),t,signal(:,3)-1.2),
grid, xlabel('t (s)'), ylabel('mV')
grid minor
legend('II','V1','V2')
title('Patient 15 - Leads II, V1, V2')

%%% For 3 minute signals, direct segments extracted from the ECG Holter
%%% recording

figure(2)
t=(1:180000);
plot(t,signal(:,2)+1.2,t,signal(:,7),t,signal(:,8)-1.2),
grid, xlabel('t (s)'), ylabel('mV')
grid minor
legend('II','V1','V2')
title('Patient 2 - Leads II, V1, V2')
```

Annex E. Information for the patient and informed consent

HOJA DE INFORMACIÓN AL PACIENTE Y CONSENTIMIENTO INFORMADO

Estudio ANVERSO: **An**álisis de **V**ariables **E**lectrocardiográficas en relación al **R**iesgo de muerte **S**úbita en pacientes con síndr**O**me de Brugada

Version 1.0 – Fecha 10/03/2016

Nos dirigimos a usted para informarle sobre un estudio de investigación en el que se le invita a participar. El estudio ha sido aprobado por el Comité Ético de Investigación Clínica correspondiente y la Agencia Española del Medicamento y Productos Sanitarios, de acuerdo a la legislación vigente, y la Ley 14/2007, de 3 de julio, de Investigación biomédica.

Nuestra intención es tan solo que usted reciba la información correcta y suficiente para que pueda evaluar y juzgar si quiere o no participar en este estudio. Para ello lea esta hoja informativa con atención y nosotros le aclararemos las dudas que le puedan surgir después de la explicación. Además, puede consultar con las personas que considere oportuno.

PARTICIPACIÓN VOLUNTARIA

Debe saber que su participación en este estudio es voluntaria y que puede decidir no participar o cambiar su decisión y retirar el consentimiento en cualquier momento, sin que por ello se altere la relación con su médico ni se produzca perjuicio alguno en su tratamiento.

DESCRIPCIÓN GENERAL DEL ESTUDIO

Estamos realizando un estudio en la Unidad de Arritmias del Hospital Clínic de Barcelona en colaboración con la Universitat Politécnica de Catalunya (UPC) que pretende evaluar si existe una relación entre las características electrocardiográficas y el riesgo de sufrir muerte súbita en pacientes con síndrome de Brugada. Para llevarlo a cabo, precisamos de su participación.

El síndrome de Brugada es una enfermedad hereditaria caracterizada por signos electrocardiográficos típicos en precordiales derechas, que predispone a la muerte súbita secundaria a taquicardia ventricular polimórfica y/o fibrilación ventricular en ausencia de cardiopatía estructural. Las alteraciones electrocardiográficas pueden ser dinámicas y ocasionalmente encontrarse ocultas, pudiéndose presentar mediante dos patrones electrocardiográficos diferenciados. En el estudio

intentaremos analizar aquellas variables electrocardiográficas que permitan predecir, de forma no invasiva, el riesgo a sufrir arritmias ventriculares malignas (y consiguientemente muerte súbita). Si se confirmase el valor pronóstico de esta prueba, el hallazgo permitiría lograr una prevención más efectiva de la muerte súbita y arritmias ventriculares malignas en pacientes con síndrome de Brugada, mediante técnicas de monitorización automáticas y no invasivas. Para confirmar o rechazar esta hipótesis, hemos diseñado un estudio en pacientes que padecen esta enfermedad, entre los que usted se encuentra, a los que se analizarán electrocardiogramas de 24 a 48 horas de duración.

Objetivos del estudio

Este estudio pretende evaluar si existe una relación entre la estratificación del riesgo de sufrir eventos de muerte súbita en pacientes con síndrome de Brugada con ciertas variables electrocardiográficas analizadas en un registro Holter de 24 horas.

¿En qué consiste?

Individuos diagnosticados de síndrome de Brugada (tipo 1 fijo, alternante o tras la administración de fármacos bloqueantes de los canales de sodio) de cualquier edad, así como de ambos sexos.

Nuestro estudio consiste en la obtención de un registro electrocardiográfico (Holter). Si acepta participar, deberá llevar un Holter durante 24 a 48 horas consecutivas. Para ello, se le colocarán en el pecho 10 electrodos (pequeños parches conductores) que se conectarán a un pequeño monitor que registrará un electrocardiograma completo durante 24 a 48 horas. Deberá cargar con el monitor Holter en un bolsillo o en una bolsa pequeña que se puede llevar puesta alrededor del cuello o de la cintura. Si desea bañarse o tomar una ducha deberá hacerlo antes de comenzar el examen, ya que no podrá hacerlo mientras se esté usando el monitor Holter. Una vez finalizado el periodo de registro deberá volver al hospital para que le sean retirados los electrodos y su información de registro sea almacenada.

