eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
BILBOKO INGENIARITZA ESKOLA
ESCUELA DE INGENIERÍA DE BILBAO

BACHELOR'S DEGREE IN TELECOMMUNICATIONS ENGINEERING

BACHELOR'S FINAL DEGREE PROJECT

AN ALGORITHM TO CLASSIFY HEARTBEATS USING THE ELECTROCARDIOGRAM

Student: Zubia, Garea, Gorka

Manager: Irusta, Zarandona, Unai

Academic year: 2018-2019

Date: Barakaldo, 27th of June 2019

SUMMARY

Cardiovascular diseases (CVD) are the leading cause of death in the world. Therefore, their early detection and prevention is an urgent task with important consequences for public health and quality of live. Low cost non-invasive monitoring techniques to assess the state of the heart like the electrocardiogram (EKG) are an essential tool for the prevention of CVDs, and the early detection of arrhythmia. Many arrhythmia are associated to disfunctions in heart rate and the nature or origin of the heartbeats. Consequently, an automatic algorithm to identify heartbeats and classify them using the EKG would be an important tool for the early detection of CVDs.

The aim of this project has been to develop and implement an EKG based supervised algorithm to discriminate normal heartbeats from heartbeats originating in the ventricles, or ventricular heartbeats. To accomplish this goal several intermediate goals have been defined and achieved; first, the adaptation of an openly available EKG database following the international standards, and the development of an easy to use graphical user interface (GUI) for the visualisation and handling of those EKG signals and their annotations. Then, two algorithms were used to detect and delineate the heartbeats from the database, namely the Hamilton Tompkins heartbeat detector and the Wavedec heartbeat delineator.

From each heartbeat fifteen characteristics were calculated, including morphological and interval features. Afterwards, these features were statistically characterised for normal and ventricular heartbeats to identify distinctive patterns that will help differentiate normal from ventricular beats. Finally, the best feature combination was determined and it included only four features. These features were used to train a machine learning logistic regression classifier to discriminate normal from ventricular beats, and a separate set of heartbeats was used to test the algorithm. The classifier correctly classified 90.98% of the ventricular beats and 85.98% of normal beats in the test set. These results are comparable or even better than some of the algorithms proposed in the literature for heartbeat classification.

In conclusion an accurate algorithm for the discrimination of normal and ventricular heartbeats was developed. Furthermore, the best EKG delineation features for the discrimination of normal and ventricular heartbeats were identified, and we showed that a classifier based only on four features achieved the best results.

Key words: **signal processing, telecommunications engineering, bioengineering, statistical learning, Machine Learning, Logistic Regression, EKG, MATLAB.**

LABURPENA

Gaixotasun kardiobaskularrak (GKB) dira mundu mailan hilkortasun kausa nagusia. Beraz, GKBak garaiz detektatzea eta hauen prebentzioa premiazko zereginak dira, biek ere osasun publikoan zein bizi-kalitatean ondorio garrantzitsuak baitituzte. Elektrokardiograma (EKG) bihotzaren egoera monitorizatzeko kostu baxuko modu ez-inbaditzailea da, eta funtsezko tresna da GKBak sahiesteko eta arritmiak garaiz detektatzeko. Arritmia asko bihotz-maiztasunaren edota bihotz taupaden izaera edo jatorriaren disfuntzioekin lotzen dira. Horrengatik, bihotz-taupadak identifikatzeko eta EKGa erabiliz bihotz-taupadak sailkatzeko algoritmo automatikoa tresna garrantzitsua izango litzateke GKBn detekzio goiztiarrerako.

Proiektu honen helburua izan da EKGn oinarritutako algoritmo bat garatu eta inplementatzea bentrikuletan sortzen diren taupadak taupada arruntetatik bereizteko. Helburu hori lortzeko, bitarteko helburuak zehaztu eta bete izan dira. Lehenik eta behin, sarbide libreko elektrokardiogramen datu-base bat nazioarteko estandarrak jarraituz egokitu da eta erabiltzaile-interfaze grafiko bat garatu da, elektrokardiograma seinaleak eta taupada anotazioak bistaratzeko eta erabiltzeko. Ondoren, bi algoritmo erabili dira datu-baseko bihotz-taupadak detektatzeko eta delineatzeko, Hamilton Tompkins taupada-detektagailua eta Wavedec taupada delineatzailea, alegia.

Taupada bakoitzeko hamabost ezaugarri kalkulatu dira, horien artean ezaugarri morfologikoak eta denborazkoan barne. Ondoren, ezaugarri horiek estatistikoki karakterizatu dira taupada normal eta bentrikularrentzako. Horrela, taupada bentrikularrak arruntetatik bereizten laguntzen duten patroi bereizgarriak identifikatu dira. Azkenik, ezaugarrien konbinaziorik onena identifikatu da, lau ezaugarriz bakarrik osatuta dagoela ondorioztatu delarik. Ezaugarri horiek ikasketa automatikoko erregresio logistiko sailkatzaile bat entrenatzeko erabili dira, taupada bentrikularrak arruntetatik bereizteko. Algoritmoa entrenatzeko eta emaitzak lortzeko taupada multzo desberdinak erabili dira. Sailkatzaileak behar bezala sailkatzen ditu taupada bentrikularren %90.98 eta taupada arrunten %85.98. Emaitza horiek, halaber, bihotz-taupadak sailkatzeko literaturan proposatutako algoritmoetako batzuen antzekoak edota hobeagoak dira.

Ondorioz, bihotz-taupada normalak eta bentrikularrak bereizteko algoritmo zehatza garatu da. Gainera, bihotz-taupadak eta bentrikularrak bereizteko elektrokardiogramaren delineazioaren ezaugarri onenak identifikatu dira, eta lau ezaugarritan oinarritutako sailkatzaileak emaitzarik onenak ematen dituela frogatu da.

Hitz gakoak: **seinaleen prozesaketa, telekomunikazio ingeniaritza, bioingeniaritza, ikasketa estatistikoa, ikasketa automatikoa, erregresio logistikoa, EKG, MATLAB.**

RESUMEN

Las enfermedades cardiovasculares (ECV) son la principal causa de muerte en el mundo. Por lo tanto, su detección precoz y prevención es una tarea urgente con importantes consecuencias para la salud pública y la calidad de vida. Las técnicas de monitorización no invasiva de bajo coste para evaluar el estado del corazón como el electrocardiograma (ECG) son una herramienta esencial para la prevención de las ECVs y la detección precoz de arritmias. Muchas arritmias se asocian a disfunciones en la frecuencia cardíaca y la naturaleza y origen de los latidos. Por consiguiente, un algoritmo automático para identificar los latidos del corazón y clasificarlos utilizando el electrocardiograma sería una herramienta importante para la detección precoz de las ECV.

El objetivo de este proyecto ha sido desarrollar e implementar un algoritmo supervisado basado en ECG para discriminar entre los latidos cardíacos normales de los latidos que se originan en los ventrículos o latidos cardíacos ventriculares. Para lograr este objetivo se han definido y cumplido varios objetivos intermedios. En primer lugar, la adaptación de una base de datos de electrocardiogramas de libre acceso siguiendo los estándares internacionales, y el desarrollo de una interfaz gráfica de usuario fácil de usar para la visualización y manejo de dichas señales de electrocardiograma y sus anotaciones. Luego, se utilizaron dos algoritmos para detectar y delinear los latidos cardíacos de la base de datos, concretamente, el detector de latidos de Hamilton Tompkins y el delineador de latidos cardíacos Wavedec.

De cada latido se calcularon quince características, incluyendo características morfológicas y de intervalo. Posteriormente, estas características se caracterizaron estadísticamente para los latidos cardíacos normales y ventriculares con el fin de identificar patrones distintivos que ayuden a diferenciar los latidos normales de los ventriculares. Finalmente, se determinó la mejor combinación de características y se incluyeron sólo cuatro características. Estas características se utilizaron para entrenar un clasificador de regresión logística de aprendizaje automático para discriminar los latidos normales de los ventriculares, y se utilizó un conjunto separado de latidos cardíacos para probar el algoritmo. El clasificador clasificó correctamente el 90,98% de los latidos ventriculares y el 85,98% de los latidos normales en el conjunto de prueba. Estos resultados son comparables o incluso mejores que algunos de los algoritmos propuestos en la literatura para la clasificación de los latidos cardíacos.

En conclusión, se ha desarrollado un algoritmo preciso para la discriminación de los latidos cardíacos normales y ventriculares. Además, se identificaron las mejores características de delineación del electrocardiograma para la discriminación de los latidos cardíacos normales y ventriculares, y se demostró que un clasificador basado sólo en cuatro características logró los mejores resultados.

Palabras clave: **procesado de señales, ingeniería de telecomunicaciones, bioingeniería, aprendizaje estadístico, aprendizaje automático, regresión logística, EKG, MATLAB.**

Contents

List of Figures

List of Tables

List of acronyms

AAMI	Association for the Advancement of Medical Instrumentation
BioRes	Bioengineering and Resuscitation
CVD	Cardiovascular diseases
EKG	Electrocardiogram
FN	False Negative
FP	False Positive
G	Working group
GUI	Graphical user interface
HT	Hamilton-Tompkins
ID	Identifier
MIT-BIH	Massachusetts Institute of Technology - Beth Israel Hospital
ML	Machine Learning
NPV	Negative Predictive Value
LR	Logistic Regression
PPV	Positive Predictive Value
SA	Sinoatrial
Se	Sensibility
Sp	Sensitivity
T	Task
TP	True Negative
TN	True Positive
VA	Ventricular arrhythmia
VF	Ventricular fibrillation
VT	Ventricular tachycardia
WP	Working package

Beat types

F	Fusion of ventricular and normal beat
N	Normal beat
S/SEVB	Supraventricular Ectopic Beat
Q	Unknown beat
V/VEB	Ventricular Ectopic Beat

1 INTRODUCTION

This bachelor's final degree project has been developed within the Bioengineering and Resuscitation (BioRes) research group of the University of the Basque Country (UPV/EHU) at the Faculty of Engineering in Bilbao. The research group's work is focused towards the application of Digital Signal Processing and Machine Learning techniques to biomedical signals recorded by monitors and defibrillators during cardiac arrest.

Nowadays, cardiovascular diseases (CVD) are the main cause of death on planet Earth, accounting for 32.26% of deaths worldwide [1]. CVDs are often left untreated because they usually occur without pain or obvious symptoms. Untreated CVDs may lead to more serious health problems, or even death in the mid to long term. One of biggest dangers associated to CVDs is they are frequently silent, that is they can be suffered without clear symptoms but with fatal consequences [2]. Fortunately, the most prevalent forms of CVD are chronic and unfatal, and are therefore entirely preventable: firstly, with a healthy lifestyle and secondly, with early and accurate detection [3].

The electrocardiogram (EKG) is a non-invasive, general purpose and low-cost powerful tool for the premature diagnosis of cardiac dysfunctions. Cardiac disfunctions can be grossly classified into two groups. On the one hand, there are the so-called lethal ventricular arrhythmia, such as ventricular tachycardia or ventricular fibrillation. These life-threatening arrhythmia require an immediate intervention, for instance with a defibrillator. Today, this is a highly developed area which already has increasingly reliable detectors that save the lives of thousands of people every year [4, 5]. On the other hand, there are chronic non-lethal arrhythmia that may not require immediate intervention, but can be deleterious for the health of the patient in the mid to long term. Arrhythmia can be caused by a wide number of reasons, a general cause being the aging of the heart. Arrhythmia are not usually life-threatening, but should be diagnosed and treated in their early stage to prevent future problems [3]. This project focuses on the early detection of arrhythmia.

An initial stage in the detection of arrhythmia is the identification of heartbeats and their classification using the EKG. From this information, it is possible to determine when the patient suffers abnormal heartbeats, which frequently preclude arrhythmia. A posterior analysis will determine whether the patient has an arrhythmia or not. At present, there are multiple studies that have proposed algorithms for EKG heartbeat detection [6] and classification [7–10].

Given the importance of early detection, it would be extremely helpful to have an accurate method for the detection and classification of heartbeats using non-invasive techniques like the EKG. In addition, the heartbeats detected and classified by this method could be used as input by more complex algorithms to diagnose arrhythmia in a simple, automatic and cheap way.

2 BACKGROUND

2.1 Cardiovascular diseases in figures

CVDs are a major health problem, and are the underlying cause of about a third of the deaths worldwide, as shown in **Figure 1**. They affect the heart by constricting the arteries and reducing the amount of blood the heart receives, which makes the heart work harder. Currently, more than 17.56 million people die annually as a consequence of these pathologies [11]. As an illustration, deaths caused by CVD double those caused by all types of cancer together (see **Figure 1**). CVDs are therefore a major global health problem.

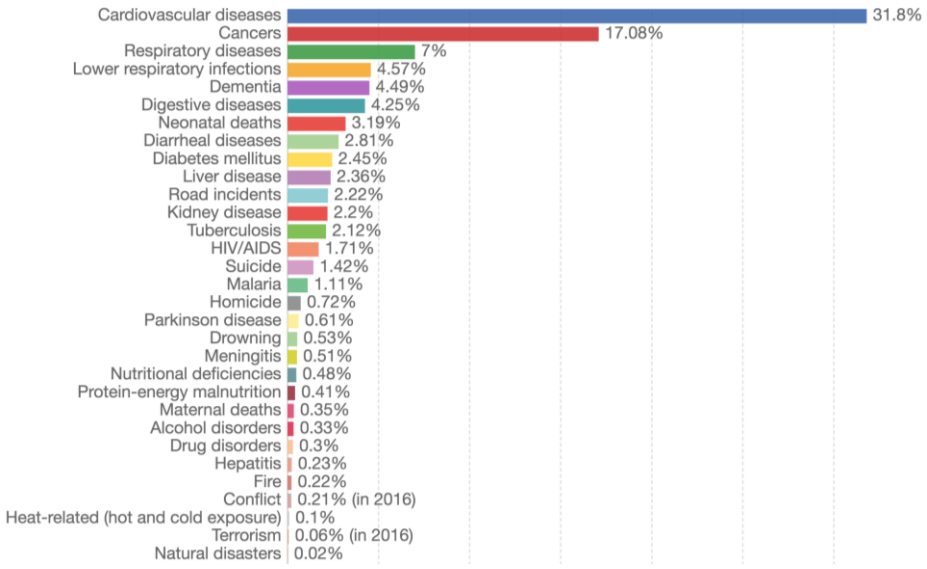


Figure 1: Share of deaths by cause in the world, 2017 [1].

In the European Union, 37% of people die of CVD related problems [12]. A detailed proportion by country is shown in ¡Error! No se encuentra el origen de la referencia.. Similarly, in the Basque Country, approximately 27% of the population passes away due to CVDs [13]. This means that every year more than 5,800 people die in the Basque Country of CVDs [14].

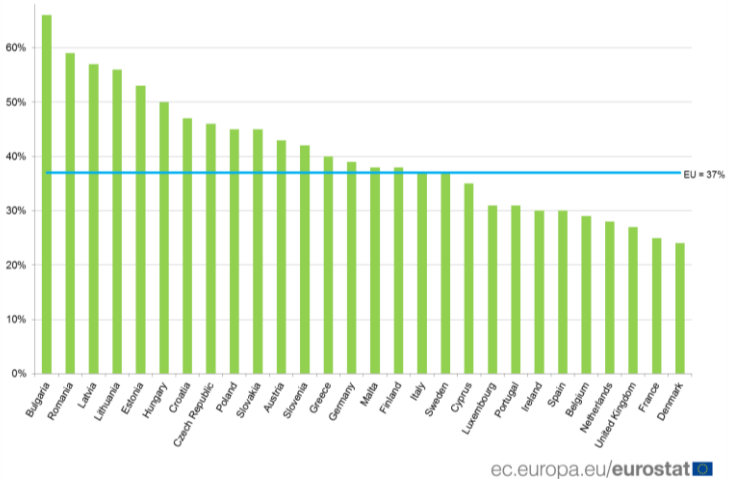


Figure 2: Proportion of deaths due to CVD in the EU member states in 2014

Prevention and treatment of CVDs is of great importance for health systems, since more than 90% of cardiovascular accidents are preventable [15]. Currently the annual cost of treatment of cardiovascular diseases in Spain is of over 9000M € [16], a burden that keeps increasing as population grows older. Hence, measures to prevent and/or more efficiently treat CVDs would be of great value.

One important step in that direction is the early detection of CVDs, not only to reduce the cost generated by treatment and hospitalization, but primarily to save lives. Probably the most popular approach to the early detection of CVDs is the electrocardiogram (EKG), because it provides a precise but low cost and non-invasive measure of the state of the heart.

2.2 The heart. From mechanics to electrical signals

The heart is the muscle responsible to pump blood into the body. The heart produces electrical impulses at regular intervals that trigger a sequence of associated mechanical movements, as can be seen in the **Figure 3**. These impulses originate in the sinoatrial (SA) node, the natural pacemaker of the heart, approximately once a second. They propagate through the atria (upper chambers) and ventricles (lower pumping chambers) originating blood flow. The electric impulses in the heart represent the different stages of a heartbeat: rest (1), stimulation (2,3,4) and recovery (5, 6). The generation and conduction of the electrical impulse of a heartbeat gives rich information on the state of the heart muscle, so the analysis of the electrical activity of the heart can be used to detect abnormal heartbeats.

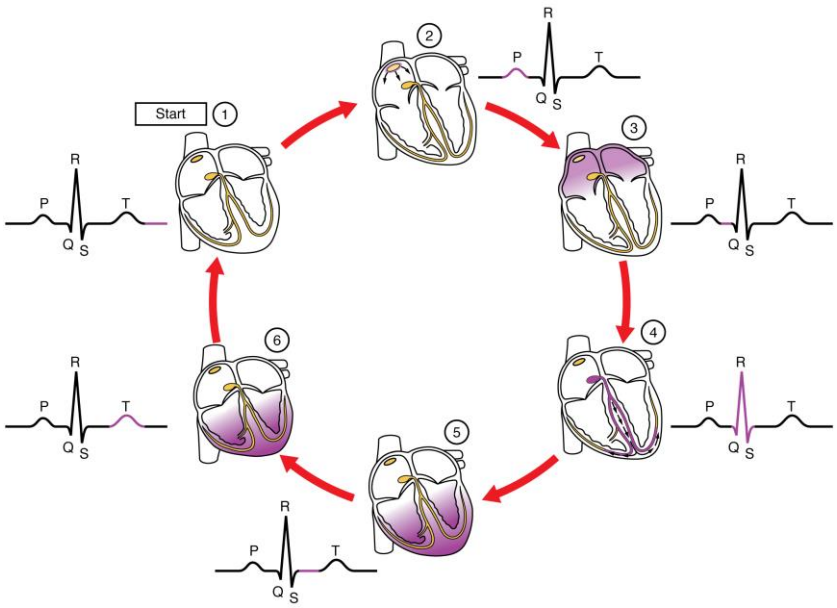


Figure 3. Physiological heartbeat sequence: (1) rest, (2) sinoatrial node (SA node) starts the electric impulse, (3) atrial contraction or diastole, (4) atrial relaxation or systole, (5) ventricular systole and (6) ventricular diastole [17].

2.3 The EKG: a key tool for the early diagnosis of CVDs

The EKG provides useful information about the heart's function, and its signal is registered by placing two electrodes in the body of the subject. The EKG is the time evolution of the electric impulses that stimulate the heart and produce its contraction, but recorded on the body's surface. A typical example of 10 s normal EKG activity is shown in **Figure 4**, in which the sequence described in **Figure 3** repeats at an approximate rate of 60 heartbeats per minute.

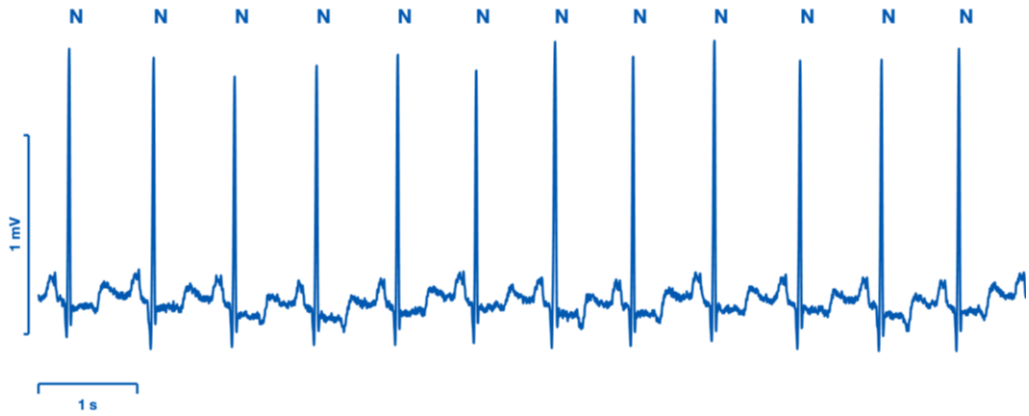


Figure 4: a 10 s interval of a normal/healthy 30 min EKG record, in which all beats are normal (N).

The EKG is a low-cost, non-invasive tool very well suited for either diagnosis or continuous monitoring of the patient. The waveform or signal shape of the EKG depends on where the electrodes are placed, and these different placements are called leads. Each lead picks up the same electrical activity of the heart, but from a different position. This permits to see the heart's electrical conduction system from many distinct angles.

As shown in **Figure 5**, different equipment is used for EKG acquisition such as 12-lead electrocardiograph (left), Holters (center) or defibrillators (right). The Holter monitor is a small outpatient electronic device that records and stores the patient's electrocardiogram for at least 24 hours. It is often used in patients with suspected cardiac arrhythmia [18]. A defibrillator is a medical device designed to analyse the heart rhythm, identify deadly arrhythmias and administer an electric shock in order to restore a healthy heart rhythm [19]. Holters are designed for the early detection of arrhythmia, while defibrillators are designed to prevent death in an emergency situation.

Figure 5: a 12-lead electrocardiograph (left) [20], a Holter (centre) [21], and a defibrillator (right).

2.4 Arrhythmia detection

In most cases cardiac arrhythmia are not lethal, but an early diagnosis of arrhythmia is convenient, either to remove the arrhythmia through surgery (ablation for instance) or to establish preventive measures to avoid side effects, like medication to avoid strokes. Abnormal beats are associated to arrhythmia or can trigger arrhythmia in the future, so a system based on the analysis of the EKG to classify heart beats would provide a very useful non-invasive diagnostic tool.

A typical example of abnormal beats are the ventricular ectopic beats (VEB) shown in **Figure 6**. These beats are spontaneously generated by active foci in the ventricles, and produce

an inefficient ventricular contraction, in which the heart cannot properly pump blood through the body. A succession of rapid VEB produces a ventricular tachycardia (VT) as shown in **Figure 7**. VT is a sign of heart dysfunction and is associated with a 38% of mortality within a year from its first appearance. Moreover, VT frequently degenerates into ventricular fibrillation (VF) which produces cardiac arrest [22]. VF is deadly if not treated with a defibrillator within minutes. VF triggers 85% of cardiac arrests [23], and roughly 10% of cardiac arrest patients survive [24]. As shown in **Figure 8**, during VF there is no effective ventricular contraction, the ventricles are fibrillating or quivering without order.

