Universidade de Coimbra

Faculdade de Ciências e Tecnologia

Departamento de Física

Algorithm development for physiological signals analysis and cardiovascular disease diagnosis - A data mining approach

João André Peixoto Vieira

Coimbra, 2011

Algorithm development for physiological signals analysis and cardiovascular disease diagnosis - A data mining approach

Scientific Supervisors

PhD Professor Carlos M. B. A. Correia

PhD João Manuel Rendeiro Cardoso

Supervisor

MsC Vânia Gomes de Almeida

Dissertation submitted to the
Faculty of Sciences and Technology
In partial fulfillment of the requirements
For the MSc degree in Biomedical Engineering

Physics Department
Faculty of Sciences and Technology
University of Coimbra

Coimbra, September 2011

This work was supported by *Fundação para a Ciência e Tecnologia* through the grant PTDC/SAU-BEB/100650/2008

COMPETE

QUADRO
DE REFERÊNCIA
ESTRATÉGICO
NACIONAL

UNIÃO EUROPEIA
Fundo Europeu
de Desenvolvimento Regional

Aos meus pais que sempre me apoiaram.

Acknowledgements

I would like to thank my scientific supervisors, Prof. Dr. Carlos Correia and Dr. João Cardoso for their support and for receiving me in such a rich environment as GEI, allowing me to develop skills in a wide range of matters.

A special thank to supervisor, Eng. Vânia Almeida, for the constant support, availability and friendship along this year.

Also, to my good friend Luís Martins, I will never forget your support and encouragement when I most needed it. To all my friends for making my life “outside the walls” much better, for all the laughing and for making who I am today. To my dear Inês Fonseca, for everything along these two years.

Most important of all, I would like to thank to all my family, my sister and especially my father and my mother for all the sacrifices they made to support me during these five years helping me whenever I needed.

Abstract

According to the World Health Organization, cardiovascular diseases (CVD) are the leading cause of death worldwide. The pulse waveform analysis is the basis of non-invasive methods to address this problem. Clinical relevant information, extracted from waveforms, allows the quantification of important parameters, e.g., pulse wave velocity (PWV), augmentation index (AI), heart rate (HR) and cardiac output (CO).

A versatile platform capable of different hemodynamic parameters measurements is being developed in order to achieve a more comprehensive diagnosis of the cardiovascular system. A method based in spatial features extraction from APW has been developed, consisting in the attribution of several fiduciary points and vectors to the APW. The APW signal pre-processing includes baseline shift removal and a morphological analysis for anomalous beats detection and elimination.

Data mining tools, such as classification and clustering techniques, have been used for relationship identification and pattern recognition in cardiac pulse waveform signals, leading to knowledge extraction from complex data. A classification model based in Random Forest algorithm was constructed in order to distinguish healthy people from people with a predisposition to develop cardiovascular diseases.

Keywords: Arterial pressure waveform, space temporal patterns, piezoelectric sensor, data mining, arterial stiffness.

Resumo

Segundo a Organização Mundial de Saúde, as doenças cardiovasculares são a principal causa de morte em todo o mundo. A análise da forma da onda de pressão arterial é a base de métodos não invasivos para encarar este problema. Informação com relevância clínica extraída da onda de pressão arterial permite a quantificação de parâmetros como a velocidade de onda de pulso, o índice de augmentação, frequência cardíaca e débito cardíaco.

Uma plataforma versátil capaz de medir diferentes parâmetros hemodinâmicos está a ser desenvolvida de forma a produzir um diagnóstico mais abrangente sobre o sistema cardiovascular. Foi desenvolvido um método que recorre à extracção de características espaciais da onda de pressão arterial que consiste na identificação de vários pontos e vectores desta onda. O pré-processamento do sinal de onda de pressão arterial inclui a remoção da linha de base e uma análise morfológica para a detecção de pulsos anómalos e consequente eliminação.

Foram utilizadas ferramentas de mineração de dados, como algoritmos de classificação e aglomeração, de forma a identificar e reconhecer padrões em sinais de onda de pressão arterial, levando à extracção de conhecimento a partir de dados complexos. Um modelo de classificação foi construído com base no algoritmo de aprendizagem automática Random Forest, de forma a distinguir indivíduos com predisposição a desenvolverem doenças cardiovasculares.

Palavras-chave: Onda de pressão arterial, padrões temporais e espaciais, sensor piezoelectrico, mineração de dados, rigidez arterial.