BENEFICIOS Y RIESGOS DERIVADOS DE SU PARTICIPACIÓN EN EL ESTUDIO

Beneficios esperados para los participantes en el estudio

La participación en este registro no prevé ningún tipo de compensación. El estudio, debido a su característica de ser observacional, y por lo tanto limitado al registro de cómo se maneja y trata su enfermedad, no prevé, de acuerdo a la normativa, ningún tipo de seguro. Su participación en el estudio no le supondrá ningún gasto. Como posible beneficio, se le realizará una evaluación continua

de su ritmo cardíaco durante 24-48 horas. Los datos le serán remitidos de forma codificada a su domicilio y, en caso de hallazgos relevantes serán valorados por su médico de referencia.

Beneficios esperados para la comunidad

Los resultados derivados del análisis de los datos podrían mejorar la gestión futura de los pacientes con síndrome de Brugada.

Alternativa a la participación

Usted es libre de no participar en el estudio. En este caso usted recibirá, sin embargo, todas las terapias estándar previstas para su condición, sin penalización alguna, y los médicos continuarán siguiéndole con la debida atención médica, incluso si no hubiese otras terapias disponibles.

CONFIDENCIALIDAD

El tratamiento, la comunicación y la cesión de los datos de carácter personal de todos los sujetos participantes se ajustará a lo dispuesto en la Ley Orgánica 15/1999, de 13 de diciembre de protección de datos de carácter personal. Los datos se recogerán en un fichero de investigación de centro y se tratarán única y exclusivamente en el marco de su participación en este estudio.

De acuerdo a lo que establece la legislación de protección de datos, usted puede ejercer los derechos de acceso, modificación, oposición y cancelación de datos, para lo cual deberá dirigirse a su médico del estudio.

Los datos recogidos para el estudio estarán identificados mediante un código y solo su médico del estudio/colaboradores podrá relacionar dichos datos con usted y con su historia clínica. Por lo tanto, su identidad no será revelada a persona alguna salvo en caso de urgencia médica o requerimiento legal.

El acceso a su información personal quedará restringido al médico del estudio/colaboradores, autoridades sanitarias (Agencia Española del Medicamento y Productos Sanitarios), al Comité Ético de Investigación Clínica y personal autorizado por el promotor, cuando lo precisen para comprobar los datos y procedimientos del estudio, pero siempre manteniendo la confidencialidad de los mismos de acuerdo a la legislación vigente.

OTRA INFORMACIÓN RELEVANTE

Si usted decide retirar el consentimiento para participar en este estudio, ningún dato nuevo será añadido a la base de datos y, puede exigir la destrucción de todas las muestras identificables previamente retenidas para evitar la realización de nuevos análisis.

También debe saber que puede ser excluido del estudio si el promotor o los investigadores del estudio lo consideran oportuno, ya sea por motivos de seguridad, por cualquier acontecimiento adverso que se produzca por la medicación en estudio o porque consideren que no está cumpliendo con los procedimientos establecidos. En cualquiera de los casos, usted recibirá una explicación adecuada del motivo que ha ocasionado su retirada del estudio

Al firmar la hoja de consentimiento adjunta, se compromete a cumplir con los procedimientos del estudio que se le han expuesto.

Cuando acabe su participación recibirá el mejor tratamiento disponible y que su médico considere el más adecuado para su enfermedad, pero es posible que no se le pueda seguir administrando la medicación del estudio. Por lo tanto, ni el investigador ni el promotor adquieren compromiso alguno de mantener dicho tratamiento fuera de este estudio.

IDENTIFICACIÓN DE LOS INVESTIGADORES

Si tiene preguntas sobre el estudio, póngase en contacto con la Dra. Elena Arbelo Lainez (Unidad de Arritmias, Servicio de Cardiología, Instituto Cardiovascular, Hospital Clínic de Barcelona. Teléfono: 93 227 5551).

HOJA DE CONSENTIMIENTO DE PARTICIPANTE

PROYECTO ANVERSO (Version 1.0 – Fecha 10/03/2016)

Yo, *(nombre y apellidos del participante)* .....

- He leído la hoja de información que se me ha entregado sobre el estudio.

- He podido hacer preguntas sobre el estudio.

- He recibido suficiente información sobre el estudio.

- He hablado con: *(nombre del investigador)* .....

- Comprendo que mi participación es voluntaria.

- Comprendo que puedo retirarme del estudio:

    - Cuando quiera.

    - Sin tener que dar explicaciones.

    - Sin que esto repercuta en mis cuidados médicos.

De conformidad con lo que establece la Ley Orgánica 15/1999, de 13 de Diciembre de Protección de Datos de Carácter Personal (artículo 3, punto 6 del Real Decreto 223/2004), declaro haber sido informado de la existencia de un fichero o tratamiento de datos de carácter personal, de la finalidad de la recogida de éstos y de los destinatarios de la información.

Presto libremente mi conformidad para participar en el estudio.

Firma del participante Firma del investigador

Fecha: ____/____/____ Fecha: ____/____/____

Deseo que me comuniquen la información derivada de la investigación que pueda ser relevante para mí salud: ☐ SI ☐ NO

Firma del participante Firma del investigador

Fecha: ____/____/____ Fecha: ____/____/____

Annex F. Statistics drawn from sub-datasets 3 and 4

Comparison of means between Symptomatic and Asymptomatic groups were performed using unpaired *t*-test. A p-value <0.05 was considered statistically significant. All analysis were performed using Statistics and Machine Learning Toolbox, Matlab.