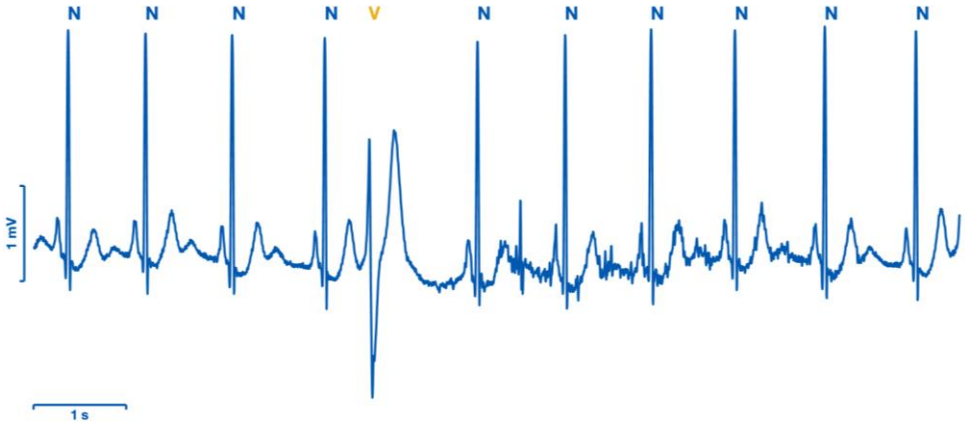


Figure 6: example of ventricular ectopic beat, labelled with V.

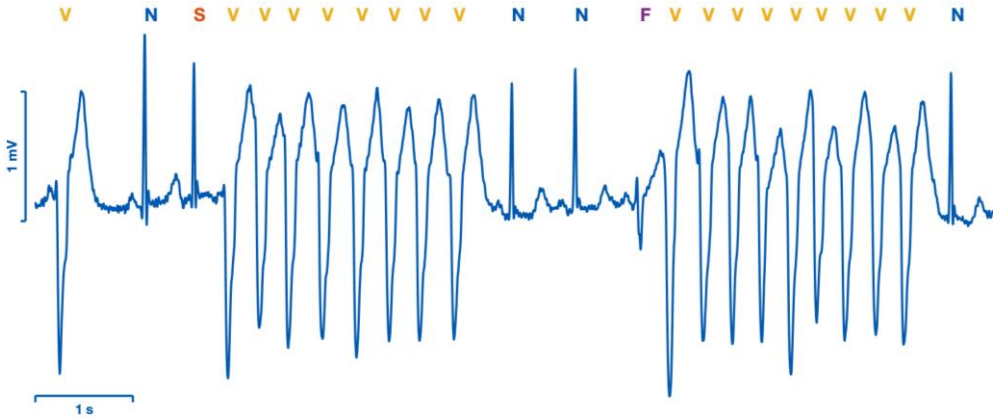


Figure 7: example of a burst of VEB, leading to ventricular tachycardia.

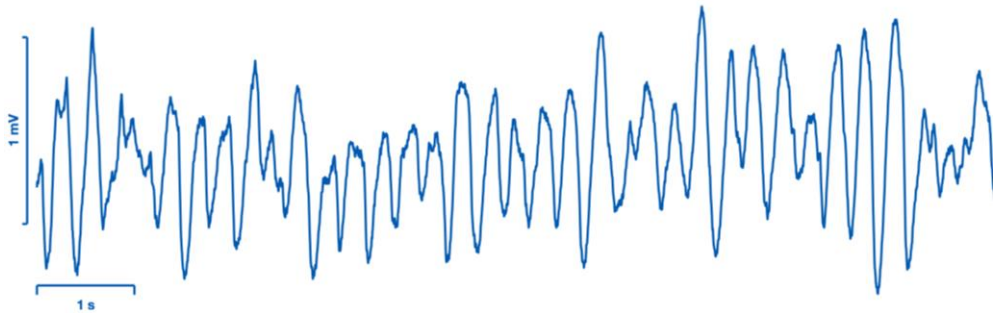


Figure 8: example of ventricular fibrillation. The EKG form is irregular, every beat is different.

Low cost and non-invasive methods or algorithms for the detection of ventricular beats would be of great value for the identification of subjects at risks of suffering a cardiac arrest. The EKG obtained from Holter devices provide a platform for the development of such algorithms.

3 OBJECTIVES AND SCOPE OF THE STUDY

The main objective of this bachelor's final degree project is to **develop an algorithm to discriminate between normal and ventricular heartbeats using the electrocardiogram.** In order to achieve this goal, the following secondary objectives have been defined:

1. To **obtain and annotate/revise an EKG database,** the database should be rich enough to contain a varied number of pre-annotated heartbeats. Pre-annotation is a requisite since these annotations are made by consensus among clinicians. Enrolling clinicians for annotation, and the time they would needed to annotate thousands of heartbeats is beyond the scope of this project.

   As a side objective, the revision of the data requires the creation of a graphical user interface (GUI) for the visualisation and handling of the EKG signals and their annotations.

2. The algorithm will be based on automatic tools to detect heartbeats and characterise their physiological variables of interest like durations or waveform amplitudes. So, another objective will be the **integration of standard EKG tools for the detection and delineation of heartbeats** into the project and the data revision GUI.

3. Once heartbeat variables of interest (features) are available, the third objective will be the **statistical characterisation of the features**, with the purpose of identifying distinctive patterns that will help differentiate normal from ventricular beats.

4. The next step will then be to apply statistical learning techniques to generate **models for the classification of the heartbeats**. These models will be based on basic machine learning techniques.

5. The final objective will be the **evaluation of the models**, and the generation of the final estimates of how the algorithm will classify heartbeats.

4 BENEFITS

The main outcome of this project will be a heartbeat classification algorithm using the EKG. And a secondary, but also important, deliverable will be a GUI. This will be used for the visualization and revision of the EKG signals; also, for the heartbeat annotations and features (like QRS duration for instance). The successful completion of these objectives will have the following main benefits:

- The algorithm is a first step towards the automatic classification of heartbeats, and tackles the discrimination of the two most important types of beats, namely normal and ventricular beats. Its potential lies in that it will set the framework for future developments of more complete heartbeat classification systems, and of algorithms and procedures to diagnose complex arrhythmia.
- The development of easy to use GUI for the semi-automatic annotation/revision of EKG data can be deployed in the future in the generation of larger datasets. Such datasets could then be enhanced to boost the accuracy of future heartbeat detection and classification algorithms, and of the arrhythmia detection algorithms based on them.

4.1 Technical benefits

The completion of the objectives of the project will produce several technical benefits related to the research conducted by the BioRes group. First, it will consolidate the use of standard EKG tools like heartbeat detection and delineation algorithms. These tools can then be used and adapted for the databases used by the BioRes group, which contain cardiac arrest cases monitored with defibrillators. The EKGs obtained during cardiac arrest are substantially different from those seen in Holter recordings, because this equipment have different acquisition characteristics (bandwidth, resolutions, ...) but mainly because the conditions of the patients are different.

However, a through characterisation of how these tools work on non-cardiac arrest patients will be a first step towards understanding how these tools should be used for the EKG seen during cardiac arrest. This will produce important technical advances for BioRes, that will use those tools in the future in their research projects.

4.2 Social benefits

As shown in **2.4**, CVDs are the most important cause of mortality. The developments of low cost, non-invasive and general-purpose tools for the early detection and prevention of CVDs would yield important benefits for the population. It can contribute to prevent potentially lethal complications derived from untreated or unmonitored heart dysfunctions.

In this project we will develop an algorithm to distinguish normal from ventricular heartbeats. This algorithm can therefore contribute to identify ventricular dysfunctions in the heart, and to early detection of CVDs derived thereof. This could contribute to advance effective treatments and prevent future fatalities.

4.3 Economic benefits

As mentioned in **2.1**, the cost of late detection of CVDs is extremely high. With this method, it may be possible to detect ventricular dysfunctions earlier. This could lead to low cost interventions in the patient that avoid costly hospitalisations or even surgery, which may be inevitable when an untreated and unmonitored CVD produces a life-threatening event like a stroke or heart failure. Moreover, since the presented algorithm works with EKG signals, it is a

very cheap solution when compared to more advanced and costly techniques like cardiac imaging based on echocardiography, for instance.

In addition, the development of user-friendly tools like the EKG visualisation and annotation GUI will speed up the time needed to construct annotated datasets in the future. These datasets can be used to improve these types of solutions, since the algorithms improve as more data are available, to infer the statistical patterns to differentiate the targeted conditions (ventricular and normal beats). Also, thanks to the heartbeat detector algorithm, diagnostic errors will be avoided in noisy EKG scenarios. Both reducing errors and shortening analysis times will result in improved productivity.

In summary, using this method will help on the early detection of CVDs. Consequently, it will be possible to design interventions or preventive measures earlier, minimising their effects on the patient's health and reducing the associated costs.

5 State of the Art

This section contains a summary of the most important topics relevant for the development of this project. It starts by reviewing the typical EKG morphology of a heartbeat and its most influential changes, which result in the different types of heartbeats. Then we review the main signal processing algorithms that allow the automatic detection and characterisation of those heartbeats, with the objective of the creation of an automatic algorithm to classify the heartbeat types based on the EKG morphology.

5.1 Normal heartbeat in the EKG

Each normal heartbeat is reflected as a new cycle on the patient's EKG signal. One typical cycle and its constituent waves and intervals is shown in **Figure 9**. The cycle represents the succession of two different processes: the atrial depolarisation/repolarisation and the posterior ventricular depolarisation/repolarisation. Therefore, the EKG of a normal heartbeat has the following characteristics [25][26]:

- **P-wave:** represents the depolarisation and contraction of the atria.
- **PQ segment:** ~0.1 s pause in the atrioventricular (AV) node to let the blood flow from the atria to the ventricles.
- **PQ interval:** time interval for the depolarisation/repolarisation of the atria. Atrial repolarisation is not visible as is masked by ventricular depolarisation.
- **QRS complex:** ventricular depolarisation, formed by three waves:
  - Q-wave. The beginning of the QRS complex and its first inferior deflection.
  - R-wave. First superior deflection, which is larger than the P-wave because ventricular activity is predominant over atrial activity.
  - S-wave. The inferior deflection during ventricular depolarisation.
- **ST segment:** the pause after the QRS complex. There is no mechanical activity here.
- **T-wave:** the repolarisation of the ventricles so they can be stimulated in the following heartbeat.
- **QT interval:** refers to the time interval for the depolarisation/repolarisation of the ventricles. This happens simultaneously for both ventricles.

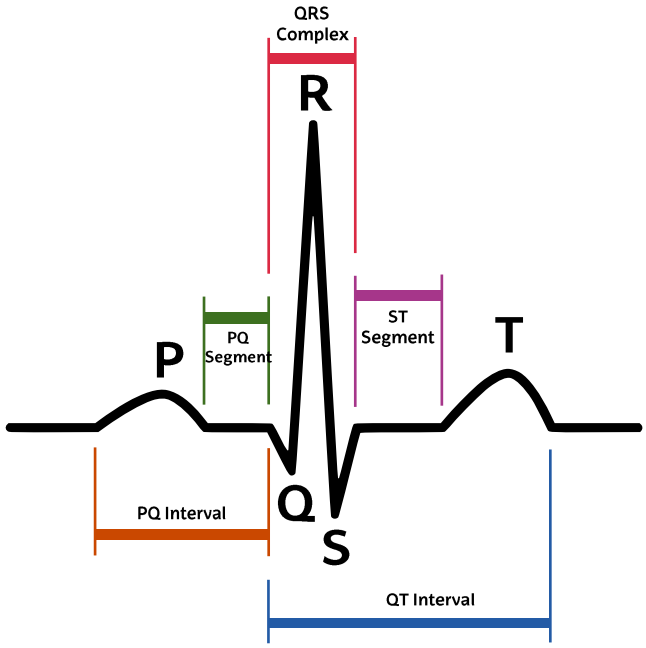


Figure 9: a normal heartbeat's EKG [27] , and its most important waves and intervals.

5.2 Types of heartbeats

Heartbeats may be very different from the normal/typical example of **Figure 9**. These differences are the result of abnormalities in the electrical conduction system of the heart; therefore, their identification is important for the early detection of CVDs. These differences can be of many types, but the most clinically important ones are:

- **P-wave existence:** sometimes the heartbeats instead of starting in the SA node, start directly from the ventricles, so there is no auricular depolarisation (no P-wave).
- **QRS morphology:** when a heartbeat starts in one ventricle rather in the SA node, that ventricle depolarises before the other; thus, the QRS deflections are bigger than the ones from a normal heartbeat, and have longer durations since both ventricles do not depolarise simultaneously.
- **T-wave existence:** at times, the QRS complex overlaps the precedent T-wave. This may unchain serious arrhythmias.

These changes produce different types of heartbeats that differ greatly in morphology from normal heartbeats. Broadly speaking, the main heartbeat types can be classified into the 5 classes shown in **Figure 10**. These classes are normal beats (N), supraventricular ectopic beats (S) with no P wave, ventricular ectopic beats (V) originated in the ventricles, fusion beats (F) originated from several sources ventricular or atrial, and unknown beats (Q), which comprise a wide category of beats of unknown sources such as beats originated from a pacemaker.

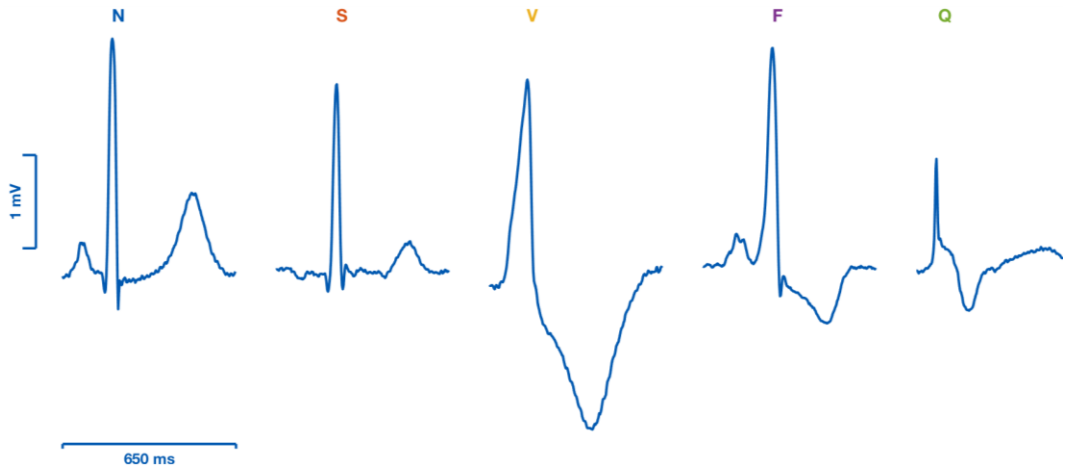


Figure 10: from left to right, normal (N), supraventricular ectopic beat (S), ventricular ectopic beat(V), fusion beat (F) and unknown beat (Q).

From a medical perspective the heartbeat class grouping shown in **Figure 10** corresponds to many conduction abnormalities. It is not the objective of the project to review all the possible abnormalities. The Association for the Advancement of Medical Instrumentation (AAMI) is a non-profit organisation founded in 1967 and is the main source of international standards for the healthcare devices industry [28]. AAMI defined the grouping shown in **Figure 10** starting from a more detailed classification of conduction abnormalities. The exact mapping proposed by the AAMI is shown in **Table 1** as reference.

Table 1: AAMI heartbeat classification, detailed conduction abnormalities for the five main types of heartbeats.

N	S	V	F	Q
Any heartbeat not in the S, V, F or Q classes	Supraventricular ectopic beat	Ventricular ectopic beat	Fusion beat	Unknown beat
Normal beat (NOR)	Atrial premature beat (AP)	Premature ventricular contraction (PVC)	Fusion of ventricular and normal beat (fVN)	Paced beat (P)
Left bundle branch block beat (LBBB)	Aberrated atrial premature beat (aAp)	Ventricular escape beat (VE)		Fusion of paced and normal beat (fPN)
Right bundle branch block beat (RBBB)	Nodal (junctional) premature beat (NP)			Unclassified beat (U)
Atrial escape beats (AE)	Supraventricular premature beat (SP)			
Nodal (junctional) escape beat (NE)				

From an algorithmic point of view, the AAMI standard simplifies the classification of heartbeats, since multiple types of abnormalities are now grouped into five broad categories. This is important in applications in which there is a limited number of EKG leads (one or two) like in this project. Some of the very detailed abnormalities can only be identified using 12-lead EKGs.

5.3 Automatic EKG algorithms for heartbeat classification

Given the importance of having well-characterised EKG signals to facilitate the further identification of CVDs, numerous studies have been conducted for the classification of heartbeats in recent years. Each of these studies has tried to bring its own approach to the problem.

One of the most commonly accepted approaches is the three-stage approach described in **Figure 11**, which is comprised of: a heartbeat detection stage, heartbeat waveform delineation to obtain the beat's characteristics, and classification based on heartbeat characteristics. The following sections introduce these three topics.

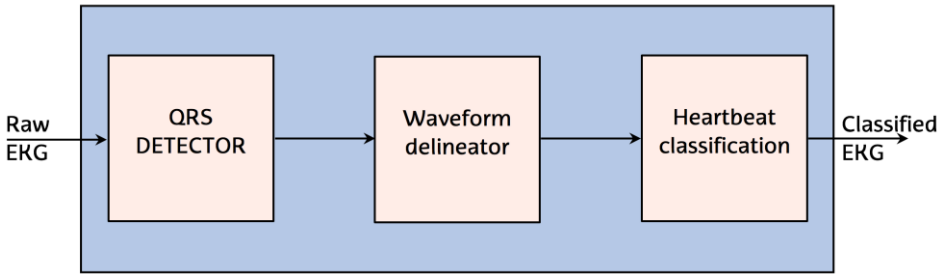


Figure 11: general scheme for the classification of heartbeat. The approach is based on three types of algorithms. First the detection of the heartbeats (QRS detector), then the characterisation of the heartbeat (waveform delineator), and finally the classification of the heartbeat.

The following sections introduce these three topics.

5.3.1 QRS detection algorithms

The automatic detection of heartbeats in the EKG is known as QRS detection, since the QRS complex is the most salient waveform in the heartbeat The QRS complex is therefore the easiest to detect wave or complex (see **Figure 9)**. QRS detection is the first step towards the heartbeat classification, since any heartbeat classification algorithm must first identify when a heartbeat occurs.

Many signal processing algorithms have been proposed for QRS detection, Kohler et al [29] provide an excellent review and introduction to the topic. Some of those approaches include the derivative of the signal (QRS is the EKG interval with largest slopes), wavelet-based algorithms, adaptive filters or methods based on the Hilbert transform. However, all of them follow a common structure, which is shown in **Figure 12.**

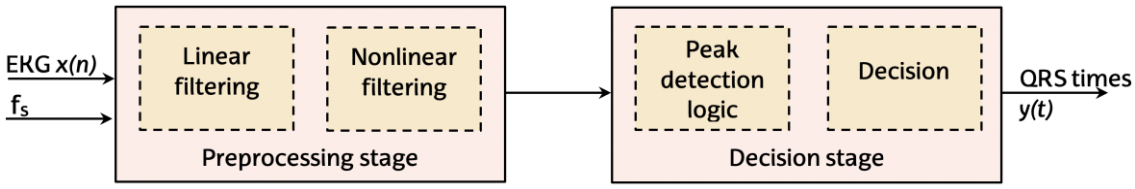


Figure 12: common structure of the QRS detectors, adapted from [29].

The algorithms first preprocess the EKG to remove sources of noise like movement, respiration or power-line interferences. Then, a peak enhancement and detection algorithm is used, and a decision logic (adaptive or fixed) is applied to identify peaks corresponding to actual heartbeats. The accuracy of the algorithms is measured by comparing the QRS detections of the algorithm to the real positions of heartbeats (marked by specialists) on well-defined EKG databases. In that way, it is possible to determine the following occurrences:

- TP, true positives. These are detected QRS that really are QRS.
- FN, false negatives. These are undetected QRS that actually are QRS.
- FP, false positives. These are detected QRS that are not QRS.

The standard performance metrics to measure the goodness of the QRS detector are then the sensitivity (Se) and positive predictive value (PPV) defined as:

$$Se = \frac{TP}{TP + FN} \quad (1)$$

$$PPV = \frac{TP}{TP + FP} \quad (2)$$

The sensitivity measures to what existing present heartbeats are detected [30], while the PPV measures how confident we can be on the detection done by the algorithm [31]. Hence higher Se and PPV will mean a better QRS detection algorithm. Nevertheless, to know if the obtained results are reliable, the algorithm has to be tested and compared to other algorithms using standard databases for QRS detection. The Physionet platform provides several open and reliably annotated databases for this purpose [32].

5.3.2 Waveform segmentation algorithms

Once the beats are correctly located with the QRS detection algorithm, they can be delineated to extract its main characteristics. The typical waveforms and intervals to be detected are the ones shown in **Figure 13**. Delineation is the determination of peaks and limits of individual QRS waves, P and T waves. In general, these algorithms start from a previous QRS location and set a temporary search window to find the waves around the next QRS location point, identifying the relevant points shown in **Figure 13**. After defining the search window, several techniques can be used to identify the start and end of each of the relevant waves. Some

approaches have made use of continuous and discrete wavelet transforms [33–35], signal envelope techniques [36], second order derivatives [37], or low-pass differentiation [38–40].

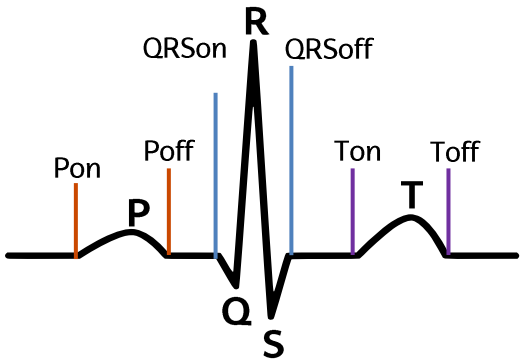


Figure 13: a delineated normal heartbeat's EKG.

The waveform segmentation algorithms return an estimation of the time points and values [41] listed in **Table 2**.

Table 2: characteristic time points of a heartbeat's EKG.

Time point	Description
Pon	indicates where the P-wave should start.
P	indicates where the P-wave peak should be
Poff	indicates where the P-wave should end.
QRSon	indicates where the QRS signal should start
Q	indicates where the Q-wave peak should be
R	indicates where the R-wave peak should be
QRS	indicates where the QRS signal should be centred. Coincides with R-wave peak position
S	indicates where the S-wave should be centred
QRSoff	indicates where the QRS signal should finish
Ton	indicates where the T-wave should start
T	indicates where the T-wave peak should be
Ttype	classifies the heartbeat's T signal type
Toff	indicates where the T-wave should finish

Once the characteristic timepoints are identified, the calculation of the standard EKG time intervals or waveform characteristics like P, QRS, and T wave duration and amplitude is straight forward. These characteristics are essential for heartbeat classification, because different beat morphologies present statistically significant differences in the duration and amplitude (even presence) of these waves.

5.4 Heartbeat classification

When heartbeats have been detected and its waveforms and characteristic values calculated, the final step is to classify the heartbeat. Although the AAMI proposes five broad categories, this project will focus on the two most important classes, normal beats (N) and ventricular beats (V). Ventricular beats are particularly important because they can precede or be an indication of future ventricular arrhythmia.