Contents

List of Figures

List of Tables

Acronyms

AGE	Advanced Glycation Endproducts
AI	Augmentation Index
APW	Arterial Pressure Waveform
ARFF	Attribute-Relation File Format
BMI	Body Mass Index
BN	Bayesian Networks
CV	Cardiovascular
CVD	Cardiovascular Disease
CWT	Continuous Wavelet Transform
DBP	Diastolic Blood Pressure
DN	Dicrotic Notch
DP	Dicrotic Peak
ECG	Electrocardiogram
FN	False Negative
FP	False Positive
GUI	Graphical User Interface
HEE	Human Expert Engineer
HR	Heart Rate
ISH	Isolated Systolic Hypertension
MMP	Matrix Metalloproteinase
PP	Pulse Pressure
PWV	Pressure Wave Velocity
PZ	Piezoelectric
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
RMSE	Root Mean Square Error
RP	Reflected Point
SBP	Systolic Blood Pressure
SNR	Signal-to-Noise Ratio
SP	Systolic Peak
TN	True Negative
TP	True Positive

1. Introduction

1.1. Motivation

According to the World Health Organization, cardiovascular diseases (CVDs) are the leading cause of death worldwide and continue to expand. Developing feasible surveillance methods to assess the pattern and trends of major cardiovascular diseases is expected to improve the early diagnoses, leading to early treatment, a vital procedure for CVD prevention, anticipating serious complications, such as cardiac stroke.

Hypertension is already a highly prevalent cardiovascular risk factor worldwide, the effective diagnosis of hypertension contributes to extend and enhance life. However, hypertension remains inadequately managed everywhere [1]. The assumption that hypertension is exclusively due to an increase in vascular resistance is untrue and neglects the effect of the pulsatile pressure and flow. The contribution of arterial shape is essential in the management of cardiac pathologies, e.g. for the arterial pressure values, for the same value of mean arterial pressure several shapes of the blood pressure curve may be recorded [2].

Arterial pressure waveform (APW) analysis is a non-invasive method that accurately determines the cardiac status based on the morphological analysis of a complex signal, allowing the extraction of clinical relevant information. APW is determined by the interaction of the heart pump with the multiple vessels that comprises the arterial tree and contains a vast amount of pathophysiological information concealed in its morphology.

Cardiac catheterization still is the “gold standard” method for CVDs diagnosis, but this procedure is highly invasive and costly, making it impractical in large population trials. The development of non-invasive methods to address this purpose remains a main interest for many researchers. Among the most widely methods to obtain APW non-invasively are applanation tonometry, where a superficial artery is flattened against a bone structure, and ultrasounds, which could provide an alternative to this purpose by recording diameter waveforms [3] but time resolution limits its utilization due the low frame rate associated to this technique [4]. The use of piezoelectric (PZ) sensors for APW measurement was reported by several authors [5, 6], having a good performance in *in vivo* acquisitions performed in the carotid artery.

In recent years, computer-aided diagnose methodologies based in data mining approaches have been widely used, with great developments in medical applications. Due to great the data amount and complexity in cardiac setups, data mining techniques may be useful to deepen knowledge about CVDs and to perform early diagnosis [7].

1.2. Goals

This work aimed to develop a versatile platform tool for APW analysis and algorithms for time-spatial features extraction from APW signals, in order to evaluate data parameters using data mining techniques for patterns recognition and model construction. This process allows the analysis of extensive amounts of time-spatial parameters to discover hidden relationship and patterns in data. It can also be used to predict the outcome of future observations or to assess the potential risk to develop a cardiovascular stiffness, allowing diagnosis prediction.

1.3. Thesis contents

This dissertation is divided in seven chapters. In the first chapter the framework and objectives of this project are referred, describing the main reasons for its implementation, the goals and lastly the organizational structure of the entire thesis.

In the second chapter, a theoretical background is presented, addressing the main concepts of the cardiovascular system, arterial stiffness and its determinant factors, the main instruments and indexes available for non-invasive arterial stiffness assessment and the main concepts of electrocardiography.