Statistics are drawn from the shared data of sub-datasets 3 and 4, where beats with biomarkers whose values belonged in the upper and lower 1% of their specific type of biomarker were removed (above percentile 99 and below percentile 1).

Number of patients = 45, total beats: 11244, where Symptomatic = 8 (1807 beats) and Asymptomatic = 37 (9437).

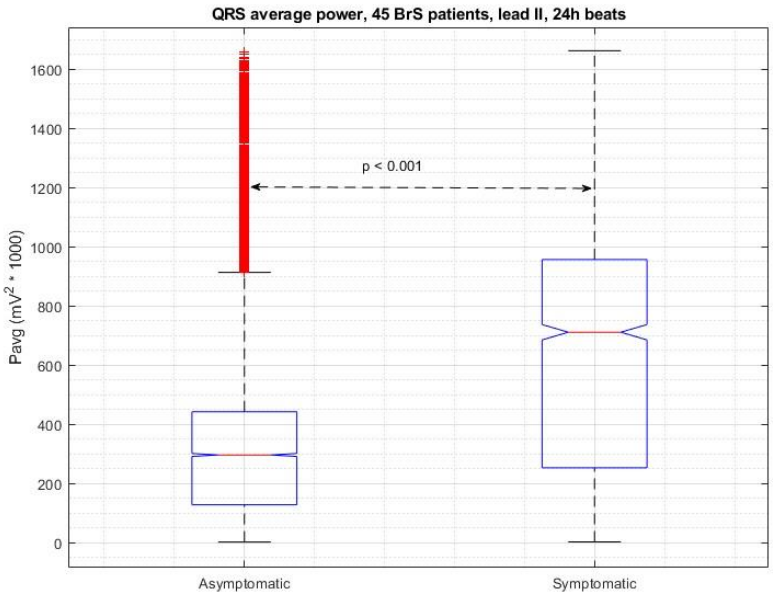


Figure F.1. Boxplot of QRS average power in Asymptomatic vs. Symptomatic group, measured in lead II.

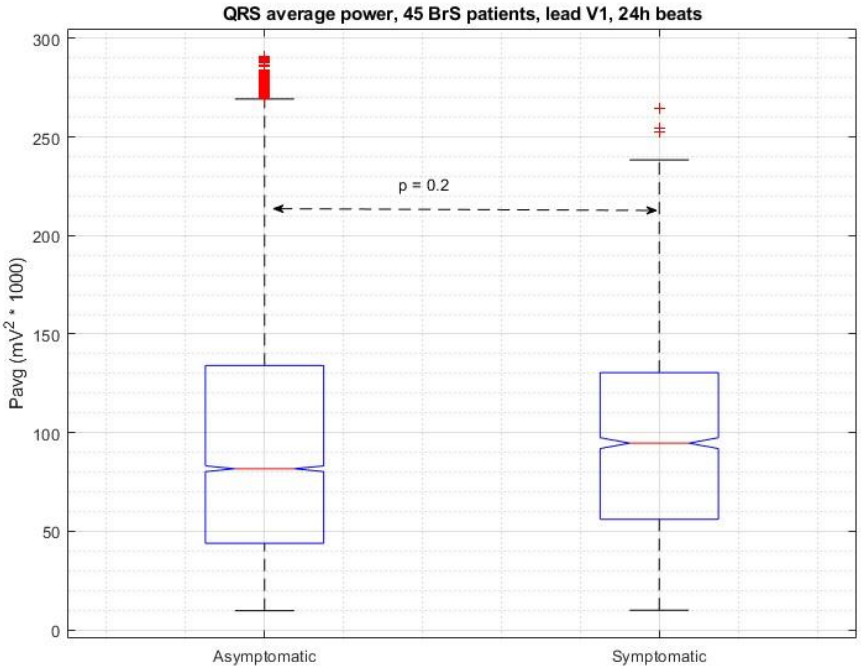
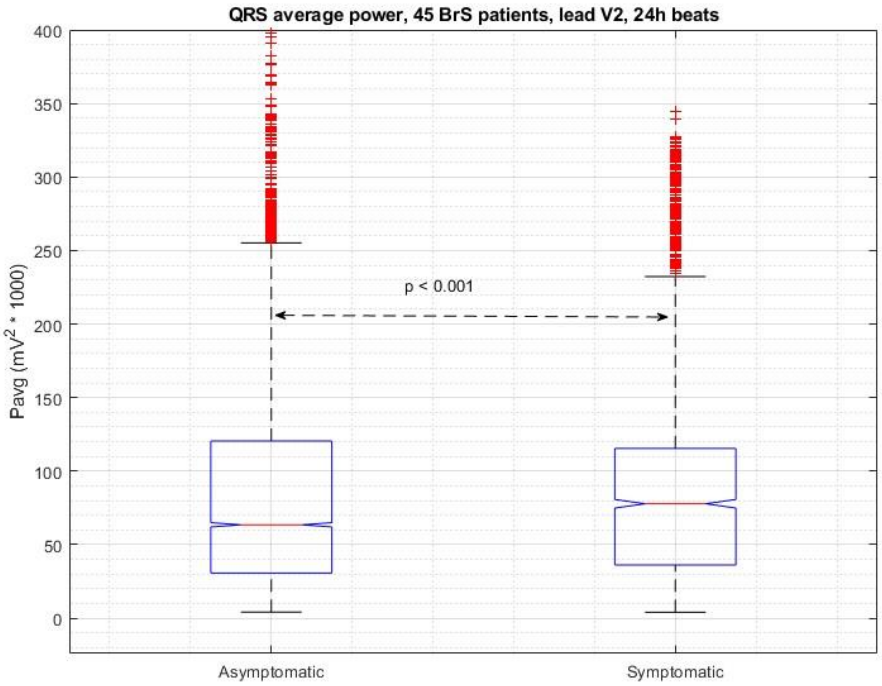


Figure F.2. Boxplot of QRS average power in Asymptomatic vs. Symptomatic group, measured in lead V1.