So, for the purposes of this project, the classification problem is restricted to two-classes or a binary classification problem. Classification problems and their evaluation are better understood in terms of confusion matrices, which for binary problems take the form shown in **Figure 14**. The confusion matrix compares the output of the binary classifier to the labels assigned to the beats by expert cardiologists. In binary problems, there is a positive class (presence of disease, V beats) and a negative class (absence of disease, N beats). This comparison produces TP, FP and FN as defined for QRS detectors, and in this case also true negatives (TN), since the absence of disease (N beats) can be identified. Contrary to QRS detectors, here a beat is always either ventricular or normal (in a QRS detector there are no true negatives since absence of beats is not detected).

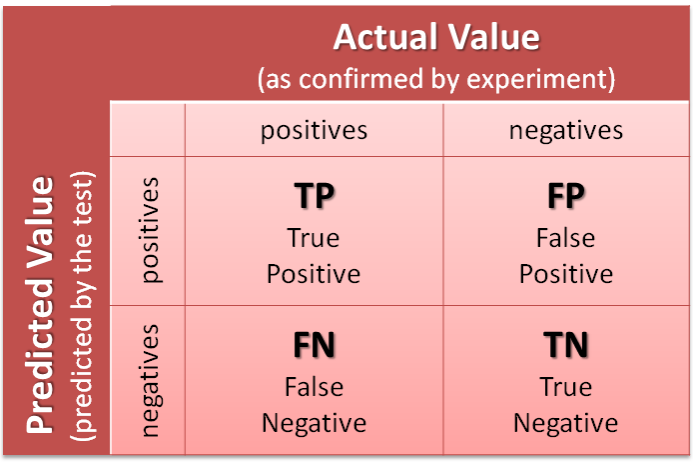


Figure 14: confusion matrix of a binary classification problem [42].

Based on this matrix the relevant performance metrics are now equations **28**- **30**.

$$Sp = \frac{TN}{TN + FP} \tag{3}$$

$$NPV = \frac{TN}{TN + FN} \tag{4}$$

Since TNs are now present, two new performance metrics can be defined. On the one hand, the specificity (Sp) measures the proportion of actual negatives that are correctly identified as such [30], [43]. On the other hand, the negative predictive value (NPV) measures the confidence that a negative detection actually corresponds to a negative [44].

Machine learning is a subfield of statistical learning that infers patterns from data that can be used for instance in classification problems. In our case, machine learning can be used to learn the characteristic values of the EKGs waves and intervals for ventricular and normal beats, and then use that information to automatically classify those beats. There are many classification algorithms that range from simple linear techniques like logistic regression, to complex algorithms like support vector machines (SVM), random forest (RF) or neural networks (NN) [45-46].

Given a set of n characteristics $(X_1, X_2, \dots, X_n)$ for an instance Y, the objective of the classification algorithm is to produce $\hat{Y} = f(X_1, X_2, \dots, X_n)$, an accurate estimate of the value of that instance. In our case $Y = \{0,1\}$ has two possible values, $0 = N$ and $1 = V$, and the characteristics X_i are the beat's characteristics obtained using a delineator. The objective is to obtain the function $f(\cdot)$ that will produce the best estimates of Y. A visual representation of the function $f(\cdot)$ for a problem with two characteristics is shown in **Figure 15**, in which a linear decision boundary was obtained.

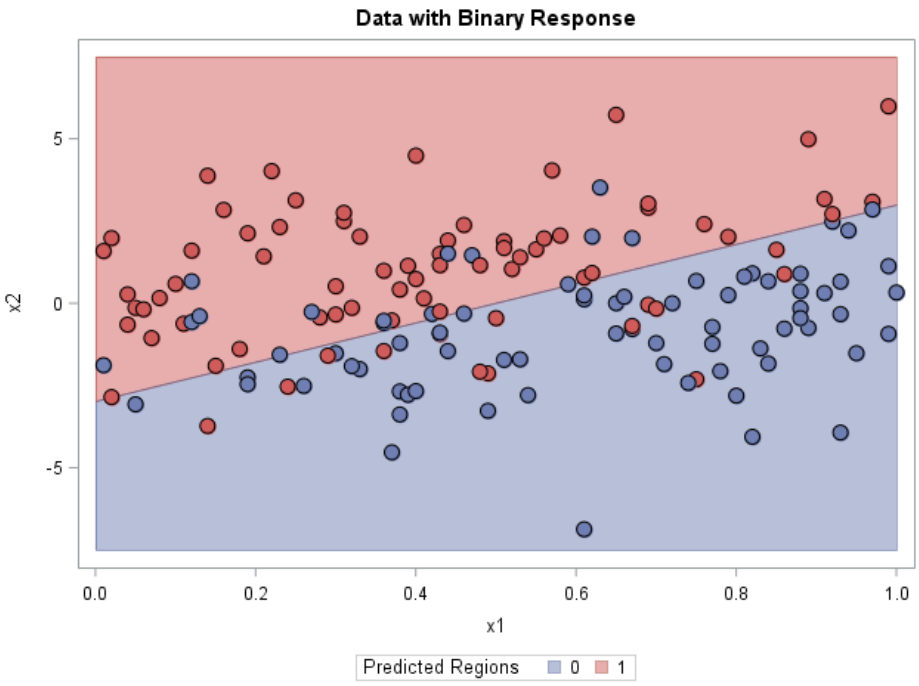


Figure 15: an example of a classifier with a linear decision boundary.

6 Analysis of alternatives

In order to achieve the objectives mentioned in **Section 3** in the most efficient way, this chapter examines the different alternatives that have been considered in the development of the project.

In the first place, the most suitable software suite was chosen for the programming of the different algorithms and the development of tools to view the EKG and the heartbeat annotations. Secondly, the existing alternatives for the automatic QRS detector were analysed. Thirdly, we reviewed the different waveform delineators. Finally, the alternatives for the machine learning techniques for the classification of beats from the electrocardiogram are analysed.

6.1 Software suite

For the software suite, four possible options for the development of the project were considered: MATLAB, Octave, C programming language and Python. The characteristics and distinctions of each will be explained below.

6.1.1 MATLAB

The MATLAB software suite is a mathematical software tool integrated in an integrated development environment (IDE). MATLAB has its own programming language (M language).

As an interpreted language, it offers a wide range of facilities to the user, even if the user is not an expert. However, this also affects execution speed, because execution times are slower for interpreted languages than for compiled languages.

MATLAB offers very useful tools for the user, such as toolboxes, in which the developed applications and functions can be found for tools ranging from signal processing to machine learning. It also has a comprehensive help guide [47] and a technical support website [48]. Another advantage of the MATLAB software suite is its high-quality graphics. In addition, the development of Graphical User Interfaces (GUI) will play an important role in this project. MATLAB offers an advanced and efficient environment for developing GUIs, called GUIDE.

The major disadvantage of MATLAB is its price. The license price for a single user is about 700 € and the price of each toolbox is 200 € [49]. Bearing in mind that it is necessary to use of at least three toolboxes, the total price for the user of the MATLAB platform could reach 1300 €.

6.1.2 OCTAVE

GNU Octave is a high-level language, similar to MATLAB and compatible with it but also independent. Despite the many similarities, there are many differences that should be considered [50]. The main advantage of this software suite is that is free open source [51].

Besides the differences in programming language, there are also disparities in the available resources, i.e. the essential tools needed for the development of GUIs are quite limited in Octave.

Also, the help guide is not as specific and easy to use and the technical support is not as comprehensive, so that problem solving can be made more difficult.

6.1.3 C Programming language

The C programming language is aimed to the implementation of operating systems and is widely used to create applications and software systems.

The main advantage of the C language is its fast runtime. Compared to the interpreted languages, MATLAB and Octave, it is more effective. However, it does not offer facilities for handling matrixes and consequently, it's not as easy to work with large signal databases as in MATLAB and Octave.

6.1.4 Python

Python is the language of choice in machine learning projects, with many available libraries. In addition, it is free, easier than C programming language and allows the creation of GUIs.

Although Python is very extended and provides all the tools necessary for the development of the project, one of its main disadvantages is the learning curve for the development of the project. In fact, Python is not taught along the bachelor's degree in telecommunications engineering.

6.1.5 Software suite selection criteria

6.1.5.1 Ease of use

The complexity of the software is identified as an important feature.

6.1.5.2 Algorithm development time

It is important to develop and implement algorithms in an easy and effective way, so that the development of the algorithms does not delay the project.

6.1.5.3 Organisation and data visualisation

This project will work with large amounts EKG data and heartbeat annotations to produce statistically reliable algorithms. Consequently, the programming language must allow and easy and flexible management of large sets of data, and the possibility to easily and dynamically visualise the data.

6.1.5.4 Learning curve

This project is a Bachelor's Final Degree Project. In consequence, the hours available for learning new tools as programming languages are very limited.

6.1.5.5 Computation time

The runtime of the algorithms is important in real-time applications; however, the objective of this project is to propose working solutions that can be implemented in efficient programming languages in the future.

6.1.5.6 Price

The price of the program is another factor to consider.

6.1.5.7 Final decision

From a weighted assessment of the above parameters, we concluded that the software suite best fitted our needs was MATLAB. A detailed disaggregation of the ponderations is presented in **Table 3**.

Table 3: software suite selection criteria breakdown.

Criteria	Weight	MATLAB	Octave	C	Python
Ease of use	2/10	2/10	2/10	0.5/10	2/10
Algorithm development time	2/10	2/10	1.5/10	0.5/10	2/10
Organisation and data visualisation	2/10	2/10	1.5/10	0.5/10	2/10
Learning curve	2/10	2/10	1.5/10	2/10	0.5/10
Computation time	1/10	1/10	1/10	2/10	1/10
Price	1/10	0/10	1/10	1/10	1/10
Total	10/10	**9/10**	8.5/10	6.5/10	8.5/10

6.2 EKG database

6.2.1 MIT-BIH Arrhythmia database

The MIT-BIH (Massachusetts Institute of Technology-Beth Israel Hospital) arrhythmia database consists of 48 two-channel EKG recordings of an approximate duration of 30 minutes each. The signals were digitalised at 360 samples per second per channel and the resolution for digitisation is 11-bit over a 10-mV range [52].

One of the major advantages of the MIT-BIH database is that it is the most spread database for EKG signal analysis, and that it comes with annotated heartbeats which contain beat labels in the format of the AAMI standard.

6.2.2 CU Ventricular Tachyarrhythmia database

The Creighton University (CU) Ventricular Tachyarrhythmia database is compounded of 35 single-channel EKG recordings of about eight-minutes each. Its sampling frequency is 250 Hz and the resolution for digitisation is 12-bit over a 10-V range (10mV nominal) [53].

Its advantages are that it contains recording from patients suffering from ventricular tachycardia, ventricular flutter, and ventricular fibrillation, so it could be useful for ventricular heartbeat classification. However, heartbeats are not identified or annotated.

6.2.3 AHA database

The American Heart Association (AHA) database comprises 80 two-channel EKG recordings. The sampling frequency used for this database was 250 Hz per channel and the resolution for digitisation is 12-bit over a 10-mV range.

The greatest advantage of the AHA database is that contains EKG recordings for ventricular arrhythmia and that, each record of AHA's long version, has a duration of 2.5 h.

However, its major disadvantage is that it is not available for the public, and its cost is high. Furthermore, the heartbeat annotations for each recording are incomplete.

6.2.4 Database selection criteria

6.2.4.1 Accessibility

This is the most important point, because we need a widely spread and reliable database to work with.

6.2.4.2 Annotations

The annotations from the cardiologists are essential to use them as ground truth marks to train our classifier. Annotating beats is a costly process and requires involving external specialists in the project.

6.2.4.3 Number of recordings

With a bigger database, the better results we will get, because our classifier will have more different data to train with.

6.2.4.4 Final decision

From a weighted assessment of the above parameters, we concluded that the database that best fitted our needs was the MIT-BIH Arrhythmia database. A detailed disaggregation of the ponderations is presented in **Table 4.**

Table 4: database selection criteria breakdown.

Criteria	Weight	MIT-BIH	CU	AHA
Accessibility	4/10	4/10	4/10	1/10
Annotations	4/10	3/10	1/10	1/10
Number of recordings	2/10	1/10	1/10	2/10
Total	10/10	**8/10**	6/10	4/10

6.3 QRS detection algorithm

Regarding QRS detection algorithms, several alternatives are proposed: combined methods (Hamilton-Tompkins), open algorithms available from the Physionet platform (SQRS, WQRS), and methods based on advanced signal processing.

6.3.1 Hamilton-Tompkins algorithm

One widely used QRS detection algorithm is the Hamilton-Tompkins (HT) algorithm, which is a combined method based on the EKG slope, squared and averaged by an integrator, with adaptive amplitude and noise thresholds. **Figure 16** shows the block diagram of the HT QRS detector.

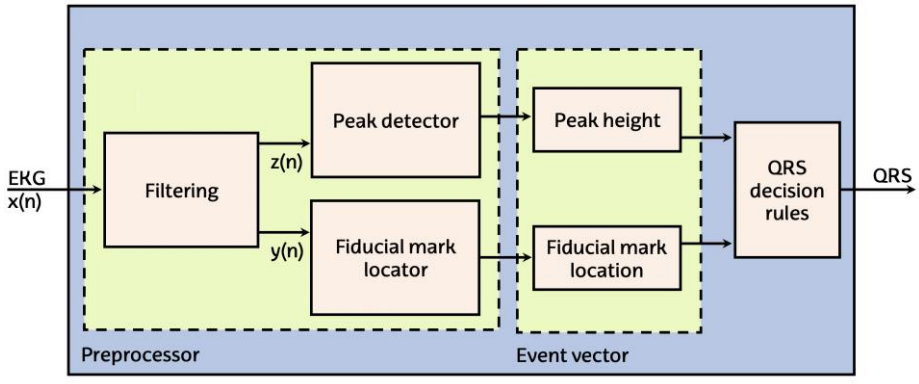


Figure 16: diagram of HT QRS detector algorithm, adapted from [6].

When Hamilton and Tompkins applied their QRS detector to the MIT-BIH database, the Se and PPV of the algorithm were 99.69% and 99.77%, respectively [6].

In terms of programming time, this algorithm is implemented in MATLAB by the BioRes Research group, which saves a lot of time and could be therefore deployed as a tool directly.

6.3.2 Physionet algorithms (SQRS and WQRS)

The PhysioToolkit suite contains numerous tools and algorithms for physiological signal processing, including two free QRS detectors: SQRS and WQRS.

The SQRS method uses the characteristic steep slope of the QRS complex for its detection [54–57]. The EKG signal is sampled at the configurable value of 250 Hz, then low pass filtered and finally its derivative is calculated by applying the first difference. After this, using the slope of that signal, the QRS complex is detected.

On the other hand, the WQRS method (based on the length of the signal) is divided into three different parts, as can be seen in the **Figure 17**.

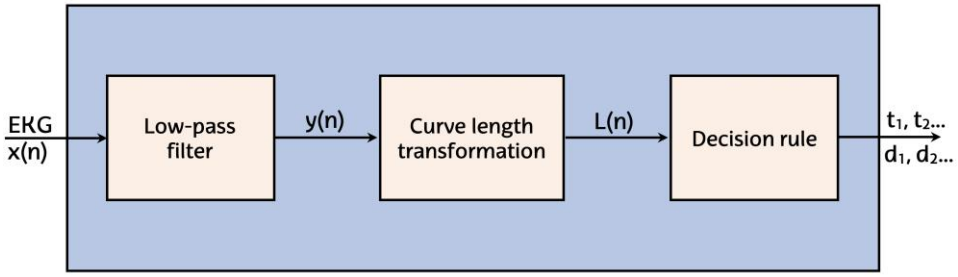


Figure 17: common structure of WQRS algorithm.

The WQRS algorithm has been applied to the EKG signals of the MIT-BIH database, achieving a Se and PPV values of 99.65% and 99.77%, respectively [58]. This means that this algorithm has a very high accuracy. In addition, it is a widely tested algorithm.

The main disadvantage of the WQRS algorithm is that, as it is developed in Java and the code is not open, wrappers must be used to make calls to the algorithm from MATLAB, which makes it difficult to use the algorithm. There are currently no versions of WQRS in MATLAB.

6.3.3 Methods based on the advanced signal processing

There are also methods based on advanced signal processing, for example wavelets and rules for determining QRS. In Köhler [29] a detailed review of QRS can be found. In this section, a robust QRS algorithm is analysed against noise [59]. This is a technique for the detection of QRS complexes in electrocardiographic signals that are is based on a characteristic obtained by counting the number of zero crosses per segment.

It is well known that zero crossing methods are robust against noise and are particularly useful for finite precision arithmetic. This detection method includes this robustness and provides a high accuracy even in cases of signals of very noisy electrocardiogram. In addition, due to the simplicity of detecting and counting zero crosses, it provides a computationally efficient solution to the problem of QRS detection. The excellent performance of the algorithm is confirmed by a Se of 99.70 % and a PPV of 99.57% against the MIT-BIH database.

6.3.4 QRS detection algorithm selection criteria

6.3.4.1 Precision

The accuracy of the algorithm is a very important criteria to be taken into account, because the methods of automatic heartbeat classification using the EKG are based on the correct detection of QRS complexes.

6.3.4.2 Development time

The time stamps returned by the algorithms do not occur at the peak of the R wave, but a little after. However, for this project we are interested in marking the precise instant of the R wave. Hence, we will have to develop patches to correct the marks.

6.3.4.3 Accessibility

It is also very important to have MATLAB versions of the algorithm, to reduce the time needed to implement it in the tools developed in this project.

6.3.4.4 Final decision

From a weighted assessment of the above parameters, we concluded that the QRS detection algorithm that best fitted our needs was the one developed by HT algorithm [6]. A detailed disaggregation of the ponderations is presented in **Table 5.**

Table 5: QRS detection algorithm selection criteria breakdown.

Criteria	Weight	Physionet	HT	Advanced signal processing
Precision	4/10	3.7/10	3.75/10	3.65/10
Development time	3/10	1/10	2/10	1/10
Accessibility	3/10	1/10	3/10	1/10
Total	10/10	5.7/10	**8.75/10**	5.65/10

6.4 EKG delineation algorithm

6.4.1 Wavedec algorithm

The Wavedec algorithm is an EKG delineation algorithm base on the wavelet decomposition of the EKG introduced by Martinez et al [41]. The wavelet decomposition analyses the signal in time and frequency by decomposing the signal in non-overlapping frequency sub-bands. The characteristic waves of the EKG occupy different frequency bands, and this property is used by wavelet-based algorithms to identify the waves and its characteristic time-points.

6.4.2 Low-pass differentiation

Low-pass differentiation (LPD) algorithms are widely used for noisy EKG signals segmentation, such as the ones recorded by the Holters. The EKGs obtained by Holters generally present a low signal-to-noise ratio (SNR), related to muscular activity and variations in the electrode to skin contact.

This kind of algorithms is used to return the Q-T interval [40], instead of all the characteristic points of a heartbeat's EKG, so their functionality is limited when compared to the Wavedec algorithm.

6.4.3 Second order derivatives

Algorithms based on second order derivatives (2^{nd} derivatives) are often used to map ventricular arrhythmias. As seen in **Figure 7**, the beats of ventricular arrhythmias usually consist of just abnormal QRS complexes. Therefore, these algorithms do not return the points described in section **5.3.2**, but the points where the QRS begins (onset) and ends (offset).

6.4.4 EKG delineation algorithm selection criteria

6.4.4.1 Accessibility

One important point is to be able to get the algorithm easily so that it can be implemented as soon as possible.

6.4.4.2 Accuracy

It is essential for this work to accurately measure the characteristic points of the EKG signal. These characteristics and others derived from them will be used to classify the beats, so they must be accurate estimates of the characteristics of the heartbeat.

6.4.4.3 Computation time

We are also interested in a fast operation time of the algorithm. However, as this project is not conceived to develop a real time application, this is not the most determining factor.

6.4.4.4 Final decision

From a weighted assessment of the above parameters, we concluded that the waveform segmentation algorithm that best fitted our needs was the Wavedec algorithm. A detailed disaggregation of the ponderations is presented in **Table 6.**

Table 6: EKG delineation algorithm selection criteria breakdown.

Criteria	Weight	WC	LPD	2nd derivatives
Accessibility	4/10	4/10	2/10	2/10
Accuracy	4/10	3/10	1/10	1/10
Computation time	2/10	1/10	2/10	2/10
Total	10/10	**8/10**	5/10	5/10

6.5 Machine Learning classifier

The objective of the project is to develop a heartbeat classification algorithm to differentiate normal and ventricular heartbeats. Heartbeats will be mapped to a set of characteristics (features) fed to a classification algorithm. Several classification algorithms are available, including Logistic Regression (LE), Support Vector Machine (SVM) or Random Forest (RF) classifiers.

6.5.1 Logistic Regression

Logistic regression models produce a linear decision boundary, which is simply a linear combination of the characteristics that optimally separates the two groups. The algorithm estimates the coefficients of the linear combination. Linear combinations are reasonably easy to train (learn from data), and easily interpretable. The contribution of each characteristic to the decisions of the algorithm is related to the coefficient of each characteristic.

6.5.2 Support Vector Machine

SVMs are advanced classification algorithms in which a non-linear decision boundary is determined. SVMs are therefore very flexible and can accurately represent complex decision boundaries. Nonlinear decision boundaries are obtained using non-linear transformations of the data (for instance gaussian or radial basis functions) to go from complex representations to linear representations in higher dimensional spaces. A set of vectors is then determined to obtain a decision boundary with a decision margin, a maximum margin separation [60]. One of the disadvantages of SVMs is interpretability since the non-linear transformation disguises the contribution of each characteristic to the decision of the classifier.

6.5.3 Random Tree Forest

These algorithms are based on the idea that is possible to fit many weak (inaccurate) classifiers that independently decide. When the decisions of these classifiers are aggregated, by a majority vote, the classifier becomes robust and very accurate. In particular, random forest (RF) classifiers are based on the aggregation of hundreds to thousands of decision trees. These trees are trained using random selections of data and characteristics during training, producing uncorrelated trees.

RF classifiers are one of the most accurate machine learning algorithms. RF algorithms are hard to train because they have several configurable parameters, they are also hard to interpret because the contribution of each variable to the decision is hidden among many trees.

6.5.4 Machine Learning classifier selection criteria

The model that we choose will have to be one that is accurate but interpretable, since one of the important characteristics of the proposed solution is that is based on physiologically meaningful characteristics of the EKG.

6.5.4.1 Ease of training

Considering that we are new to this topic (ML classifiers) and the bachelor's final degree project has an established deadline, we have to look for easy to train models. In addition, the used MIT-BIH Arrhythmia database is limited.

6.5.4.2 Interpretability

Another important aspect is the interpretability of the solution, so that decisions can be traced back to meaningful EKG characteristics.

6.5.4.3 Accuracy

It is important to that the classifier is accurate in order to achieve reliable results.

6.5.4.4 Final decision

From a weighted assessment of the above parameters, we concluded that the ML classification algorithm that best fitted our needs was the Logistic Regression. A detailed disaggregation of the ponderations is presented in **Table 7.**

Table 7: ML classifier selection criteria breakdown.