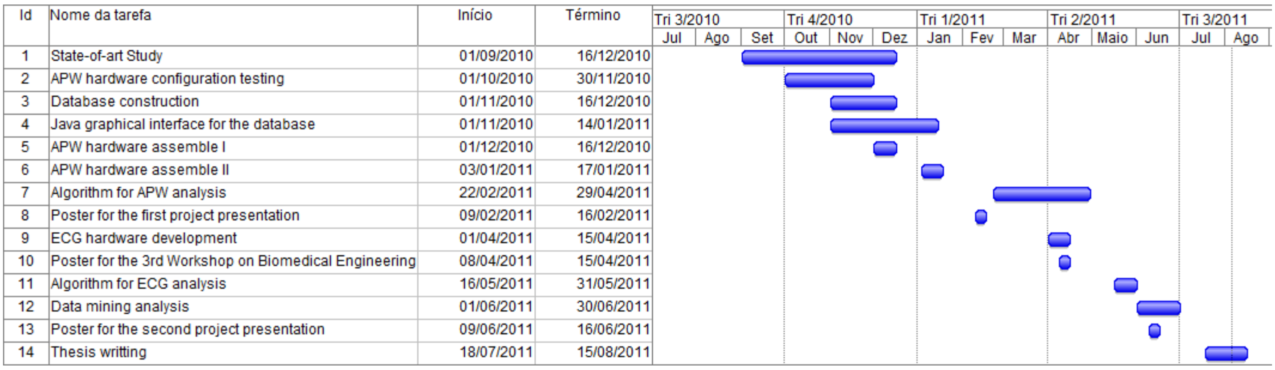
In the third chapter a quick approach of the data mining concept is given, with particular interest to the most common classification and clustering algorithms.

The fourth chapter presents all the developed hardware. The developed acquisition system is composed by an APW module and an ECG module. The APW multiprobe construction is also addressed.

All software methods are presented in the fifth chapter. The signal processing part includes APW and ECG signal analysis. Particularly, APW analysis explains the methods for data pre-processing, bad pulse identification and removal and spatial features extraction. The constructed database and interface are also reported. Lastly, the data mining procedures for model construction and clustering are approached. Results obtained from the followed methodology are presented in chapter six.

Finally, in the last chapter the final remarks are drawn and a summary of the developed work is given. A vision of possible future developments is also briefly discussed.

Table 1: Gantt diagram of project tasks planning;



Id	Nome da tarefa	Início	Término
1	State-of-art Study	01/09/2010	16/12/2010
2	APW hardware configuration testing	01/10/2010	30/11/2010
3	Database construction	01/11/2010	16/12/2010
4	Java graphical interface for the database	01/11/2010	14/01/2011
5	APW hardware assemble I	01/12/2010	16/12/2010
6	APW hardware assemble II	03/01/2011	17/01/2011
7	Algorithm for APW analysis	22/02/2011	29/04/2011
8	Poster for the first project presentation	09/02/2011	16/02/2011
9	ECG hardware development	01/04/2011	15/04/2011
10	Poster for the 3rd Workshop on Biomedical Engineering	08/04/2011	15/04/2011
11	Algorithm for ECG analysis	16/05/2011	31/05/2011
12	Data mining analysis	01/06/2011	30/06/2011
13	Poster for the second project presentation	09/06/2011	16/06/2011
14	Thesis writting	18/07/2011	15/08/2011

2. Theoretical Background

In this chapter, a theoretical background will be presented, addressing the main concepts of the cardiovascular system, arterial stiffness and its determinant factors, the main instruments and indexes available for non-invasive arterial stiffness assessment and the main concepts of electrocardiography.

2.1. Cardiovascular system

In order that every cell in the human body easily exchange products, energy and momentum with the environment, the physiologic system is endowed with the cardiovascular system that is mainly composed by the heart, blood and blood vessels [8].

2.1.1. Heart

Acting like a pump, the heart is one of the most vital organs in the human body. With each heartbeat, blood is pushed into the arteries and through veins, delivering oxygen to and removing carbon dioxide from organs, tissues and cells [9].

The heart consists of a tough muscular wall, the myocardium, which alternating contractions and relaxations cause the heartbeats and blood pumping. The myocardium is covered in the outside by a thin layer of tissue, the pericardium, while the inside is covered by the endocardium. The heart is divided by the interventricular septum, a tough muscular wall [9].

In order to separate the arterial blood from the venous, the heart is divided in four chambers (two atria and two ventricles). The heart valves are membranous structures that facilitate the circulation of blood through the heart in one direction, from the atria to the ventricles (atrioventricular valves) and those for the pulmonary artery and the aorta (sigmoid or semilunar valves). The tricuspid valve makes the connection between the atrium and right ventricle, the left mitral valve ensures the connection of the right ventricle to the left ventricle, the pulmonary semilunar valve allows the blood to flow from the right ventricle to the left pulmonary artery and the aortic valve ensures the connection from the left ventricle to the aorta [9, 10]. Figure 1 shows a cross section of the human heart.