.

Figure F.3. Boxplot of QRS average power in Asymptomatic vs. Symptomatic group, measured in lead V2.

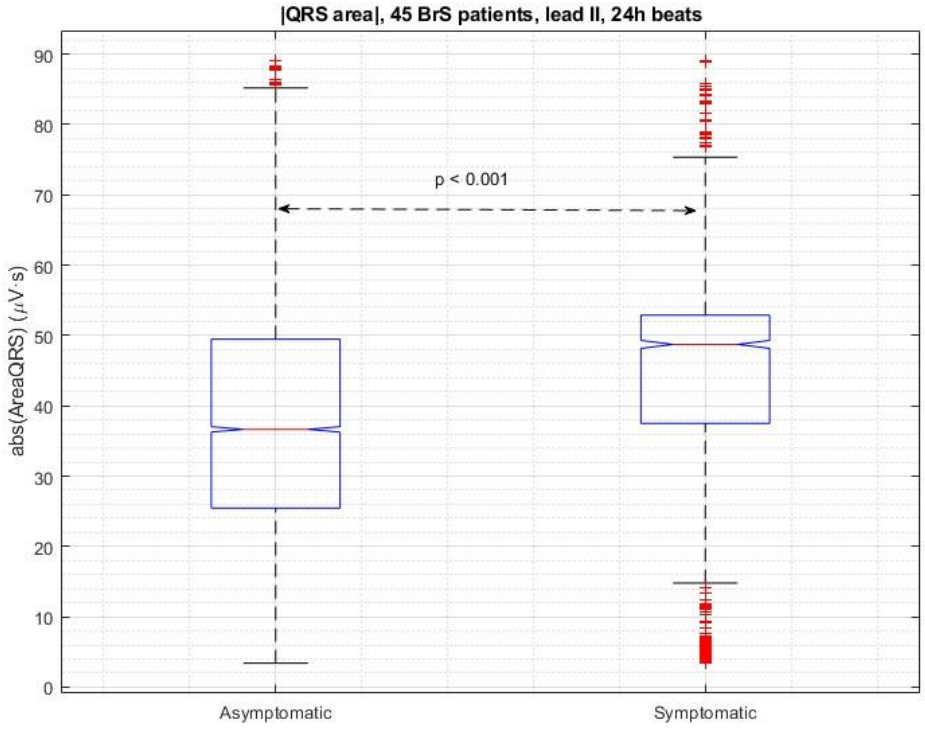


Figure F.4. Boxplot of the absolute value of QRS area in Asymptomatic vs. Symptomatic group, measured in lead II.

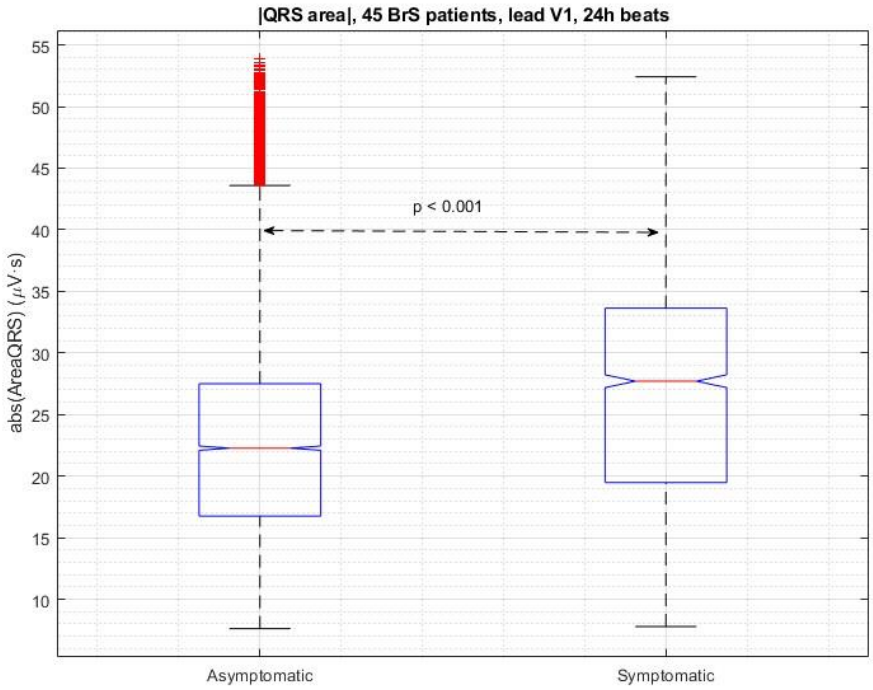


Figure F.5. Boxplot of the absolute value of QRS area in Asymptomatic vs. Symptomatic group, measured in lead V1.