Criteria	Weight	LR	SVM	RTF
Ease of training	4/10	4/10	3/10	2/10
Interpretability	4/10	3/10	2/10	1/10
Accuracy	2/10	1/10	1.5/10	2/10
Total	10/10	**8/10**	6.5/10	5/10

7 Description of the solution

In order to effectively carry out and manage this project, its development has been divided in several stages. First, we adapted the MIT-BIH Arrhythmia database to a compatible MATLAB format and the AAMI standards. Second, we detected the QRS complexes and delineated the heartbeats to obtain their features. Third, the most significant features were selected and used to classify the beats into normal and ventricular beats, using a logistic regression classifier. Finally, we evaluated and interpreted the results.

7.1 General architecture

The overall scheme with the main stages of the development of the project is shown in **Figure 18.** The following sections describe the details of the building blocks shown in **Figure 18.**

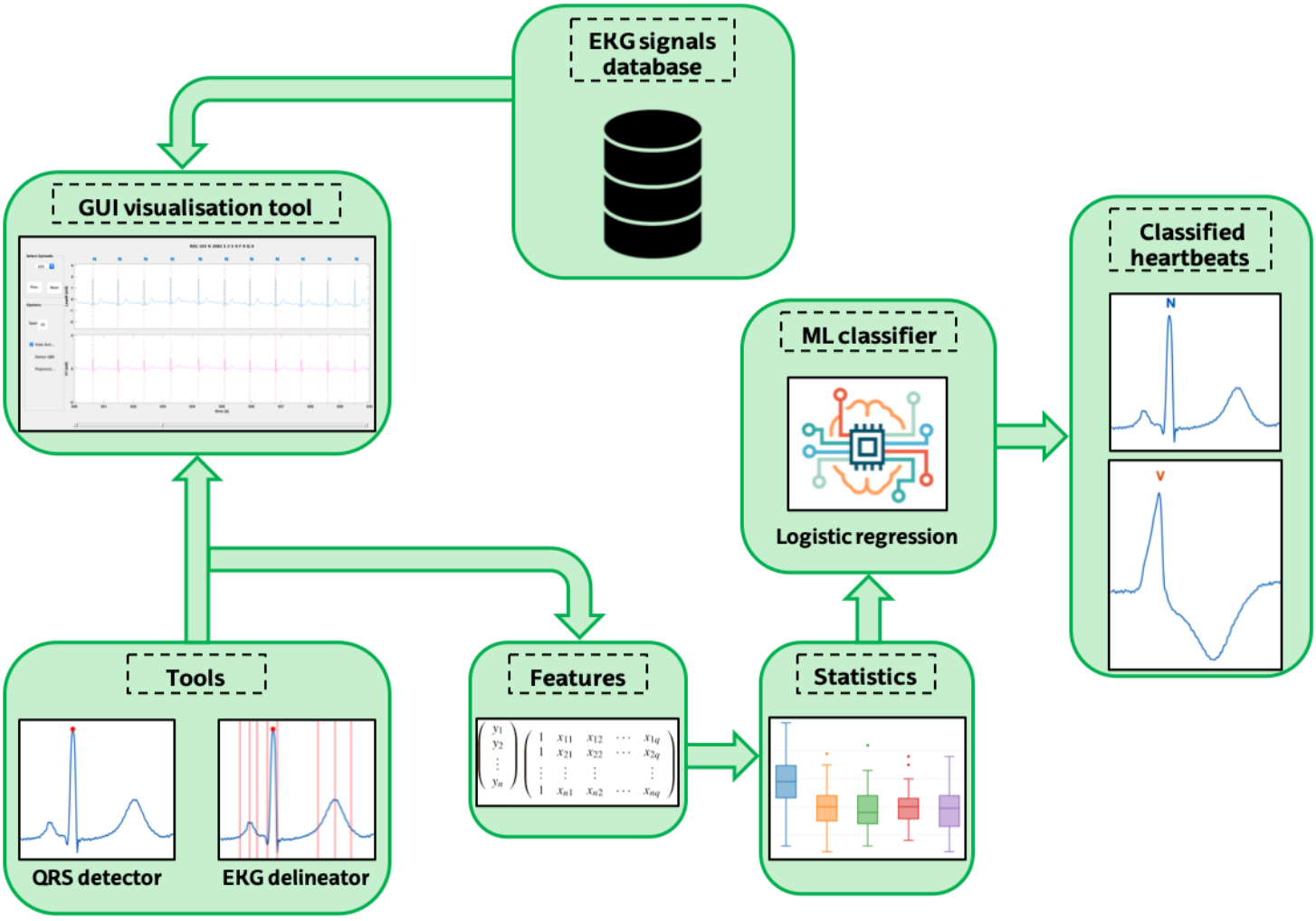


Figure 18: Block diagram of the main stages of the project.

7.2 EKG signals database

7.2.1 Conversion of the MIT-BIH Arrhythmia Database to MATLAB

The database chosen in this project is Physionet's open access MIT-BIH Arrhythmia Database. Its main characteristics have been described in section **6.2.1**. This database was downloaded from Physionet's webpage [52]. The MIT-BIH Arrhythmia database contains 48 recordings (see **Table 35** of Appendix I). Each of these recordings is composed of 3 files, which are explained in **Table 8**.

Table 8: breakdown of the file types, format and their description. Adapted from [61].

File type	File format	Description
MIT Signal	.dat	These are binary files which contain the digitalised signals' samples. However, these cannot be correctly interpreted without their corresponding header files.
MIT Header	.hea	These are short text files that describe the contents of associated signal files
MIT Annotation	.atr	These are binary files containing annotations (labels that generally refer to specific samples in associated signal files). Annotation files should be read with their associated header files.

However, these files are not directly readable in MATLAB. Hence the first step is to adapt the MIT-BIH Arrhythmia Database to MATLAB's file format. We have done this using the Physionet's Physiobank ATM toolkit, which is shown in **Figure 19**.

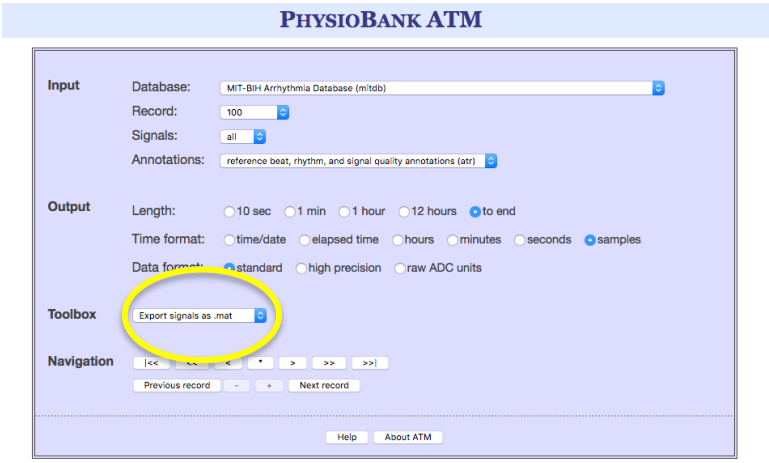


Figure 19: Physionet's Physiobank ATM tool.

As can be appreciated in **Figure 19,** after doing this, we get an '.mat' (which can be perfectly opened with MATLAB) file per recording. Another adaptations have been made in MATLAB to add the annotations to the '.mat' file of each recording. Finally, we obtain 48 ready-to-use '.mat' file recordings. Each of the generated '.mat' files is a variable of type double described in **Table 9**.

Table 9: description of the generated '100' recording's content. The same applies to remaining 47.

File name	File type	File content	File content's channels (leads)
100.mat	Double	s_ecg	Lead II
			V1

Moreover, in order to ease the management of the 48 recordings, another '.mat' file was created, containing additional relevant data for all 48 recordings. This is shown in **Table 10.**

Table 10: analysis of the 'metadata.mat' file.

File name	File type	File content	File content	Description
metadata.mat	1x48 Structure	data	name	The name of the recording
			t_ann	The time-stamps (s) for the annotations of the recording. These coincide with the R-peaks occurrence times.
			ann	The annotations that indicate the types of beats of the recording
			fs	Sample frequency of the recording

7.2.2 Database creation

Once we had MATLAB ready-to-use signals, the next task was to create the database with heartbeat annotations adjusted to the AAMI standard. These annotations are very important because they will be used as ground truth marks. The reason for doing this adjustment is that the database contains 109,494 beats, annotated by cardiologists into 15 different classes. Using the AAMI standard simplifies the classification of heartbeats; we pass from having 15 to just 5 heartbeat types. In addition, using this standard allows us to compare the results of this project with the available literature and therefore assess the goodness of the results. The mapping of the MIT-BIH Arrhythmia Database heartbeat types [62] to the AAMI heartbeat classes is shown in **Table 11**.

Table 11: AAMI heartbeat classification for the MIT-BIH database.

AAMI heartbeat class	N	S	V	F	Q
Description	Any heartbeat not in the S,V,F or Q classes	Supraventricular ectopic beat	Ventricular ectopic beat	Fusion beat	Unknown beat
MIT-BIH heartbeat types	Normal beat (NOR)	Atrial premature beat (AP)	Premature ventricular contraction (PVC)	Fusion of ventricular and normal beat (fVN)	Paced beat (P)
	Left bundle branch block beat (LBBB)	Aberrated atrial premature beat (aAp)	Ventricular escape beat (VE)		Fusion of paced and normal beat (fPN)
	Right bundle branch block beat (RBBB)	Nodal (junctional) premature beat (NP)			Unclassified beat (U)
	Atrial escape beats (AE)	Supraventricular premature beat (SP)			
	Nodal (junctional) escape beat (NE)				

The general overview of the number of heartbeats per type is shown in **Table 12**. A more detailed review is attached in **Table 38** of Appendix II). As shown in the table, the normal beats are by far the most numerous ones, followed by the unknown and the ventricular beats. This implies that in the design of the classifier we will have to address the class imbalance between normal and ventricular heartbeats.

Table 12: general overview MIT-BIH Arrhythmia Database after AAMI classification.

AAMI heartbeat class	N	S	V	F	Q
Full MIT-BIH database	90402	3010	7236	803	8043

7.3 Signal and annotation visualisation tool

The second stage of the project was the development of a software tool to visualise the recordings and annotations in the database. This has been done using MATLAB's Guide (GUI design environment). The tool had to be user-friendly to be used by non-technical operators such as physicians while providing a fast and clear interface to display the EKG and their annotations. A general overview of the GUI is shown in **Figure 20.**

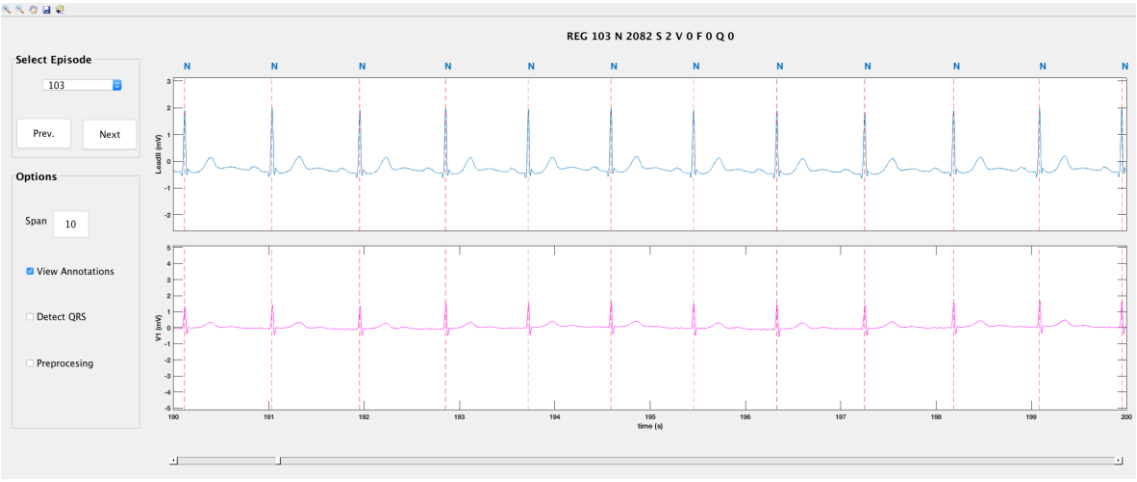


Figure 20: general overview of the GUI.

The GUI's interface is divided in three parts. The first one is the "Select Episode" area. As shown in **Figure 21**, clicking the central button opens a pop-up menu. Here, we can select the desired recording of the database. To improve the user experience, we have added two buttons, "Prev." and "Next", to shift between the previous and next recordings.

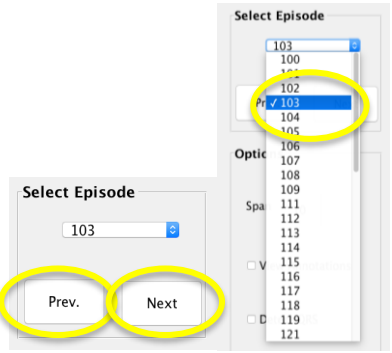


Figure 21: the GUI's "Select episode" area (left). The pop-up displayed (right).

The second set of tools was designed to control the GUI, and it is called "Options". With the "Span" textbox, we can modify the span of the recording's time-interval between 10 and 120s. The "View annotations" checkbox changes the annotation's visibility. The annotations are displayed in their time stamps over the first channel's plot. They follow the colour palette shown in **Table 13**. At those points, red vertical dashed lines are also drawn. When this checkbox is unclicked, the annotations and the vertical lines disappear, as shown in **Figure 22**.

Table 13: annotation colour distribution.

Annotation	Colour
N	Blue
S	Red
V	Orange
F	Violet
Q	Green

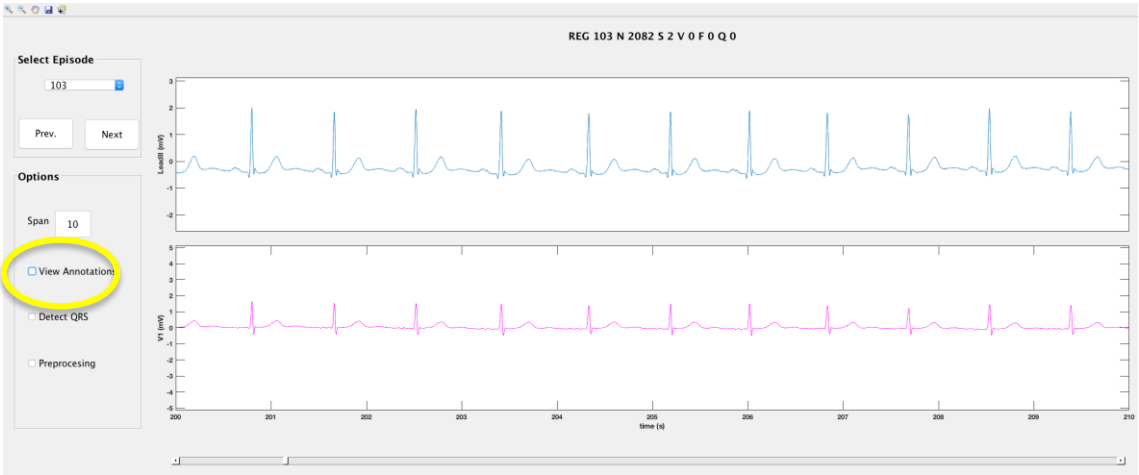


Figure 22: the GUI's general overview with the "View annotations" checkbox unclicked.

Finally, the GUI integrates the Hamilton Tompkins QRS detection algorithm (see **Section 7.4.1**). When the "Detect QRS" button is clicked, the HT algorithm automatically locates the QRS complexes using the EKG signal from channel 1 (Lead II). This helps not only for heartbeats classification, but also to locate them when they are not clear.

The third part of the GUI is the display section shown in **Figure 23**. Here, the GUI plots the selected episode's channels (Lead II and V1) at the same time. On the top, there is a label that displays the number of the register and heartbeat classification according to the annotations of the cardiologists. A key element is the slider at the bottom of the plots, which allows moving forward and backward along the time range of the recording.

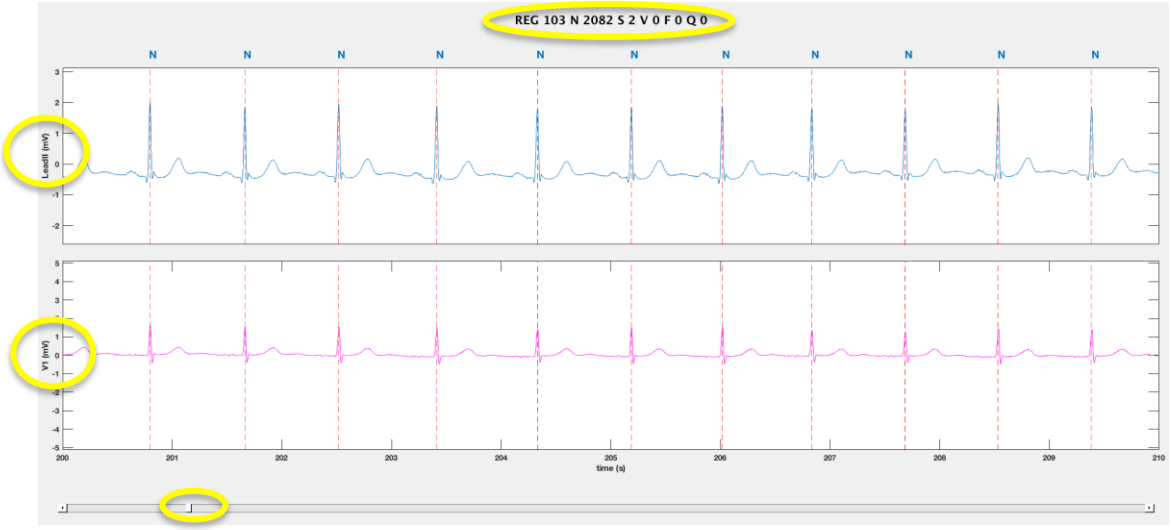
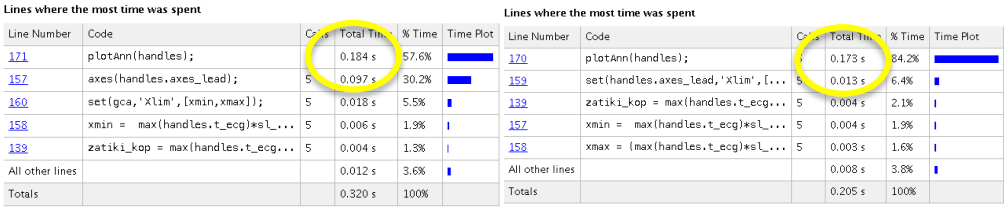


Figure 23: visualisation area of the GUI.

A key characteristic of the GUI is its rapid response to the user's commands. Some elements or programming styles result in long execution times in GUIs, so we used MATLAB's profiler tool to correct inefficient and slow code. Some functions related to signal and annotation display slowed down the GUI's performance. These delays could be of some milliseconds (see an example in **Figure 24**) or of some seconds (see an example in **Figure 25**), and were mostly related to inefficient coding for the display of the annotations.



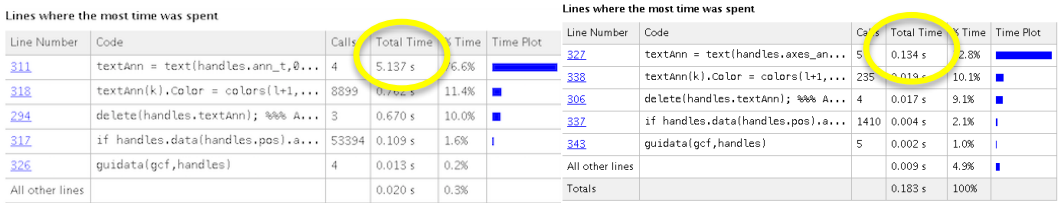
Lines where the most time was spent

Line Number	Code	Calls	Total Time	% Time	Time Plot
171	plotAnn(handles);		0.184 s	57.6%	
157	axes(handles.axes_lead);	5	0.097 s	30.2%	
160	set(gca,'Xlim',[xmin,xmax]);	5	0.018 s	5.5%	
158	xmin = max(handles.t_ecg)*sl_...	5	0.006 s	1.9%	
139	zatiki_kop = max(handles.t_ecg...	5	0.004 s	1.3%	
All other lines			0.012 s	3.6%	
Totals			0.320 s	100%	

Lines where the most time was spent

Line Number	Code	Calls	Total Time	% Time	Time Plot
170	plotAnn(handles);		0.173 s	84.2%	
159	set(handles.axes_lead,'Xlim',[...	5	0.013 s	6.4%	
139	zatiki_kop = max(handles.t_ecg...	5	0.004 s	2.1%	
157	xmin = max(handles.t_ecg)*sl_...	5	0.004 s	1.9%	
158	xmax = (max(handles.t_ecg)*sl_...	5	0.003 s	1.6%	
All other lines			0.008 s	3.8%	
Totals			0.205 s	100%	

Figure 24: Example of improvement of tens of millisecond in execution time from left to right.



Lines where the most time was spent

Line Number	Code	Calls	Total Time	% Time	Time Plot
311	textAnn = text(handles.ann_t,0...	4	5.137 s	76.6%	
318	textAnn(k).Color = colors(l+1,...	8899	0.765 s	11.4%	
294	delete(handles.textAnn); %%% A...	3	0.670 s	10.0%	
317	if handles.data(handles.pos).a...	53394	0.109 s	1.6%	
326	guidata(gcf,handles)	4	0.013 s	0.2%	
All other lines			0.020 s	0.3%	

Lines where the most time was spent

Line Number	Code	Calls	Total Time	% Time	Time Plot
327	textAnn = text(handles.axes_an...	5	0.134 s	72.8%	
338	textAnn(k).Color = colors(l+1,...	235	0.019 s	10.1%	
306	delete(handles.textAnn); %%% A...	4	0.017 s	9.1%	
337	if handles.data(handles.pos).a...	1410	0.004 s	2.1%	
343	guidata(gcf,handles)	5	0.002 s	1.0%	
All other lines			0.009 s	4.9%	
Totals			0.183 s	100%	

Figure 25. Example of improvement of a few seconds in execution time from left to right.

7.4 Heartbeat detection and delineation tools

For the QRS detection and heartbeat delineation we used two algorithms, the Hamilton-Tompinks and the Wavedec. The following two sections explain the implementation details of these two EKG processing tools.

7.4.1 Hamilton Tompkins QRS detector

The Hamilton Tompkins algorithm was applied to the EKG database to detect the QRS complexes. We have used an improved version of the Hamilton Tompkins algorithm that uses IIR filters instead of FIR filters with similar amplitude frequency responses. The HT algorithm package was provided by BioRes Research group.

On the other hand, the HT algorithm returns the time stamps where the R-peaks should occur, although the heartbeat detections are not normally placed at the R-peak, as shown in **Figure 26**. Therefore, the values had to be corrected and the detections of the algorithm had to be placed at the R-peaks by doing a local maximum search around the returned heartbeat detection.

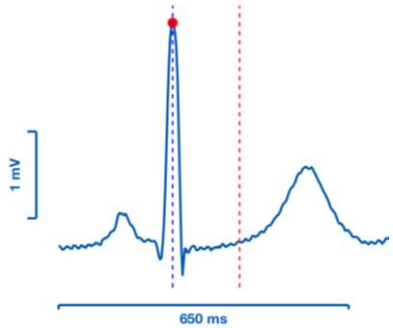


Figure 26: the blue dashed vertical line represents the time instant of the R-peak and the red one the detection of the HT algorithm. The red dot indicates the R-peak of the QRS complex.

In addition, sometimes the HT algorithm fails to detect some QRS complexes (False Negatives), while in other times it returns the time stamps for non-existent QRS complexes (False Positives). As a consequence, we had to assess the algorithm's detections to the actual annotations in the database, assuming that detections within 200ms of the annotated beats correspond to True Positives.

7.4.2 Wavedec heartbeat delineator

We have used this algorithm to extract the basic characteristics of the heartbeats, by delineating the heartbeats associated to the QRS complexes detected by the HT algorithm. The Wavedec algorithm was developed by BSiCoS Research Group from the University of Zaragoza [63]. However, in this work we have modified the MATLAB version of the Wavedec code for the BioRes Research Group to make it compatible to the data structure created for the MIT-BIH database.

The Wavedec algorithm takes as input a single channel EKG, its sampling frequency and the samples corresponding to the R-peak locations. And for each heartbeat, the software returns the different interval features described in **Table 14**.

Table 14: Wavedec's return parameters and its contents for each detected heartbeat.

Characteristic	Description
Pon	indicates where the P-wave should start.
P	indicates where the P-wave peak should be
Poff	indicates where the P-wave should end.
QRSon	indicates where the QRS signal should start
Q	indicates where the Q-wave peak should be
R	indicates where the R-wave peak should be
Rprima	Indicates the position, if exists, of an additional R peak.
QRS	indicates where the QRS signal should be centred. Coincides with R-wave peak position
S	indicates where the S-wave should be centred
QRSoff	indicates where the QRS signal should finish
Ton	indicates where the T-wave should start
T	indicates where the T-wave peak should be
Ttipo	Indicates the T signal type based on the number and polarity of the found maxima. There are six possible T wave morphologies (see **Figure 31**).
Toff	indicates where the T-wave should finish
QRSmainpos	Indicates if the QRS is positive (normal).
QRSmaininv	Indicates if the QRS is negative (inverted).
Pprima	Indicates the position, if exists, of an additional P-peak.

As shown in the table, Wavedec does not only return the basic characteristic information of each heartbeat described in **5.3.2**, but also additional information, such as the type of the T-wave (see **Figure 31**) or if the QRS complex is inverted or not.

All these basic characteristics will be used in the next stage to calculate more significant features intended for the later heartbeat classification between N and V.

7.5 Heartbeat classification features

At this stage of the project, heartbeat's significant features have been calculated for all the records in the MIT-BIH database using the HT algorithm followed by the Wavedec algorithm. The heartbeat's features can be grouped into different categories as described in the following sections.

7.5.1 Heartbeat time-interval features

First, the features related to the time-intervals have been computed. **Figure 27** shows the three essential time-intervals features, their definition is given in **Table 15**.

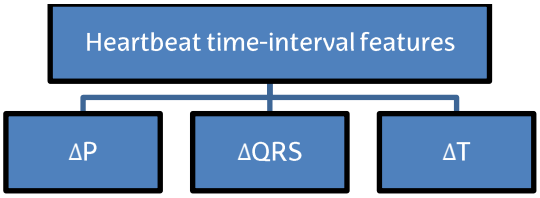


Figure 27: the calculated heartbeat time-interval features.

These features were determined from the time differences shown in **Table 15**.

Table 15: definition of the time-interval features.

Heartbeat time-interval features	Calculations
ΔP	Poff-Pon
ΔQRS	QRSoffset- QRSonset
ΔT	Toff-Ton

7.5.2 EKG morphology features

Secondly, we obtained the EKG morphology-features, namely the amplitude of the EKG at the peak points of its characteristic waveforms. These are listed in **Figure 28.**

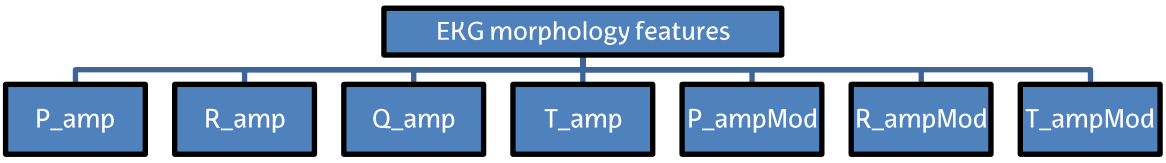


Figure 28: the calculated EKG morphology features.

To estimate these features, we followed the procedures described in **Table 16.** The amplitude of the S waves was not calculated because S waves are most of the times not detected by Wavedec.

Table 16: the EKG morphology features and how we calculated them.

EKG morphology features	Calculations
P_amp	The value of the EKG signal's first channel at P time instant
R_amp	The value of the EKG signal's first channel at R time instant
Q_amp	The value of the EKG signal's first channel at Q time instant
T_amp	The value of the EKG signal's first channel at T time instant
P_ampMod	The absolute value of the difference of the signals maximum and minimum absolute values between the Pon and Poff time instants
R_ampMod	The absolute value of the difference of the signals maximum and minimum absolute values between the QRSon and QRSoff time instants
T_ampMod	The absolute value of the difference of the signals maximum and minimum absolute values between the Ton and Toff time instants

One important addition to the features obtained from Wavedec was the definition of modified EKG wave amplitudes. The difference between the amplitude definition (with respect to 0) and the modified amplitude for the R-wave is illustrated in **Figure 29.**

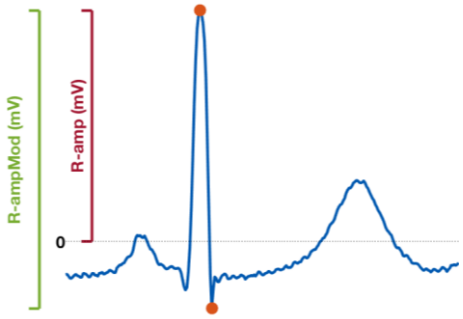


Figure 29: an example of the R_amp (green) and R_ampMod (red), with orange dots marking the maximum and minimum of the QRS complex.

7.5.3 Wave existence features

Since the ventricular beats usually do not have P waves, another significant feature is the existence or not of the P and T waves, as shown in **Table 17**.

Table 17: the P- and T-wave features and how we calculated them.

Wave existence	Calculations
P_pr	If a heartbeat has its Poff instant before its QRSon value, it indicates that there is a P-wave in that heartbeat
T_pr	If a heartbeat has its Ton instant after its QRSoff value, it indicates that there is a T-wave in that heartbeat

7.5.4 QRS inversion feature

As mentioned above, the Wavedec algorithm returns a value to indicate if the QRS of the delineated heartbeat is inverted or not. However, this is not very accurate, so we tried to make our own QRS inverted detector, based on the percentile values. We analysed the 20 % percentile of the R-peak in all the episodes. The results are shown in **Figure 30.**

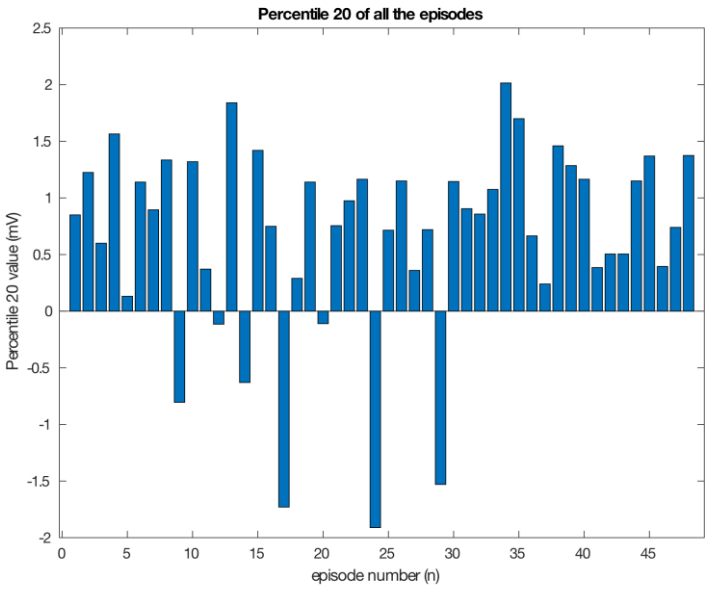


Figure 30: 20 % percentile of the R-peaks amplitude of the whole database.

The 20% percentile is significantly negative in those recording with high number of inverted QRS complexes. For that reason, we developed a self-made algorithm to correlate the amplitude of the R-peaks with the optimum percentile value of the whole episode. We applied this algorithm to the database to determine if the QRS were inverted or not.

7.5.5 Type of T wave feature

One of the output parameters of Wavedec for each heartbeat is the type of T wave. The different types of T waves are shown in **Figure 31**. The differences between the types of T wave are so pronounced that we can use this feature to distinguish normal and ventricular beats. A close look at **Figure 31** *b/d* and *e/f* reveal that normal and ventricular beats have different T wave morphology.

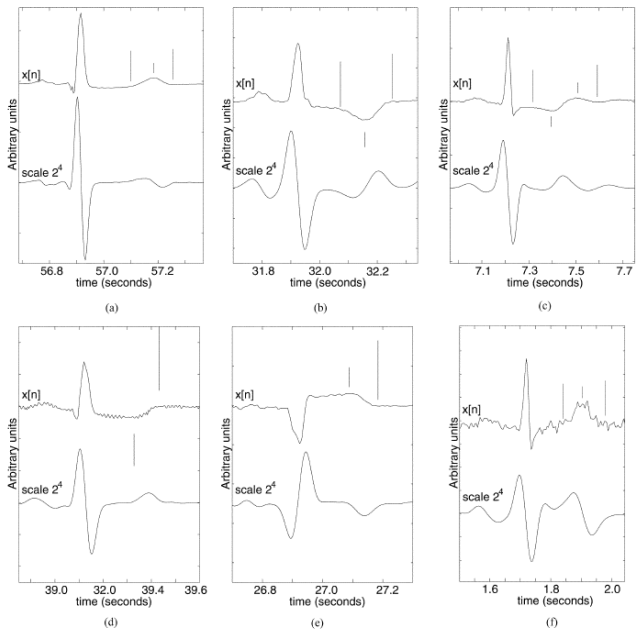


Figure 31: six different types of T-wave returned by Wavedec. Extracted from Martínez et al [41].

7.6 Data preparation for the classifier

All the heartbeat related data must be grouped in proper data types to be used to train and test the classifiers. For that purpose, we created a feature matrix $\boldsymbol{X}$ that contained all the features of the heartbeats in the database, and a ground truth class label vector $\boldsymbol{y}$ with the type of heartbeats. The two data variables $\boldsymbol{X}$ and $\boldsymbol{y}$ will be used later for the development of the Logistic Regression classifier.

The feature matrix $\boldsymbol{X}$ is organised as follows. Each row corresponds to a heartbeat, and each column to a feature, as shown in **Table 18**. So, the element X_{ij} corresponds to feature j of heartbeat i. Since the MIT-BIH database contains109,494 heartbeats, and each heartbeat was characterised using 15 features, matrix X is a 109494x15 matrix.

Table 18: general overview of the X matrix. N = 109,494.

P_pr_1	T_pr_1	T_tp_1	QRS_inv_1	ΔP_1	ΔQRS_1	ΔT_1	P_amp_1	R_amp_1	Q_amp_1	P_ampMod_1	R_ampMod_1	T_ampMod_1
P_pr_2	T_pr_2	T_tp_1	QRS_inv_2	ΔP_2	ΔQRS_2	ΔT_2	P_amp_2	R_amp_2	Q_amp_2	P_ampMod_2	R_ampMod_2	T_ampMod_2
P_pr_3	T_pr_3	T_tp_1	QRS_inv_3	ΔP_3	ΔQRS_3	ΔT_3	P_amp_3	R_amp_3	Q_amp_3	P_ampMod_3	R_ampMod_3	T_ampMod_3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
P_pr_N	T_pr_N	T_tp_N	QRS_inv_N	ΔP_N	ΔQRS_N	ΔT_N	P_amp_N	R_amp_N	Q_amp_N	P_ampMod_N	R_ampMod_N	T_ampMod_N

The ground truth heartbeat label vector y is organised similarly, as a column vector in which row i represents the true heartbeat annotation for beat i. In addition, we also created a pat_ID vector to give each heartbeat a label between 1-48 that identifies the original register from the MIT-BIH database.

7.7 Statistical analysis

Before introducing the data into the classifier, it is important to visually and analytically assess the statistical differences in the distributions of the values of the features for normal and ventricular heartbeats. The basic approach is to compute central measures of tendency (mean/median) and dispersion (standard deviation/percentile ranges). In the case of binary features like T/P wave existence this will be done using bar plots. For continuous variables we used boxplots and scatterplots.

A boxplot is a standardised method to graphically represent a series of numerical data across its quartiles. Thus, the main reason to use the boxplot, is that it shows at a glance the median and quartiles values of the data. A detailed visual explanation of the boxplot is displayed on **Figure 32**.

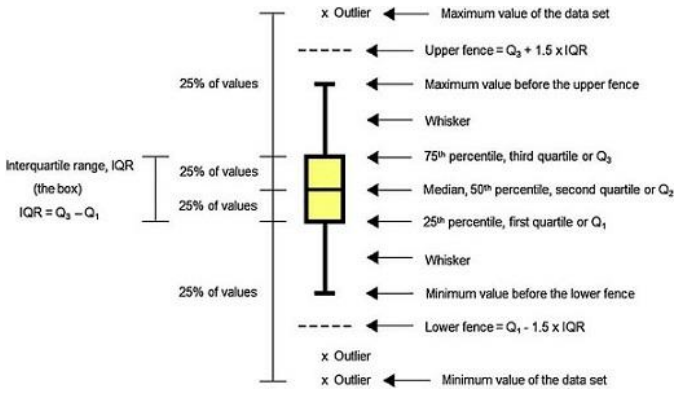


Figure 32: boxplot visual explanation extracted from [64].

As shown in the figure, on each box the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' or 'x' symbol.

To compare the statistical distributions of a variable between the two classes we used the Mann-Whitney test. This is a nonparametric test applied to two independent samples and returns a measure of the probability that two sets of samples of different sizes have the same median [65]; this is the p value. When the p-value is low we can be confident that the medians are different. In this work we assumed that $p < 0.05$ represented statistically significant differences in medians between N and V heartbeats [66]. An example of the comparison of the distributions (in terms of boxplots) and their statistical differences for the R-peak amplitude of N and V heartbeats is presented in **Figure 33**.

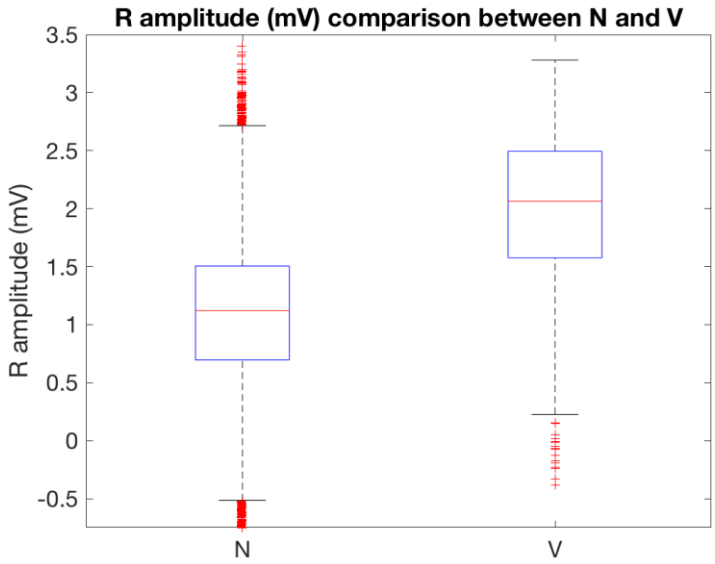


Figure 33: example of boxplot for R_amp feature.

The visual assessment of the boxplots already indicates the statistical differences will be significant, since the medians and ranges are well separated. This is confirmed by the Mann-Whitney test that in this case returns a value $p < 0.05$.

7.8 The logistic regression classifier

7.8.1 Preparation of the data

Before the classifier is trained and tested the data has to be first separated into two sets, one for training (to learn from data) and a separate one for testing (to evaluate its performance on unseen data). The objective is to design a classifier that is accurate on new data, not on data it has already tested. In addition, there is a strong class imbalance between N and V beats, and has been addressed by assigning different weights (importance) to each class, in order to penalise the misclassifications of the least frequent class (V). The whole phase of the preparation of the data is illustrated in **Figure 34.**

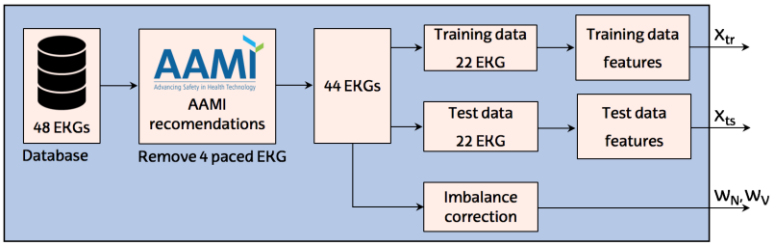


Figure 34: general block diagram of the preparation data stage.

Moreover, to compare the results of the classifier that we obtained with the available literature, we had to follow the AAMI standard, which recommends to remove the recordings containing paced beats. In our case, we had to remove 4 recordings ('102','104','107' and '217').

Once we did that, we proceeded to split the remaining recordings into two datasets, one for the training and the other for testing the classifier. Both datasets contained approximately 50,500 beats. However, the database was highly imbalanced (see **Figure 35)**, almost the 90% of the beats belong to the N class whilst just the 6%, fall into V class (the remaining 4% corresponded to the SVEB, and F classes, which were not considered in this project). So, with this class distribution an algorithm that gave always an N classification would have an accuracy of over 90%, while being completely useless.

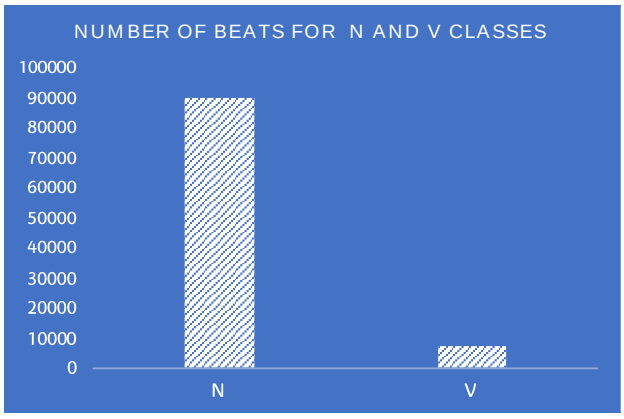


Figure 35: the number of N and V type beats for both datasets.

To minimise the impact of class imbalance we weighted the beats using the probabilistic frequencies:

$$w_N = \frac{n_N}{n_T} \qquad w_V = \frac{n_V}{n_T} \qquad (5)$$

The n_T stands for the total number of beats, while n_N and n_V indicate the number of N and V beats, respectively. This way, we got the weights for the N (w_N) and V (w_V) classes that we passed to the learning algorithm in the training process.

7.8.1.1 Training process and classification

Firstly, we used all the extracted features and the ground truth marks as the inputs for the logistic regression machine learning, together with the weights. The output of this learner is the model of the classifier. This process is illustrated in **Figure 36**.

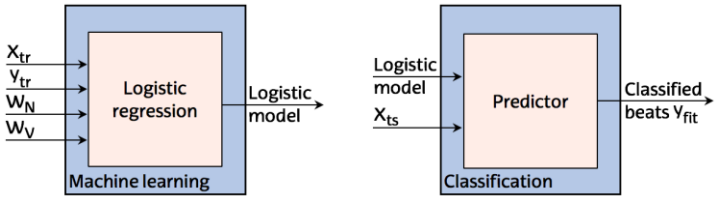


Figure 36: machine learning step (left) and classification step (right). At the output from the classification stage we have the classified beats (y_{fit}).

Secondly, we decided not to use all the extracted features but their best combination. The number of k-element combinations of n features (n=15), without repetition is obtained through the **equation 6**.

$$C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad (6)$$

So, the total number of possible combinations will be:

$$\sum_{k=1}^{n} C_{n,k} = \sum_{k=1}^{15} \frac{n!}{k!(n-k)!} \qquad (7)$$

For each of the obtained combinations we repeated the procedure described in **Figure 36**.

Finally, we evaluated the results against the ground truth annotations, y, using the confusion matrix (see **Figure 14**) and its derived parameters (Se, Sp, PPV and NPV). This phase is represented in **Figure 37.**

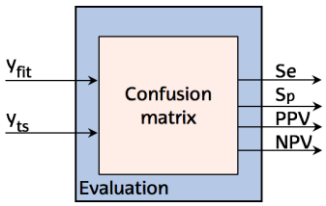


Figure 37: evaluation phase of the training and classification stage.

This process was also done by using the test data to train the classifier and the train data to test the classifier.

7.9 Summary of results

In this section we summarise the most relevant results

7.9.1 QRS detector

The HT QRS detector evaluated on the MIT-BIH database presented a Se of 99.80% and a PPV of 99.40%. The detailed performance for each of the records in the database is given in Appendix II. Although the HT algorithm is very accurate it failed in some instances. When the R-peaks from the HT were plotted in the GUI, we noticed that the HT algorithm failed when big slopes were presented in the EKG, as can be seen in **Figure 38**. In the figure there are false positive detections (FP) that appear just before and after the great depression of the EKG. This usually correspond to abrupt movement artefacts in the EKG.

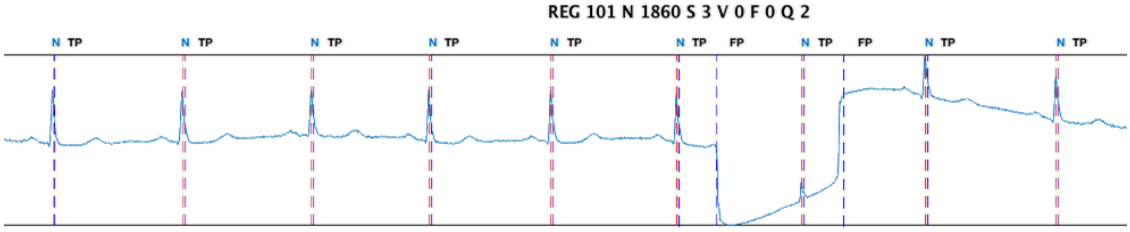


Figure 38: [171-178] s time interval of the 101 recording with the detected and corrected QRS time stamps in blue vertical lines. The test's results are in black next to each annotation.

Besides, we also observed that the HT algorithm failed to detect QRS complexes when when large changes in QRS slopes occurred between consecutive beats. An example of this is shown in **Figure 39**. We can see that the 4^{th} – 9^{th} QRS complexes are not detected by the HT algorithm, so they are FN (false negatives).