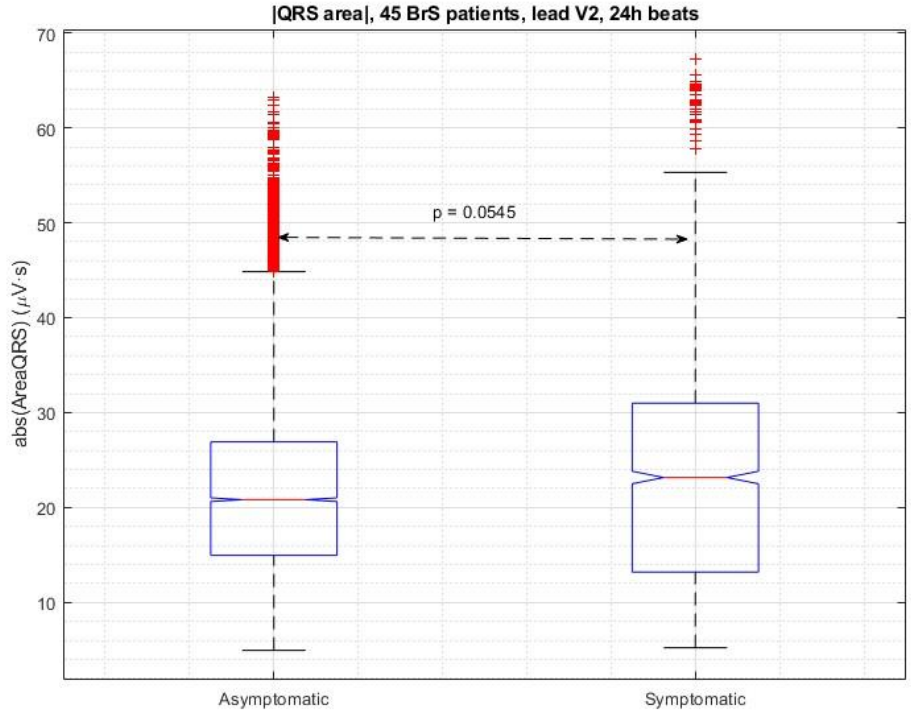


Figure F.6. Boxplot of the absolute value of QRS area in Asymptomatic vs. Symptomatic group, measured in lead V2.

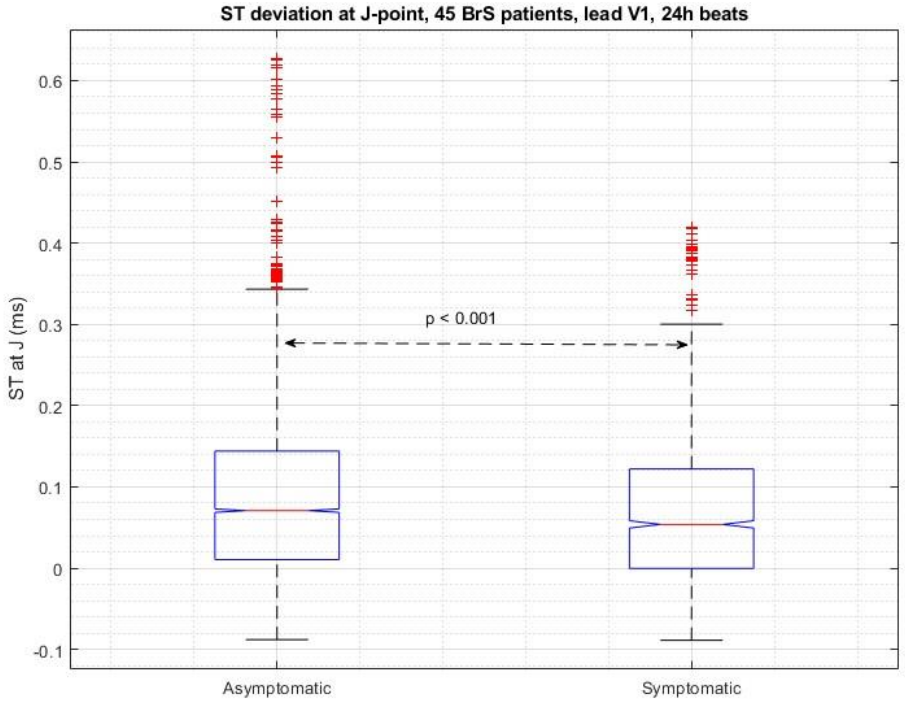


Figure F.7. Boxplot of the ST segment deviation at J-point in Asymptomatic vs. Symptomatic group, measured in lead V1.