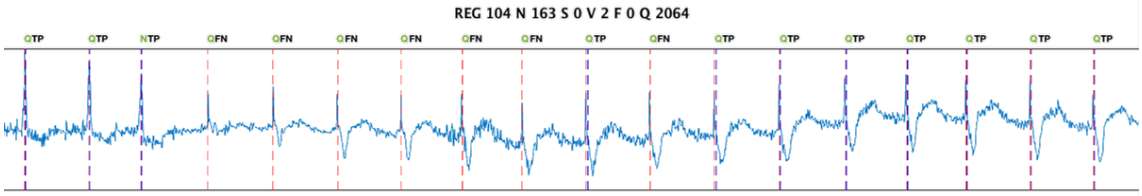


Figure 39: [255-270] s time interval of the 104 recording with the detected and corrected QRS time stamps in blue vertical lines. The test's results are in black next to each annotation.

To finish with this stage, we set the corrected values of the time stamps as the fiducial points for the next step of the processing.

7.9.2 EKG delineation and statistical evaluation

The analysis of the different features gave us information about their individual capacity of prediction. In **Figure 40** and **Figure 41.** We can see different features comparations for the N and V classes.

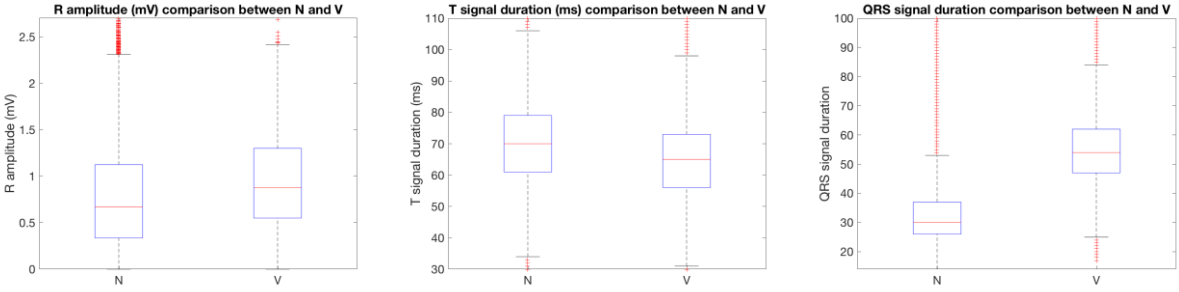


Figure 40: an example of the analysed boxplots for three representative continuous features.

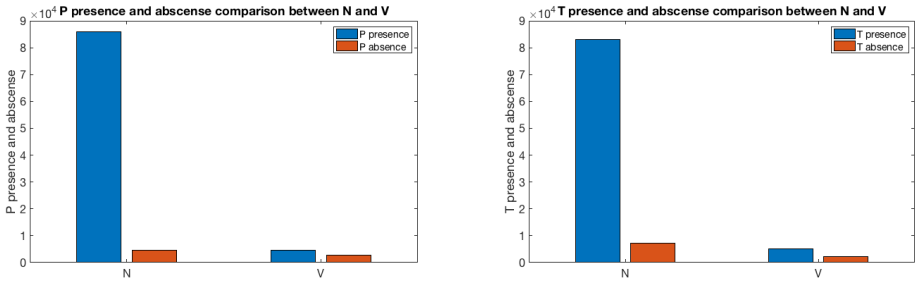


Figure 41: an example of the generated bar plots for binary features.

In **Figure 41** we observed that although in normal beats there is more relative presence of T waves than in ventricular ones, the difference is not significant, and these features will most likely not be relevant to discriminate N and V heartbeats.

The p values for the corresponding Mann–Whitney test are listed in **Table 19**.

Table 19: p-values for the Mann–Whitney test to assess differences in median of the feature values for N and V heartbeats.

Feature	p-value
P_pr	$4.11 * 10^{-11}$
T_pr	$5.68 * 10^{-6}$
T_tp	$2.41 * 10^{-22}$
QRS_inv	$5.12 * 10^{-11}$
ΔP	$1.13 * 10^{-5}$
ΔQRS	$2.97 * 10^{-250}$
ΔT	$2.47 * 10^{-107}$
P_amp	$9.82 * 10^{-88}$
R_amp	$6.22 * 10^{-248}$
Q_amp	$6.80 * 10^{-07}$
T_amp	$4.02 * 10^{-22}$
P_ampMod	$2.75 * 10^{-46}$
R_ampMod	$8.69 * 10^{-299}$
T_ampMod	$1.64 * 10^{-07}$

In terms of p values the most discriminative features (lowest p values) are R_ampMod, ΔQRS, R_amp and ΔT. These results are in line with our intuitive perception because, as can be seen in **Figure 10,** the main differences between N and V is their duration (V has longer QRS duration) and their amplitude (V has greater QRS).

7.9.3 Baseline classifier

We started using the complete set of 15 features to get the first classification results. However, we saw that they were not as good as expected with a Se of 79.99 % and an Sp of 57.94%. Therefore, we tried all the 32,767 possible combinations of features in order to get the one that maximised the Se and Sp values. The **Figure 42** presents an example of some of the achieved Sp and Se values for a different number of features (predictors).

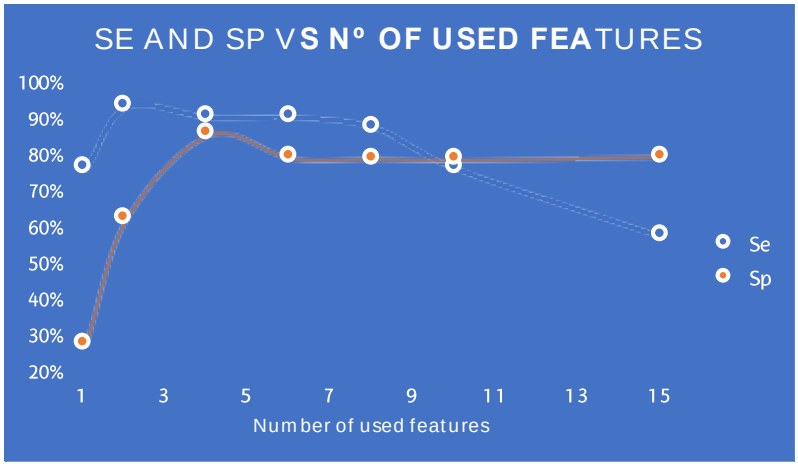


Figure 42: comparative graph for the obtained Se and Sp values in function of the used number of features.

We analysed those results and we concluded that the better results are not related to the number of the used features, but to the quality of the combination of the predictors. The feature combination that maximised the obtained results was the one that used the features displayed in **Table 20**.

Table 20: the feature combination that gets the best Se and Sp results.

	Feature
Best feature combination	ΔP
	ΔQRS
	ΔT
	RAmp

This feature combination provided a **Se of 90.98% and a Sp of 85.98%, a PPV of 0.05% and a NPV of 99.92%.** This means that the classifier detects the 90.98 % of the ventricular beats and correctly classifies 85.98% of the detected beats. This combination provides an improvement over the initial 15 feature combination of 13-points in Se and of 48.4-points in Sp.

The combination of features that offers us the best Se y Sp is formed mostly by the most significant features that revealed the Mann-Whitney test (see **Table 19**).

7.9.4 Examples of classified beats

As we have pointed out before, the classifier is not perfect. Sometimes the classifier misclassifies the heartbeats, this usually happens when the morphology of ventricular heartbeats is similar to that of normal ones, and vice versa. Some examples are shown in **Figure 43**.

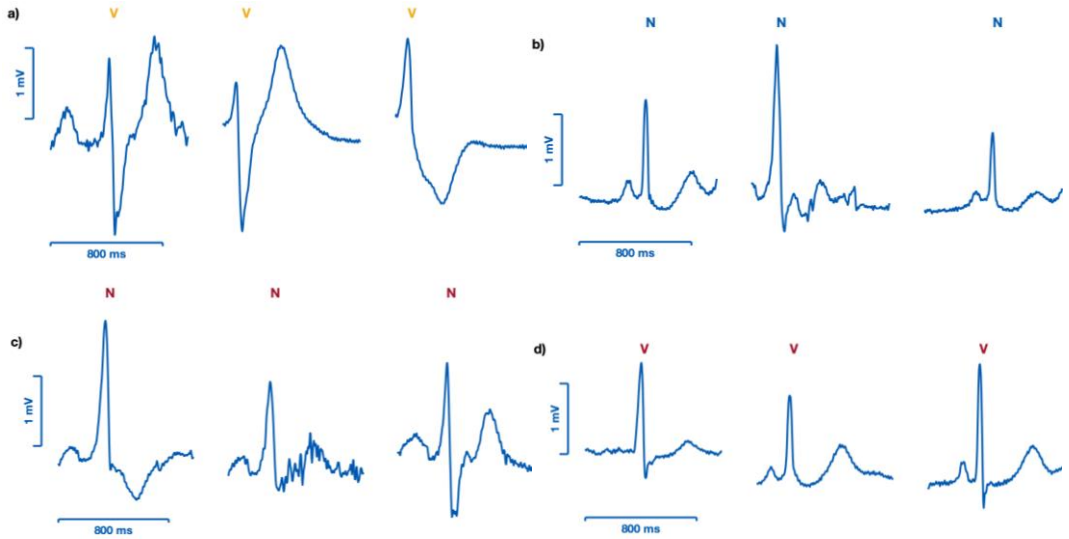


Figure 43: examples of correctly classified beats on top, and of misclassified beats below. Ventricular beats are shown on the left (a,c) and normal beats on the right (b,d).

8 METHODOLOGY

The project planning provides control techniques to manage the complexity, amount of data and deadlines of the project. This is made prior to starting the development of any project, and it is essential to define the tasks and responsibilities of all those who take part in that project. It is essential to define the limits and duration of each participant's work packages for being able to carry out the project in an efficient and controlled way. So in this section of the document, we will introduce the work team and the work packages carried out by those teams.

The work packages in which the project is divided include:

- Tasks to be performed and their deliverables.
- Milestones, which should reflect the project's checkpoints or critical dates.

In the subsequent points, the members of the working group are listed, and the different work packages are described in detail. Finally, the Gantt diagram of the project is also included.

8.1 Working group

Table 21: project's working group.

ID	Position	Name and surnames	Function
G1	Project manager	Unai Irusta Zarandona	Proposes the project, indicates the necessary stages to follow and takes care of the correction and supervision of the work.
G2	Junior engineer	Gorka Zubia Garea	It is in charge of the development the project and of writing the document.

8.2 Work packages

The following tables show the work packages that have been defined for the project. In each work package the description of that stage and the corresponding tasks are explained, specifying the duration and the start and end dates of each of them.

First/overall stage of the project:

Table 22: first work package.

WP1	Starting data	Ending data	Duration (days)
PROJECT MANAGEMENT: Monitoring and administration carried out to verify the proper development of the project.	**05/07/2018**	**15/07/2019**	**375**
T.1.1. Management, monitoring and supervision of work: Coordination and supervision work from the start to the end of the project.	05/07/2018	15/07/2019	375

Second stage of the project:

Table 23: second work package.

WP2	Starting date	Ending date	Duration (days)
BASIC TRAINING: learning the basic skills to be able to begin with the preparation of the project	**05/07/2018**	**07/09/2018**	**64**
T.2.1. Acquiring MATLAB skills: following online courses with practical tasks	05/07/2018	07/09/2018	64
T.2.2. Learning the importance of CVD: reading scientific papers about the CVD and sudden cardiac arrest	05/07/2018	07/09/2018	64

Third stage of the project:

Table 24: third work package.

WP3	Starting date	Ending date	Duration (days)
PROJECT PREPARATION: acquisition of knowledge necessary before specifying the course of the project and its development	**07/09/2018**	**30/09/2018**	**13**
T.3.1. GUI design skills: MATLAB GUI design train course delivered by BioRes	07/09/2018	14/09/2018	7
T.3.2. Project definition: Definition of the project guidelines and the work plan.	07/09/2018	15/09/2018	8
T.3.3. State of art: Search for relevant information and studies necessary for the development of the project	07/09/2018	30/09/2018	13

Fourth stage of the project:

Table 25: fourth work package.

WP4	Starting date	Ending date	Duration (days)
Project development: Different sections that have been tackled for the development of the project	**30/09/2018**	**7/04/2018**	**189**
T.4.1 Database creation: adapt the MIT-BIH Arrhythmia database to MATLAB and AAMI standards.	30/09/2018	16/10/2018	16
T.4.2 Development of the GUI: in order to be able to work with the signals of the database, which will also serve to show the results from the implemented algorithms.	16/10/2018	06/11/2018	21
T.4.3. Algorithms	06/11/2018	15/01/2019	70

T.4.3.1 QRS detector: Detection of QRS complexes using an automatic QRS detector. In this case, the Hamilton-Tompkins was used together with two additional patches to adjust the detections to the R-wave peaks, given the importance of a correct and precise detection of the beats. **T.4.3.2 Heartbeat delineation:** delineation of the heartbeats detected by the HT algorithm, using the Wavedec algorithm. This way, several features of the heartbeats were extracted. These features were fundamental for the later developed classifier.	*06/11/2018* *26/11/2018*	*26/11/2018* *15/01/2019*	*20* *50*
T.4.4 Feature extraction: analysis of the results of the delineator and obtaining new characteristics from them. These together with the previous features are the predictors used in the classifier **T.4.5 LR classification:** adaptation of the features obtained in the previous sections in order to use them to train the ML LR classifier. Evaluation of the obtained results.	15/01/2019 26/02/2019	26/02/2019 07/04/2019	42 40

Fifth stage of the project:

Table 26: fifth work package.

WP5	Starting date	Ending date	Duration (days)
DOCUMENTATION AND SUBMISSION OF THE PROJECT: wording of the project and oral presentation.	**07/04/2019**	**15/07/2019**	**99**
T.5.1. Project documentation: drafting of the document that summarises the context of the project, the objectives, scope, benefits, the description of the solution, methodology and conclusions.	07/04/2019	27/06/2019	81
T.5.2. Oral presentation of the project: preparation of the presentation, rehearsal and presentation in front of the evaluation board.	27/06/2019	15/09/2019	18

The description of the working time schedule is shown in **Table 27**. Although the project was completed in 41 weeks, there was a large variation in hours per week due to other activities done by G1 and G2 during the development of the project.

Table 27: Working time schedule of the project.

Unit	Duration
Project	41 weeks
Week	7 days
Day	3 hours*

* This is an average, there was wide variation in hours from 0 hour to 6-hour days.

This section defines the milestones and deliverables that must be met throughout the development of the project. These are shown in **Table 28** and **Table 29**.

Table 28: project milestones.

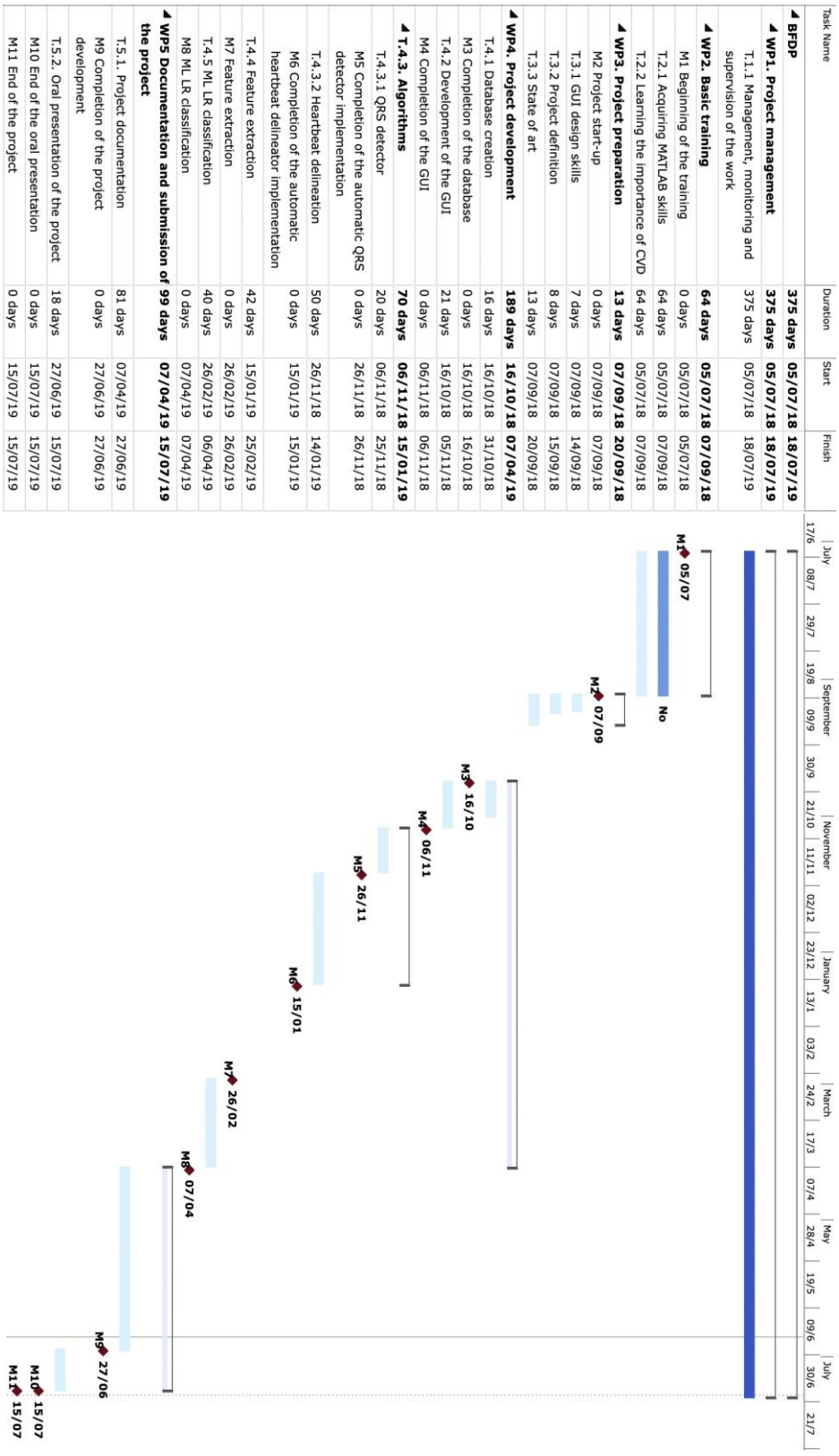
ID	Milestone	Date
M1	Beginning of the training	05/07/2018
M2	Project start-up	07/09/2018
M3	Completion of the database	16/10/2018
M4	Completion of the GUI	06/11/2018
M5	Completion of the automatic QRS detector implementation	26/11/2018
M6	Completion of the automatic heartbeat delineator implementation	15/01/2019
M7	Feature extraction	26/02/2019
M8	ML LR classification	7/04/2019
M9	Completion of the project development documentation	27/06/2019
M10	End of the oral presentation	15/07/2019
M11	End of the project	15/07/2019

Table 29: project deliverables.

ID	Deliverable	Date
D1	Database	05/07/2018
D2	Algorithms	07/09/2018
D3	Classifier	16/10/2018
M4	Documentation of the report of the project	06/11/2018
M5	Presentation	26/11/2018

8.3 Gantt diagram

Figure 44 shows the tasks and milestones together with the GANTT diagram of the project planification.



Task Name	Duration	Start	Finish
BFDP	**375 days**	**05/07/18**	**18/07/19**
WP1. Project management	**375 days**	**05/07/18**	**18/07/19**
T.1.1 Management, monitoring and supervision of the work	375 days	05/07/18	18/07/19
WP2. Basic training	**64 days**	**05/07/18**	**07/09/18**
M1 Beginning of the training	0 days	05/07/18	05/07/18
T.2.1 Acquiring MATLAB skills	64 days	05/07/18	07/09/18
T.2.2 Learning the importance of CVD	64 days	05/07/18	07/09/18
WP3. Project preparation	**13 days**	**07/09/18**	**20/09/18**
M2 Project start-up	0 days	07/09/18	07/09/18
T.3.1 GUI design skills	7 days	07/09/18	14/09/18
T.3.2 Project definition	8 days	07/09/18	15/09/18
T.3.3 State of art	13 days	07/09/18	20/09/18
WP4. Project development	**189 days**	**16/10/18**	**07/04/19**
T.4.1 Database creation	16 days	16/10/18	31/10/18
M3 Completion of the database	0 days	16/10/18	16/10/18
T.4.2 Development of the GUI	21 days	16/10/18	05/11/18
M4 Completion of the GUI	0 days	06/11/18	06/11/18
T.4.3. Algorithms	**70 days**	**06/11/18**	**15/01/19**
T.4.3.1 QRS detector	20 days	06/11/18	25/11/18
M5 Completion of the automatic QRS detector implementation	0 days	26/11/18	26/11/18
T.4.3.2 Heartbeat delineation	50 days	26/11/18	14/01/19
M6 Completion of the automatic heartbeat delineator implementation	0 days	15/01/19	15/01/19
T.4.4 Feature extraction	42 days	15/01/19	25/02/19
M7 Feature extraction	0 days	26/02/19	26/02/19
T.4.5 ML LR classification	40 days	26/02/19	06/04/19
M8 ML LR classification	0 days	07/04/19	07/04/19
WP5 Documentation and submission of the project	**99 days**	**07/04/19**	**15/07/19**
T.5.1. Project documentation	81 days	07/04/19	27/06/19
M9 Completion of the project development	0 days	27/06/19	27/06/19
T.5.2. Oral presentation of the project	18 days	27/06/19	15/07/19
M10 End of the oral presentation	0 days	15/07/19	15/07/19
M11 End of the project	0 days	15/07/19	15/07/19

Figure 44: the GANTT diagram followed through the project.

9 BUDGET OF THE PROJECT

This section presents the project's costs which are mainly the cost of human resources and that of the material used in the project, having in mind both depreciable material and consumables.

9.1 Human resources

This is the salary of each member of the team:

Table 30: hourly wage of the members of the project team.

ID	Duration	Salary (€/h)
G1	Project manager	60
G2	Junior engineer	30

Table 31 presents an economic balance sheet of the project's human resources, taking into account the work hours spent in the project and the unit cost of each participant.

Table 31: cost of the human resources.

WP task	G2 work (h)	G2 cost (€)	G1 work (h)	G1 cost (€)	Total work (h)	Total cost (€)
T.1.1	10	600	10	300	30	1200
T.2.1	10	600	10	300	20	900
T.2.2	15	900	10	300	25	1200
T.3.1	10	600	15	450	25	1050
T.3.2	10	600	25	750	35	1350
T.3.3	15	900	60	1800	75	2700
T.4.1	10	600	20	600	30	1200
T.4.2	10	600	25	750	35	1350
T.4.3	15	900	60	1800	75	2700
T.4.4	10	600	20	600	30	1200
T.4.5	15	900	60	1800	75	2700
T.5.1	30	1800	80	2400	110	4200
T.5.2	10	600	20	600	30	1200
				TOTAL	**595**	**22950**

9.2 Material resources

In this section we are presenting the tables of the costs of depreciable material and consumables.

9.2.1 Depreciable material

These are the costs and expenses of the depreciable material.

Table 32: total cost of the depreciable material.

Material	Units	Initial cost (€)	Lifespan (months)	Use (months)	Cost (€)
Toshiba laptop	1	1000	36	6	166.66
Printer	1	50	36	0.5	6.94
MATLAB with toolboxes licence	1	1300	24	10	500
Microsoft office	1	150	15	2	20
				SUBTOTAL	**520**

9.2.2 Consumables

The next table lists the cost associated with consumables.