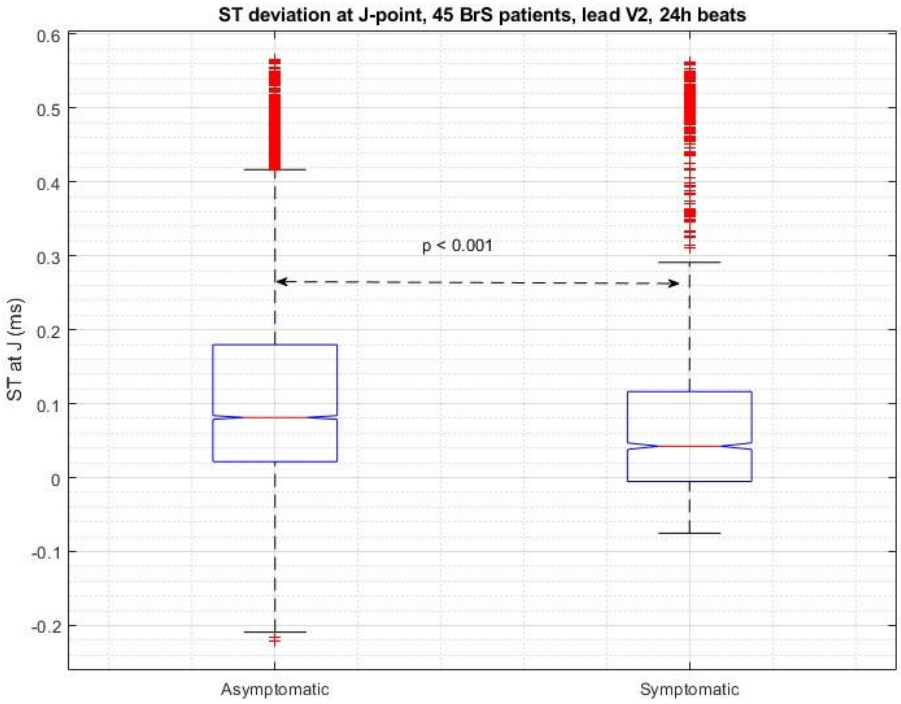


Figure F.8. Boxplot of the ST segment deviation at J-point in Asymptomatic vs. Symptomatic group, measured in lead V2.

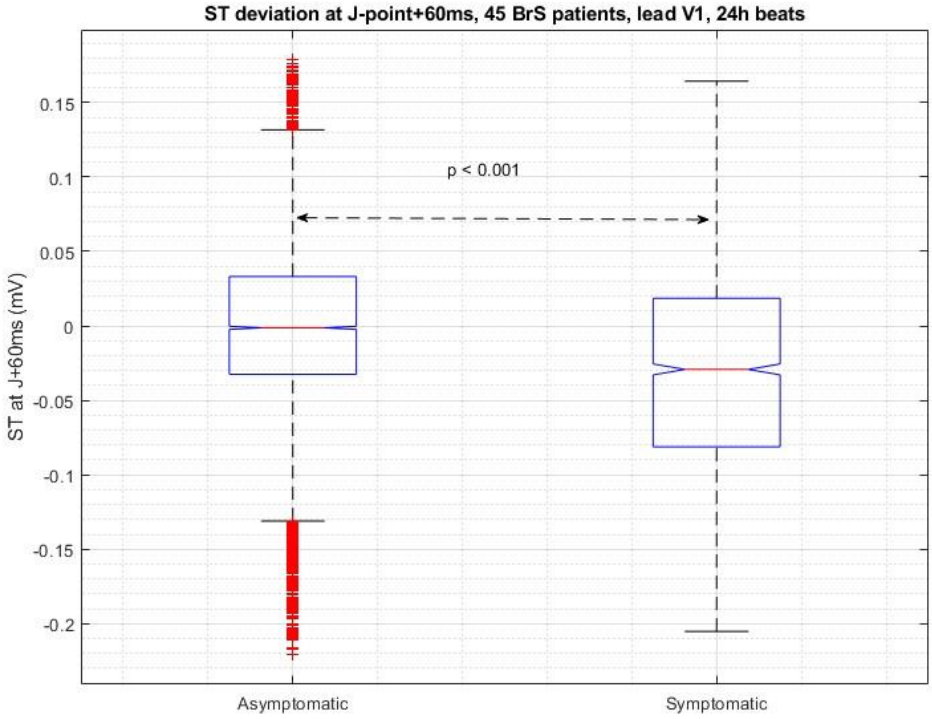


Figure F.9. Boxplot of the ST segment deviation at J-point + 60 ms in Asymptomatic vs. Symptomatic group, measured in lead V1.