Table 33: total cost of consumables.

ID	Material	Cost (€)
C1	Office supplies	50
C2	Energy bill	30
C3	Hard disk	50
	SUBTOTAL	**130**

9.2.3 Summary of the budget of the project

In the following table we summarise the total costs and expenses: human resources (work hours), depreciable material and consumables.

Table 34: summary of the total costs and expenses.

Concept	Cost (€)
Work hours	22950
Depreciations	520
Consumables	130
TOTAL	**23600**

Adding up all the costs, the total cost of the project development amounts to **twenty three thousand six hundred euros**. Most of the expenses correspond to work hours.

10 Risks analysis

The objective of this section is to identify the possible risks throughout the development of the project and to have a contingency plan to minimise their impact. As the project is finished, it can be said that the risks have been avoided. However, from the beginning it has been essential to foresee these risks, so an analysis has been made of the possible risks and the damage they could cause.

Two concepts have been considered to carry out this risk analysis. On the one hand, the chance of the risks to occur has been studied. On the other hand, the impact that these risks may have on the project has been considered. Therefore, the risk analysis will take into account the probability of occurrence and its possible incidence. These two parameters have been measured as follows:

- Probability: low, medium or high.
- Impact: low, medium or high.

The possible risks foreseen and the contingency measures that would have been foreseen to combat them are listed below.

10.1 Risk of coding errors (A)

When developing an algorithm, it is very common to find coding errors that produce the wrong execution of the program and hinder the normal progress of work. Coding errors are frequent (high probability), and can have a medium impact on the project as it can leave the project on standby for days. Also, in the worst case, it may involve rewriting the code we have worked on.

To reduce the effect of this risk, we use MATLAB's *debug* tool, to locate bugs in the code. Besides, it is recommended to run the program every few days to make sure everything is correct. If the root of the problem is not found, members of BioRes research group can be consulted.

10.2 Risk of delays (B)

It is very common to have delays in the different stages of the project, which can lead to not fulfilling the deadlines established at the beginning. This is very likely, but it has a low impact because the working group is very small and the working time is easily recoverable.

To minimise this risk, a prior planning of the work is done, well-structured and in which the tasks have some margin for completion.

10.3 Risk of data loss (C)

This risk includes any loss of information that may occur during the project, whether in the documentation of the work, the latest versions of the code for the various algorithms and software tools, and the databases and annotations created. This event has a low probability of occurrence, but in case the impact would be high.

To avoid this risk, several systems are used to backup and store the data, like hard disks or the cloud. In addition, we periodically save the files while working on them, so that, in case the software fails, a recent version of them is available.

10.4 Risk of staff leaving (D)

It is also necessary to consider the possible departures (for medical or personal reasons) of the different members of the working group. This is a rare fact (low probability), and the impact can be considered medium although it can vary depending on the responsibility that the individual has in the project and the duration of his absence.

In this case, there is no action that can be taken. In the event of termination, the project manager will decide whether it is necessary to postpone completion of the project or reassign responsibilities.

10.5 Technological risks (E)

Not having knowledge in the use of the indispensable technologies for the development of the project, having hardware/software problems or having problems in the integration of the developed interfaces or scripts, are examples of risks that can appear within this set.

Nevertheless, measures have been taken to avoid these risks. To begin with, since it was the first time that MATLAB was used in such depth by G2, a training programme was designed before the project began. In addition, a proper use and maintenance of the equipment was ensured. Therefore, in view of the above, the probability of technological risks is low and marginally influential.

10.6 Risk of excessive costs (F)

Exceeding the planned development costs is a concept that is included in the cost risks. In the case of this project, this would be possible if costs not taken into account, such as the need for more material, or changes in prices, appeared. Therefore, to deal with these events, we added a 5% margin to the budget at the start of the project for eventualities.

However, in order to develop our project, not much material has been purchased. Therefore, the probability of this risk is very low, and its impact on the project would also be small.

10.7 Summary of the risk analysis

Below in **Figure 45** is a matrix showing the probability relationship of the different mentioned risks.

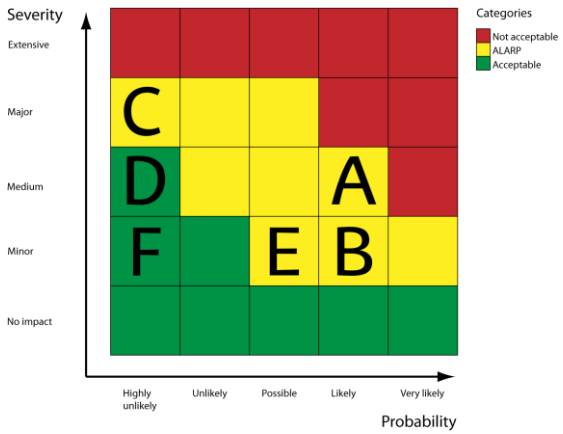


Figure 45: severity-probability matrix [67]. Green: acceptable risk. Yellow: as low as reasonably practicable risk (ALARP). Red: unacceptable risk.

11 CONCLUSIONS AND FUTURE WORK

The aim of this project has been to develop and implement a supervised algorithm to discriminate between normal and ventricular heartbeats using the electrocardiogram. The project has concluded by completing all the marked objectives and tasks. These were developed according to the proposed methodology.

In particular, an EKG database has been created following the AAMI standards from the MIT-BIH publicly available database. We have developed a user-friendly graphical user interface for the visualisation and handling of the EKG signals and their annotations. We have also implemented into this GUI, the Hamilton Tompkins QRS detector and the Wavedec delineator, to detect and segment the heartbeats from the database. The main result of this implementation has been the calculation of several features of the beats.

Afterwards, we have completed the statistical characterisation of the features, with the purpose of identifying distinctive patterns that will help differentiate normal from ventricular beats. We have concluded that the best individual predictors were the R_ampMod, ΔQRS, R_amp and ΔT.

We have used all the extracted features to develop a machine learning logistic regression classifier. The optimum feature combination included the ΔP, ΔQRS, R_amp and ΔT. The best results of this classification were **Se of 90.98% and a Sp of 85.98%, a PPV of 0.05% and a NPV of 99.92% for normal beats.**

This implies that the classifier detects the 90.98% of the ventricular beats and correctly classifies 85.98% of the normal beats. And that the confidence on the rightness of the classification is of 99.93% for normal and of 83.89% for ventricular beats. These results are an improvement on some of the previously reported results for automated heartbeat classification systems; in particular to those obtained by Chazal et al [7], which reported a Se of 77.7%, a PPV of 81.9% and a NPV of 1.2% for ventricular beats (we took the ventricular as the positives while Chazal took the normal).

A final conclusion is that **the adding all EKG features does not imply a better performance of the algorithm**. We have achieved the best results by taking 4 of the 15 features.

This method will help on the early detection of CVDs. Consequently, it will be possible to design interventions or preventive measures earlier, minimising their effects on the patient's health and reducing the associated costs.

We will finish pointing out the research lines opened by this project. The developed method tackles the discrimination of the two most important types of beats, namely normal and ventricular beats, and it is a first step towards the automatic classification of heartbeats. Therefore, it sets the framework for future developments of more complete heartbeat classification systems, and of algorithms and procedures to diagnose complex arrhythmia.

12 BIBLIOGRAPHY

[1]‘Share of deaths by cause’, *Our World in Data*. [Online]. Available: https://ourworldindata.org/grapher/share-of-deaths-by-cause-2016. [Accessed: 14-Apr-2019].

[2]Rioja Salud, ‘Enfermedades Cardiovasculares’, *riojasalud.es*. [Online]. Available: https://www.riojasalud.es/ciudadanos/problemas-de-salud/23-enfermedades-cardiovasculares. [Accessed: 30-May-2019].

[3]S. S. Menéndez, ‘Enfermedades Cardiovasculares’, Institut d’ Estudis de la Salut, Barcelona.

[4]K. Minami, H. Nakajima, and T. Toyoshima, ‘Real-time discrimination of ventricular tachyarrhythmia with Fourier-transform neural network’, *IEEE Trans Biomed Eng*, vol. 46, no. 2, pp. 179–185, Feb. 1999.

[5]S. Barro, R. Ruiz, D. Cabello, and J. Mira, ‘Algorithmic sequential decision-making in the frequency domain for life threatening ventricular arrhythmias and imitative artefacts: a diagnostic system’, *Journal of Biomedical Engineering*, vol. 11, no. 4, pp. 320–328, Jul. 1989.

[6]P. S. Hamilton and W. J. Tompkins, ‘Quantitative Investigation of QRS Detection Rules Using the MIT/BIH Arrhythmia Database’, *IEEE Transactions on biomedical engineering*, vol. BME-33, no. 12, p. 8, Dec. 1986.

[7]P. deChazal, M. O’Dwyer, and R. B. Reilly, ‘Automatic Classification of Heartbeats Using ECG Morphology and Heartbeat Interval Features’, *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 7, pp. 1196–1206, Jul. 2004.

[8]V. Mondéjar-Guerra, J. Novo, J. Rouco, M. G. Penedo, and M. Ortega, ‘Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers’, *Biomedical Signal Processing and Control*, vol. 47, pp. 41–48, Jan. 2019.

[9]K. N. V. P. S. Rajesh and R. Dhuli, ‘Classification of imbalanced ECG beats using re-sampling techniques and AdaBoost ensemble classifier’, *Biomedical Signal Processing and Control*, vol. 41, pp. 242–254, Mar. 2018.

[10]M. Hammad, A. Maher, K. Wang, F. Jiang, and M. Amrani, ‘Detection of abnormal heart conditions based on characteristics of ECG signals’, *Measurement*, vol. 125, pp. 634–644, Sep. 2018.

[11]‘Annual number of deaths by cause’, *Our World in Data*. [Online]. Available: https://ourworldindata.org/grapher/annual-number-of-deaths-by-cause. [Accessed: 14-Apr-2019].

[12]‘Heart diseases and strokes cause over 1.8 million deaths in the EU’. [Online]. Available: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20170928-1?inheritRedirect=true. [Accessed: 14-Apr-2019].

[13]Sociedad española de cardiología, ‘País Vasco presenta el tercermejor índicede mortalidad cardiovascular a nivel nacional’, *www.secardiologia.es*, Madrid, España, 14-Apr-2015.

[14]Eusko Jaurlaritza, ‘Salud pública y adicciones’, Osasun saila, 2016.

[15]appleTREE, ‘La inmensa mayoría de enfermedades cardiovasculares son prevenibles’, *Fundación Española del Corazón*. [Online]. Available: https://fundaciondelcorazon.com/prensa/notas-de-prensa/2545-inmensa-mayoria-de-enfermedades-cardiovasculares-son-prevenibles.html. [Accessed: 14-Apr-2019].

[16]L. Ontiveros, ‘¿Cuánto cuesta un enfermo cardiovascular?’, *Fundación Española del Corazón*. [Online]. Available: https://fundaciondelcorazon.com/corazon-facil/blog-impulso-vital/2208-cuanto-cuesta-enfermo-cardiovascular.html. [Accessed: 14-Apr-2019].

[17]OpenStax College, ‘ECG Tracing with Heart Contraction’. 19-Jun-2013.

[18]Fundación Española del Corazón, ‘Holter’, *Fundación Española del Corazón*. [Online]. Available: https://fundaciondelcorazon.com/informacion-para-pacientes/metodos-diagnosticos/holter.html. [Accessed: 08-Jun-2019].

[19]GRUPO DE TRABAJO DE PREVENCIÓN DE RIESGOS LABORALES, 'IMPLANTACIÓN DE DESFIBRILADORES EXTERNOS AUTOMÁTICOS Y SEMIAUTOMÁTICOS (DESAs) EN LA UNIVERSIDAD.', Conferencia de Rectores de las Universidades Españolas (CRUE), Girona, Oct. 2013.

[20]'Equipment for making electrocardiogram wires vector image on VectorStock', *VectorStock*. [Online]. Available: https://www.vectorstock.com/royalty-free-vector/equipment-for-making-electrocardiogram-wires-vector-19965594. [Accessed: 08-Jun-2019].

[21]S. GROUP, 'ECG 2.gif (Imagen GIF, 2304 × 1536 píxeles) - Escalado (38 %)'. [Online]. Available: http://www.sosgroup.co/UserFiles/Image/ECG%202.gif. [Accessed: 07-Jun-2019].

[22]L. M. Girbau and J. B. Terradellas, 'Capítulo 52 - Arritmias cardíacas', in *Farreras Rozman. Medicina Interna.*, 18th ed., vol. 1, 2 vols, Barcelona: Elsevier, 2016, p. 28.

[23]N. Rodríguez de Viguri, J. López Mesa, and M. Ruano Campos M, *Manual de soporte vital avanzado.*, 4th ed. Madrid, España: Masson, 2007.

[24]J. J. M. de Vreede-Swagemakers *et al.*, 'Out-of-Hospital Cardiac Arrest in the 1990s: A Population-Based Study in the Maastricht Area on Incidence, Characteristics and Survival', *Journal of the American College of Cardiology*, vol. 30, no. 6, pp. 1500–1505, Nov. 1997.

[25]W. Einthoven, 'Ueber die Form des menschlichen Electrocardiogramms', *Pflüger, Arch.*, vol. 60, no. 3–4, pp. 101–123, Mar. 1895.

[26]Dale. Dubin, *Electrocardiografía práctica : lesión, trazado e interpretación*, 3rd ed. México, D.F. [etc.] : Interamericana, 2007.

[27]'File:SinusRhythmLabels.svg', *Wikipedia*. .

[28]'Welcome to AAMI - Membership and Community - Association for the Advancement of Medical Instrumentation'. [Online]. Available: https://www.aami.org/membershipcommunity/content.aspx?ItemNumber=1292&navItemNumber=4603. [Accessed: 15-Jun-2019].

[29]B.-U. Kohler, C. Hennig, and R. Orglmeister, 'The principles of software QRS detection', *IEEE Eng. Med. Biol. Mag.*, vol. 21, no. 1, pp. 42–57, Feb. 2002.

[30]E. Martin *et al.*, 'Sensitivity and Specificity', in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2011, pp. 901–902.

[31]G. L. Iverson, 'Positive Predictive Power', in *Encyclopedia of Clinical Neuropsychology*, J. S. Kreutzer, J. DeLuca, and B. Caplan, Eds. New York, NY: Springer New York, 2011, pp. 1968–1970.

[32]'PhysioNet'. [Online]. Available: https://physionet.org/. [Accessed: 16-Jun-2019].

[33]Cuiwei Li, Chongxun Zheng, and Changfeng Tai, 'Detection of ECG characteristic points using wavelet transforms', *IEEE Trans. Biomed. Eng.*, vol. 42, no. 1, pp. 21–28, Jan. 1995.

[34]J. S. Sahambi, S. N. Tandon, and R. K. P. Bhatt, 'Using wavelet transforms for ECG characterization. An on-line digital signal processing system', *IEEE Eng. Med. Biol. Mag.*, vol. 16, no. 1, pp. 77–83, Feb. 1997.

[35]M. Bahoura, M. Hassani, and M. Hubin, 'DSP implementation of wavelet transform for real time ECG wave forms detection and heart rate analysis', *Comput Methods Programs Biomed*, vol. 52, no. 1, pp. 35–44, Jan. 1997.

[36]M.-E. Nygårds and L. Sörnmo, 'Delineation of the QRS complex using the envelope of the e.c.g.', *Med. Biol. Eng. Comput.*, vol. 21, no. 5, pp. 538–547, Sep. 1983.

[37]J. G. Kemmelings, A. C. Linnenbank, S. L. Muilwijk, A. SippensGroenewegen, A. Peper, and C. A. Grimbergen, 'Automatic QRS onset and offset detection for body surface QRS integral mapping of ventricular tachycardia', *IEEE Trans Biomed Eng*, vol. 41, no. 9, pp. 830–836, Sep. 1994.

[38]P. Laguna, R. Jané, and P. Caminal, 'Automatic detection of wave boundaries in multilead ECG signals: validation with the CSE database', *Comput. Biomed. Res.*, vol. 27, no. 1, pp. 45–60, Feb. 1994.

[39]G. Speranza, G. Nollo, F. Ravelli, and R. Antolini, 'Beat-to-beat measurement and analysis of the R-T interval in 24 h ECG Holter recordings', *Med Biol Eng Comput*, vol. 31, no. 5, pp. 487–494, Sep. 1993.

[40]S. H. Meij, P. Klootwijk, J. Arends, and J. R. T. C. Roelandt, 'An algorithm for automatic beat-to-beat measurement of the QT-interval', in *Computers in Cardiology 1994*, Bethesda, MD, USA, 1995, pp. 597–600.

[41]J. P. Martinez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna, 'A Wavelet-Based ECG Delineator: Evaluation on Standard Databases', *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 4, pp. 570–581, Apr. 2004.

[42]'screenshot.png (Imagen PNG, 699 × 462 píxeles)'. [Online]. Available: https://www.mathworks.com/matlabcentral/mlc-downloads/downloads/submissions/60900/versions/13/screenshot.png. [Accessed: 16-Jun-2019].

[43]'Sensitivity and specificity', *Wikipedia*. 07-May-2019.

[44]W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds., *Encyclopedia of Systems Biology*. New York, NY: Springer New York, 2013.

[45]G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 103. New York, NY: Springer New York, 2013.

[46]T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning*. New York, NY: Springer New York, 2001.

[47]'MATLAB Documentation'. [Online]. Available: https://www.mathworks.com/help/matlab/. [Accessed: 16-Jun-2019].

[48]'MATLAB Documentation - MathWorks España'. [Online]. Available: https://es.mathworks.com/help/. [Accessed: 16-Jun-2019].

[49]'New License for MATLAB Student R2019a - MathWorks España'. [Online]. Available: https://es.mathworks.com/store/link/products/student/new. [Accessed: 16-Jun-2019].

[50]'MATLAB Programming/Differences between Octave and MATLAB - Wikibooks, open books for an open world'. [Online]. Available: https://en.wikibooks.org/wiki/MATLAB_Programming/Differences_between_Octave_and_MATLAB. [Accessed: 17-Jun-2019].

[51]'About'. [Online]. Available: https://www.gnu.org/software/octave/about.html. [Accessed: 17-Jun-2019].

[52]G. B. Moody and R. G. Mark, 'MIT-BIH Arrhythmia Database'. physionet.org, 1992.

[53]F. M. Nolle and R. W. Bowser, 'Creighton University Ventricular Tachyarrhythmia Database'. physionet.org, 1992.

[54]W. P. Holsinger, K. M. Kempner, and M. H. Miller, 'A QRS Preprocessor Based on Digital Differentiation', *IEEE Transactions on Biomedical Engineering*, vol. BME-18, no. 3, pp. 212–217, May 1971.

[55]J. Fraden and M. R. Neuman, 'QRS wave detection', *Med. Biol. Eng. Comput.*, vol. 18, no. 2, pp. 125–132, Mar. 1980.

[56]J. C. T. B. Moraes, M. M. Freitas, F. N. Vilani, and E. V. Costa, 'A QRS complex detection algorithm using electrocardiogram leads', in *Computers in Cardiology*, Memphis, TN, USA, 2002, pp. 205–208.

[57]M. Okada, 'A digital filter for the QRS complex detection', *IEEE Trans Biomed Eng*, vol. 26, no. 12, pp. 700–703, Dec. 1979.

[58]W. Zong, G. B. Moody, and D. Jiang, 'A robust open-source algorithm to detect onset and duration of QRS complexes', in *Computers in Cardiology, 2003*, Thessaloniki Chalkidiki, Greece, 2003, pp. 737–740.

[59]Bert-Uwe Köhler, Hennig, C, and Orglmeister, Reinhold, 'QRS detection using zero crossing counts.', *Progress in Biomedical Research*, no. 8, pp. 138–145, 2003.

[60]C. A. González, ‘SVM: Máquinas de Vectores Soporte’, presented at the Escuela de Ingeniería Informática de Valladolid.

[61]‘Frequently Asked Questions about PhysioNet’. [Online]. Available: https://physionet.org/faq.shtml#downloading-databases. [Accessed: 19-Jun-2019].

[62]‘PhysioBank Annotations’. [Online]. Available: https://physionet.org/physiobank/annotations.shtml. [Accessed: 25-Jun-2019].

[63]‘ECG detector/delineator (Wavedet)’, *BSICoS Group Website*, 10-Apr-2015. .

[64]J. E. V. Ferreira *et al.*, ‘Graphical representation of chemical periodicity of main elements through boxplot’, *Educación química*, vol. 27, no. 3, pp. 209–216, Jul. 2016.

[65]M. Neuhäuser, ‘Wilcoxon–Mann–Whitney Test’, in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1656–1658.

[66]W. Haynes, ‘Wilcoxon Rank Sum Test’, in *Encyclopedia of Systems Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds. New York, NY: Springer New York, 2013, pp. 2354–2355.

[67]‘Risk matrices - CGE Barrier Based Risk Management Knowledge base’. [Online]. Available: https://www.cgerisk.com/knowledgebase/Risk_matrices. [Accessed: 22-Jun-2019].

APPENDIX I

1 MIT-BIH Arrhythmia Database:

The free and publicly accessible MIT-BIH Arrhythmia Database was obtained from the Physionet's website and contained the records of 48 different patients which are shown in the **Table 35**. Each of these recordings was 30 minutes long and has been sampled at a frequency of 360 Hz. In each recording there are two channels, one of EKG lead II and the other of lead V1 [52]. For the development of this work we have used lead II.

PhysioNet is supported by the National Institute of General Medical Sciences (NIGMS) and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number 2R01GM104987-09 [32].

Table 35: Recording's names of the MIT-Arrhythmia Database.

	Names of the recordings				
MIT-BIH Arrhythmia Database's recordings	**100**	**111**	**122**	**209**	**222**
	101	**112**	**123**	**210**	**223**
	102	**113**	**124**	**212**	**228**
	103	**114**	**200**	**213**	**230**
	104	**115**	**201**	**214**	**231**
	105	**116**	**202**	**215**	**232**
	106	**117**	**203**	**217**	**233**
	107	**118**	**205**	**219**	**234**
	108	**119**	**207**	**220**	
	109	**121**	**208**	**221**	

Originally, the EKGs of the MIT-BIH Arrhythmia Database contain two types of annotations. The first type indicates the type of beat, and the second is additional information about the EKG.

Table 36: heartbeat type annotations. Extracted from [62].

Code	Description
N	Normal beat
L	Left bundle branch block beat
R	Right bundle branch block beat
B	Bundle branch block beat (unspecified)
A	Atrial premature beat
a	Aberrated atrial premature beat
J	Nodal (junctional) premature beat
S	Supraventricular premature or ectopic beat (atrial or nodal)
V	Premature ventricular contraction
r	R-on-T premature ventricular contraction
F	Fusion of ventricular and normal beat

e	Atrial escape beat
j	Nodal (junctional) escape beat
n	Supraventricular escape beat (atrial or nodal)
E	Ventricular escape beat
/	Paced beat
f	Fusion of paced and normal beat
Q	Unclassifiable beat
?	Beat not classified during learning

Table 37: non-heartbeat type annotations. Extracted from [62].