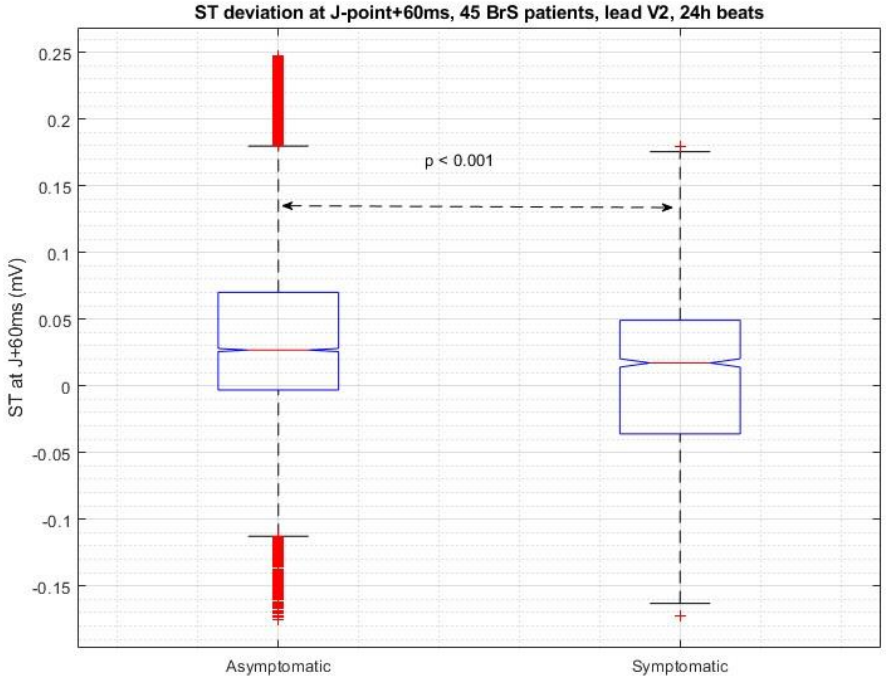


Figure F.10. Boxplot of the ST segment deviation at J-point + 60 ms in Asymptomatic vs. Symptomatic group, measured in lead V2.

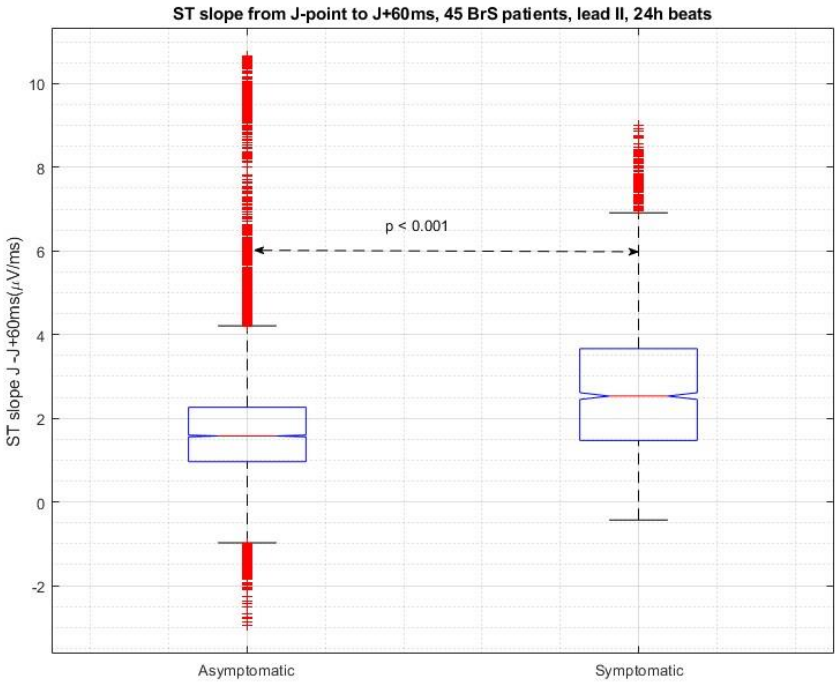


Figure F.11. Boxplot of the ST segment slope from J-point to J-point + 60 ms in Asymptomatic vs. Symptomatic group, measured in lead II.

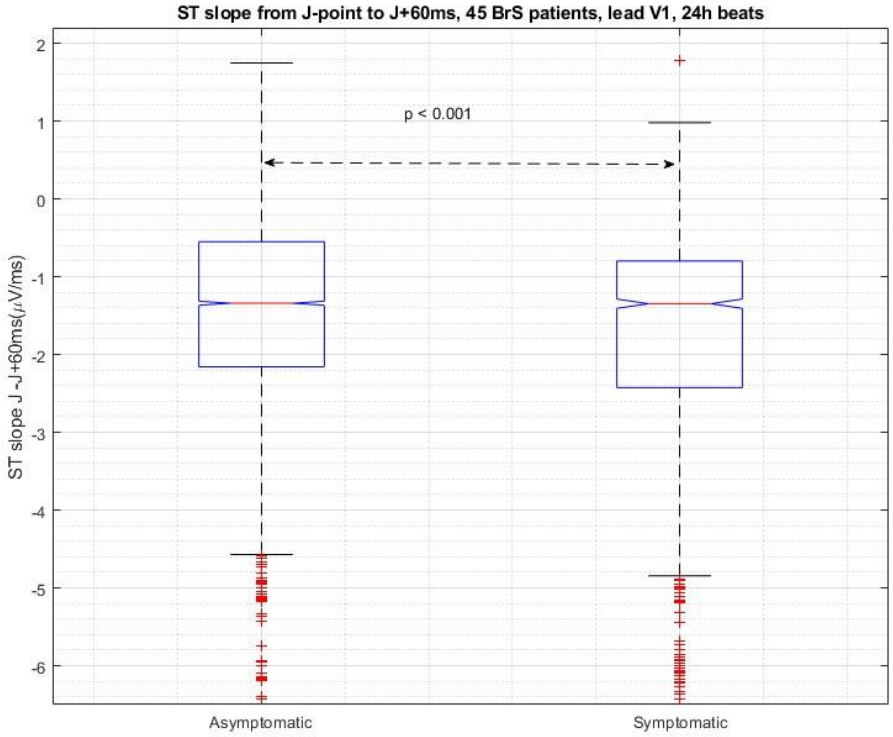


Figure F.12. Boxplot of the ST segment slope from J-point to J-point + 60 ms in Asymptomatic vs. Symptomatic group, measured in lead V1.

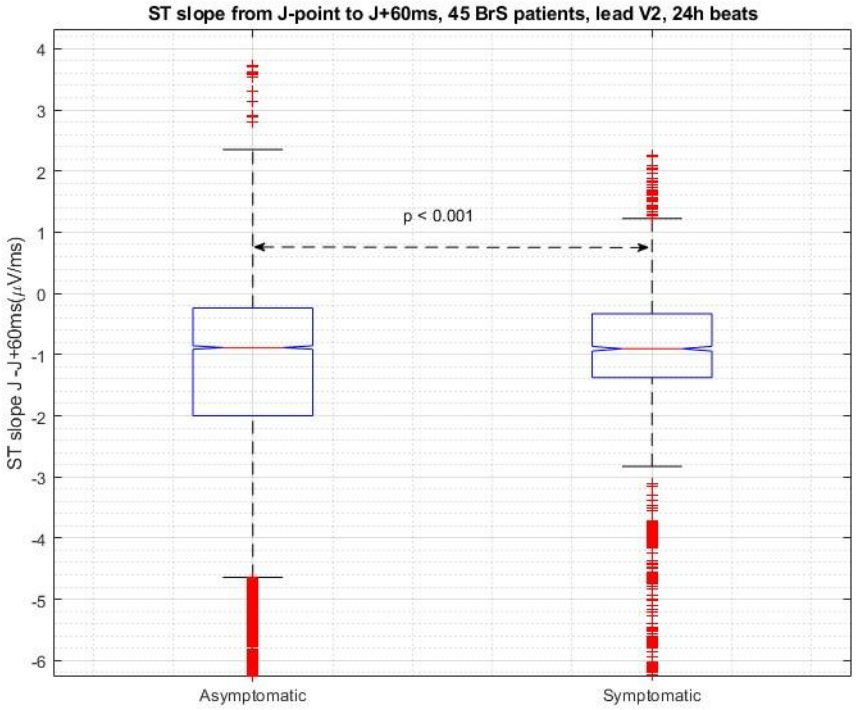


Figure F.13. Boxplot of the ST segment slope from J-point to J-point + 60 ms in Asymptomatic vs. Symptomatic group, measured in lead V2.

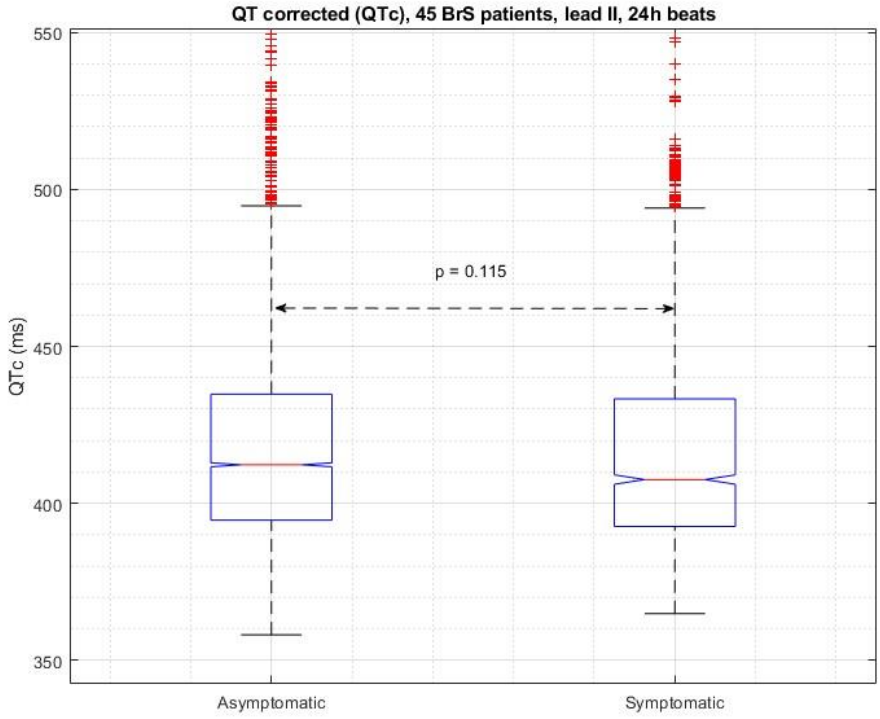


Figure F.13. Boxplot of the corrected QT (QTc) in Asymptomatic vs. Symptomatic group, measured in lead II.