Code	Description
[	Start of ventricular flutter/fibrillation
!	Ventricular flutter wave
]	End of ventricular flutter/fibrillation
x	Non-conducted P-wave (blocked APC)
(	Waveform onset
)	Waveform end
p	Peak of P-wave
t	Peak of T-wave
u	Peak of U-wave
`	PQ junction
'	J-point
^	(Non-captured) pacemaker artefact
|	Isolated QRS-like artefact
~	Change in signal quality
+	Rhythm change
s	ST segment change
T	T-wave change
*	Systole
D	Diastole

As already explained, the database annotations had to be adapted to the AAMI standard in order to be able to compare our results with other similar studies. After making this adjustment, the heartbeat distribution of the database is shown in **Table 38**.

Table 38: detailed breakdown of the MIT-BIH Arrhythmia Database after AAMI classification.

AAMI heartbeat class	N	S	V	F	Q
Full MIT-BIH database	**90402**	**3010**	**7236**	**803**	**8043**
100	2239	33	1	0	0
101	1860	3	0	0	2
102	99	0	4	0	2084
103	2082	2	0	0	0
104	163	0	2	0	2064
105	2526	0	41	0	5
106	1507	0	520	0	0
107	0	0	59	0	2078
108	1739	5	17	2	0
109	2492	0	38	2	0
111	2123	0	1	0	0
112	2537	2	0	0	0
113	1789	6	0	0	0
114	1820	12	43	4	0
115	1953	0	0	0	0
116	2302	1	109	0	0
117	1534	1	0	0	0
118	2166	96	16	0	0
119	1543	0	444	0	0
121	1861	1	1	0	0
122	2476	0	0	0	0
123	1515	0	3	0	0
124	1531	36	47	5	0
200	1743	30	826	2	0
201	1625	138	198	2	0
202	2061	55	19	1	0
203	2529	2	444	1	4
205	2571	3	71	11	0
207	1543	107	210	0	0
208	1586	2	992	373	2
209	2621	383	1	0	0
210	2423	22	195	10	0
212	2748	0	0	0	0
213	2641	28	220	362	0
214	2003	0	256	1	2

215	3195	3	164	1	0
217	244	0	162	0	1802
219	2082	7	64	1	0
220	1954	94	0	0	0
221	2031	0	396	0	0
222	2062	421	0	0	0
223	2045	73	473	14	0
228	1688	3	362	0	0
230	2255	0	1	0	0
231	1568	1	2	0	0
232	397	1383	0	0	0
233	2230	7	831	11	0
234	2700	50	3	0	0

2 QRS detection results

The Hamilton-Tompkins algorithm is a well-known QRS detector that was used to identify the QRS complexes in the EKG signals. Although its good performance is already proven, we decided to make an experimental test to assess the integrity and reliability of our results. The QRS detectors performance is measured calculating the Se and PPV values for each record. **Table 39** shows the breakdown of the results for each recording.

Table 39: experimental results of the HT algorithm for the MIT-BIH database.

Recording	Total beats	Detected beats	TP	FP	FN	Se (%)	PPV (%)
100	2273	2271	2271	0	2	99.91	100.00
101	1865	1871	1865	6	0	100.00	99.68
102	2187	2186	2186	0	1	99.95	100.00
103	2084	2082	2082	0	2	99.90	100.00
104	2229	2282	2220	62	5	99.78	97.28
105	2572	2616	2563	53	8	99.69	97.97
106	2027	2023	2023	0	4	99.80	100.00
107	2137	2137	2136	1	1	99.95	99.95
108	1763	1810	1752	58	5	99.72	96.80
109	2532	2528	2526	2	6	99.76	99.92
111	2124	2125	2123	2	1	99.95	99.91
112	2539	2540	2538	2	1	99.96	99.92
113	1795	1794	1794	0	1	99.94	100.00
114	1879	1882	1878	4	1	99.95	99.79
115	1953	1952	1952	0	1	99.95	100.00
116	2412	2392	2390	2	22	99.09	99.92
117	1535	1535	1534	1	1	99.93	99.93
118	2278	2281	2277	4	1	99.96	99.82
119	1987	1987	1986	1	1	99.95	99.95
121	1863	1864	1861	3	2	99.89	99.84
122	2476	2475	2474	1	2	99.92	99.96
123	1518	1517	1517	0	1	99.93	100.00
124	1619	1615	1614	1	5	99.69	99.94
200	2601	2616	2597	19	4	99.85	99.27
201	1963	1937	1937	0	26	98.68	100.00
202	2136	2131	2131	0	5	99.77	100.00
203	2980	3014	2966	48	10	99.66	98.41
205	2656	2652	2652	0	4	99.85	100.00
207	1860	2112	1852	260	7	99.62	87.69
208	2955	2918	2912	6	43	98.54	99.79
209	3005	3011	3004	7	1	99.97	99.77
210	2650	2645	2639	6	11	99.58	99.77
212	2748	2749	2747	2	1	99.96	99.93

213	3251	3247	3247	0	4	99.88	100.00
214	2262	2260	2258	2	4	99.82	99.91
215	3363	3361	3360	1	3	99.91	99.97
217	2208	2207	2205	2	3	99.86	99.91
219	2154	2153	2153	0	1	99.95	100.00
220	2048	2047	2047	0	1	99.95	100.00
221	2427	2421	2421	0	6	99.75	100.00
222	2483	2485	2481	4	2	99.92	99.84
223	2605	2606	2603	3	2	99.92	99.88
228	2053	2113	2050	63	3	99.85	97.02
230	2256	2257	2255	2	1	99.96	99.91
231	1571	1570	1570	0	1	99.94	100.00
232	1780	1791	1780	11	0	100.00	99.39
233	3079	3072	3072	0	7	99.77	100.00
234	2753	2749	2749	0	4	99.85	100.00
			Mean sensitivity (Se, %)				**99'80**
			Mean Positive Predictive value (PPV, %)				**99'40**

3 Detailed distributions of heartbeat features

3.1 Individual features

In this section we present the boxplots that are left in the main document but we used for the statistical characterisation of the extracted features together with the Mann-Whitney test.

3.1.1 Heartbeat time-interval features

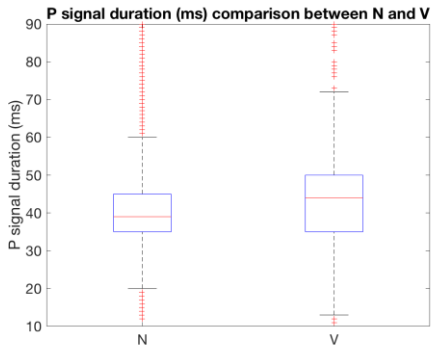


Figure 46: boxplot for the ΔP feature.

In this boxplot we can infer that the **ΔP** is not a good induvial predictor because there is a large overlap between the boxplots (distributions) for both classes.

3.1.2 EKG morphological features

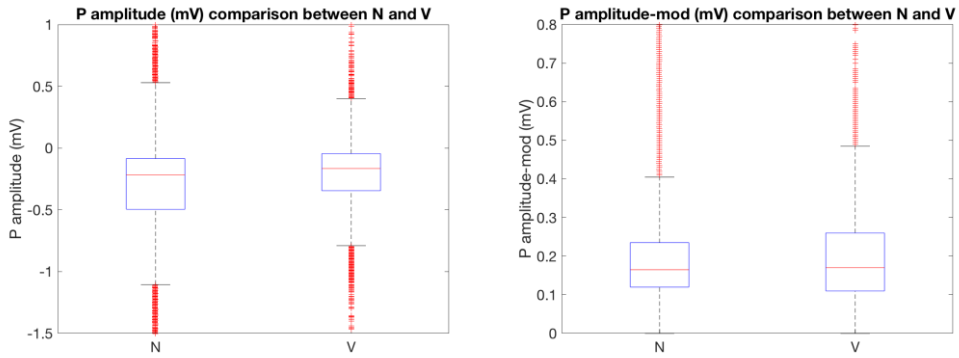


Figure 47: boxplot for the P_amp (left) and P_amp_mod (right) features.

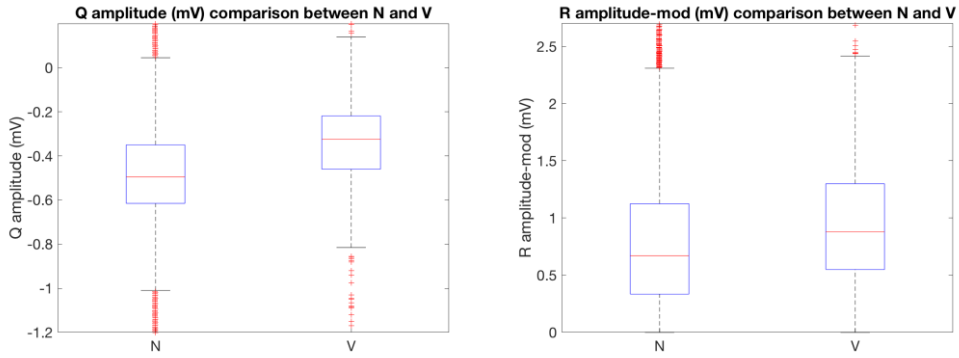


Figure 48: boxplot for the Q_amp (left) and R_amp_mod (right) features.

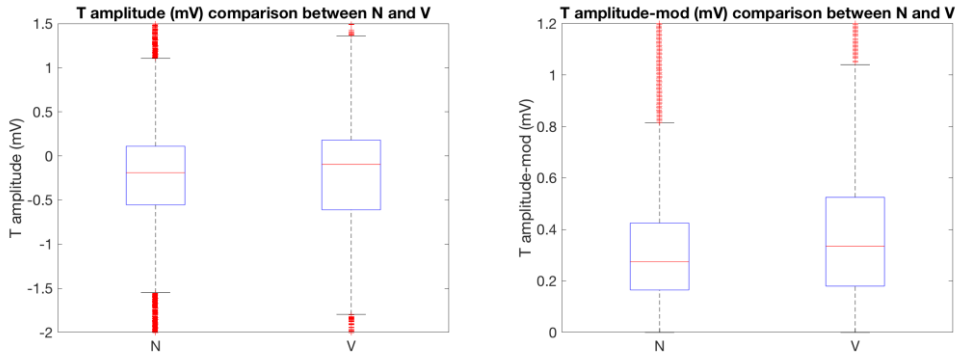


Figure 49: boxplot for the T_amp (left) and T_amp_mod (right) features.

The displayed EKG morphological features do not show significant differences between the two different types of heartbeats. So, no one of these features it is a good individual predictor.

3.1.3 T-wave type feature

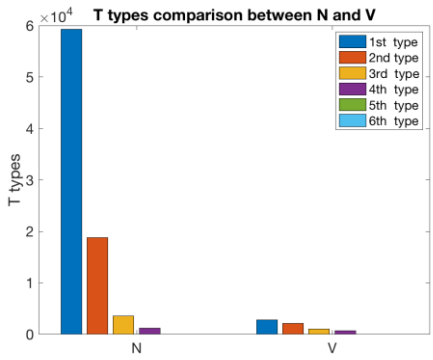


Figure 50: boxplot for the T_tp feature.

In this case, although the quantities of every type of signal are larger for the N type, it seems like the N and V heartbeats follow the same proportions, so it is not a good individual predictor.

3.1.4 QRS inversion

As we already showed in **Figure 10**, the V type beats sometimes had their QRS inverted. This is experimentally confirmed by the algorithm developed to detect QRS inversion, and its results are displayed in **Figure 51**,

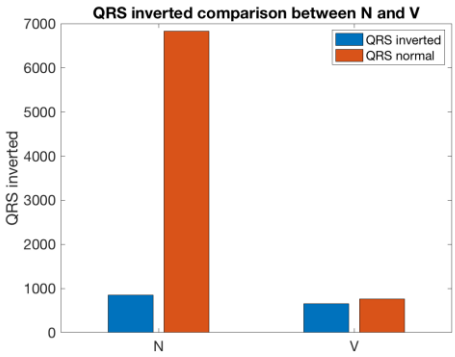


Figure 51: boxplot for the QRS_inv feature.

Even though QRS_inv seems to be a good individual predictor, training of the classifier revealed that this is not included in the combination of features that returns the better results.

3.2 Pairs of features

Aside from the individual study of the features, we also performed an analysis based on the pair aggrupation of the heartbeat interval and EKG morphology features. In particular, the scatterplots of the following combinations that were studied:

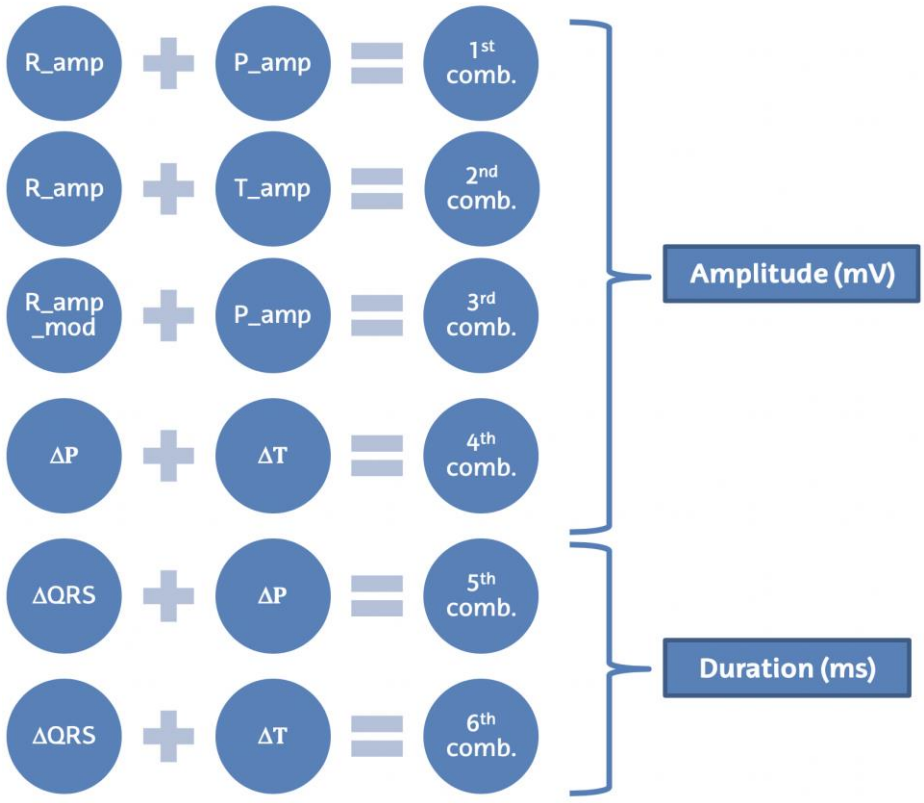


Figure 52: the studied different feature pair combinations.

In most of the cases no conclusive results could be extracted from these comparations, but there were cases, like the one shown in **Figure 53**, where interesting things could be observed.

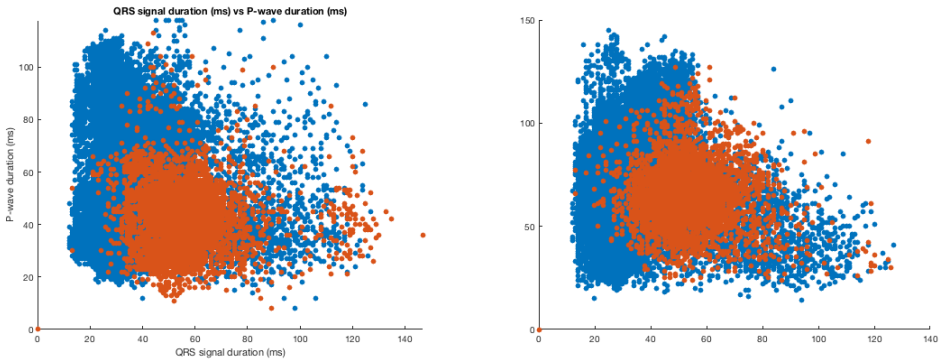


Figure 53. Left: ΔQRS vs ΔP. N heartbeats (blue) and V heartbeats (orange)**. Right:** ΔQRS vs ΔP. N heartbeats (blue) and V heartbeats (orange).

The figure shows that the values for the V heartbeats are concentrated in one circular area. Therefore, if the value for both features coincides outside this area, it will most likely be a N heartbeat. However, if the value is inside of that circle, it cannot be precisely determined whether the heartbeat is N or V.

4 Examples of classification errors

The developed classifier detects the 90.98% of ventricular heartbeats and 85.98% of normal heartbets. Although these results are very good, the algorithm fails in the 14.02% of the available heartbeats. Some examples of this classification errors are shown in the following figures, in which the output of the algorithm is marked on top of each heartbeat. Correct decisions are displayed in blue, and incorrect decisions in red.

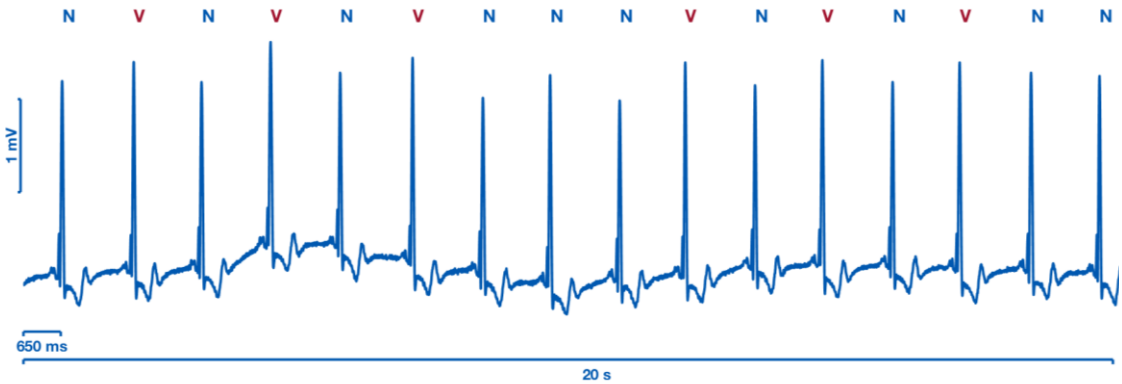


Figure 54: the first 20s from the recording '124'. The algorithm incorrectly classifies 6 of 16 displayed heartbeats.

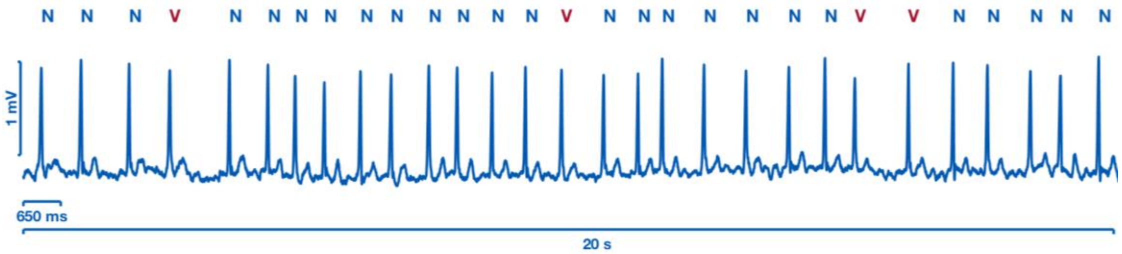


Figure 55: the last 20s from the recording '201'. The algorithm incorrectly classifies 4 of 29 displayed heartbeats.

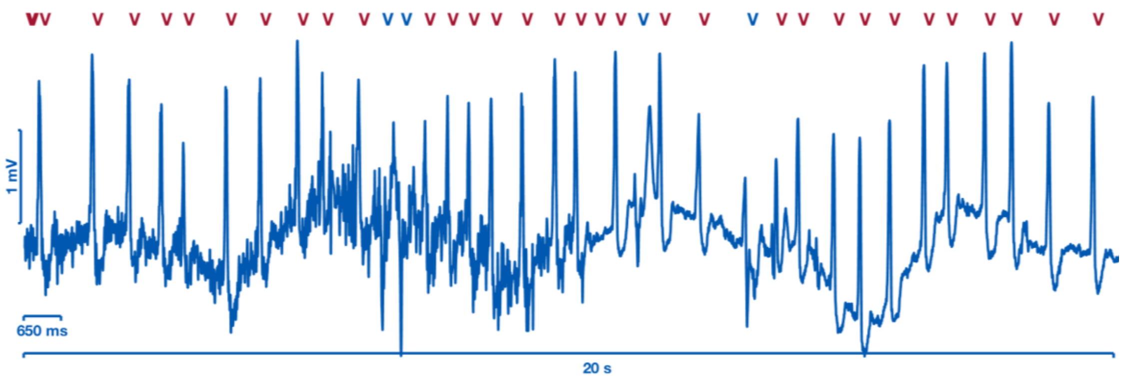


Figure 56: the first 20s from the recording '203'. The algorithm incorrectly classifies 29 of 37 displayed heartbeats.

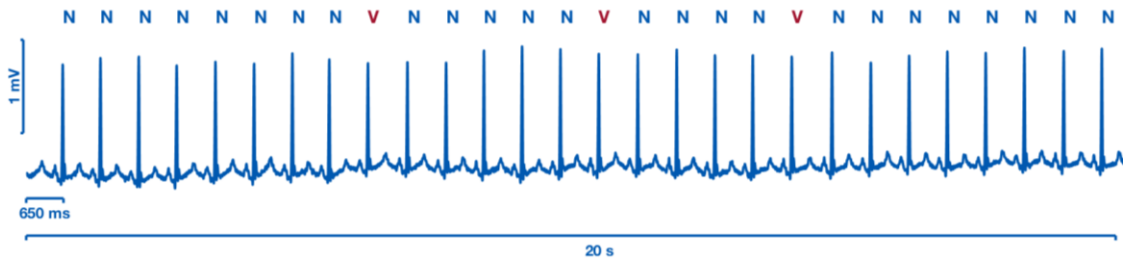


Figure 57: the first 20s from the recording '205'. The algorithm incorrectly classifies 3 of 28 displayed heartbeats.

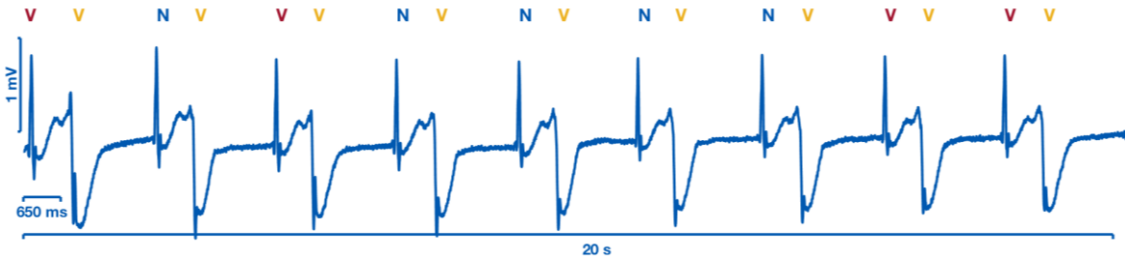


Figure 58: the first 20s from the recording '207'. The algorithm incorrectly classifies 4 of 18 displayed heartbeats.

We can appreciate that in large amplitude QRSs and durations (**Figure 54** and **Figure 57**), and noisy environments (**Figure 56),** the classifier tends to classify the heartbeat as V, although they are frequently N but with slight deviations from the standard N heartbeats.