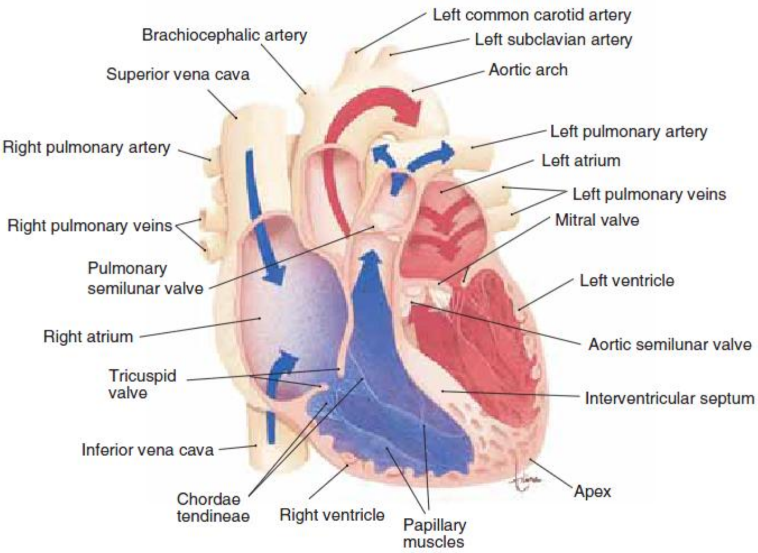


Figure 1: Cross section of the human heart [9];

Because of the anatomic proximity of the heart to the lungs, the right side of the heart does not have to work very hard to drive blood through the pulmonary circulation, so it functions as a low-pressure ($P \leq 40$ mmHg gauge) pump compared with the left side of the heart, which does most of its work at a high pressure (up to 140 mmHg gauge or more) to drive blood through the entire systemic circulation to the furthest extremes of the organism [8].

Rhythmic cardiac contractions are originated with an electrical impulse that travels from the top of the heart in the atria to the bottom of the heart in the ventricles. The period of relaxation is called diastole and the period of contraction is called systole. Diastole is the longer of the two phases so that the heart can rest between contractions [9].

2.1.2. Circulatory routes

There are two major blood circulatory routes, the systemic and the pulmonary circulation. In figure 2 shows a schematic illustration of the circulatory routes, with the deoxygenated (in blue) and oxygenated (red) blood.

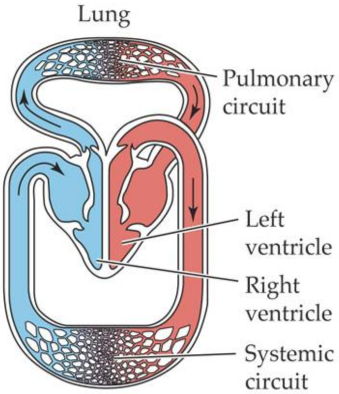


Figure 2: Deoxygenated (in blue) and oxygenated blood (in red) [11];

2.1.2.1. Systemic Circulation

The systemic circulation supplies oxygenated blood to and returns deoxygenated blood from the tissues of the body, through a circuit of vessels. The left ventricle pumps the blood from the heart through the aorta and arterial branches to the arterioles and through capillaries, where it reaches the tissue fluid, and then drains through the venules into the veins and returns, via the vena cava, to the right atrium of the heart [10].

2.1.2.2. Pulmonary Circulation

The pulmonary circulation consists of a system of blood vessels that forms a closed circuit between the heart and the lungs. The pulmonary trunk passes diagonally upward to the left across the route of the aorta. The trunk divides into two branches (the right and left pulmonary arteries) which enter the lungs. Then, the branches go through a process of subdivision, being the final branches the capillaries. At the capillaries surrounding the alveoli of the lungs, supply is replenished and its carbon dioxide content is purged. The capillaries carrying oxygenated blood join larger vessels until they reach the pulmonary veins, which carry oxygenated blood from the lungs to the left atrium of the heart [9, 10].

2.1.3. Common carotid artery

The common carotid, internal carotid, and external carotid arteries provide the major source of blood to the head and neck. Additional arteries arise from branches of the subclavian artery, particularly the vertebral artery. The common carotid arteries differ, with respect to their origins, on the right and left sides. On the right, the common carotid arises from the brachiocephalic artery as it passes behind the sternoclavicular joint. On the left, the common carotid artery comes directly from the arch of the aorta in the superior mediastinum [10].

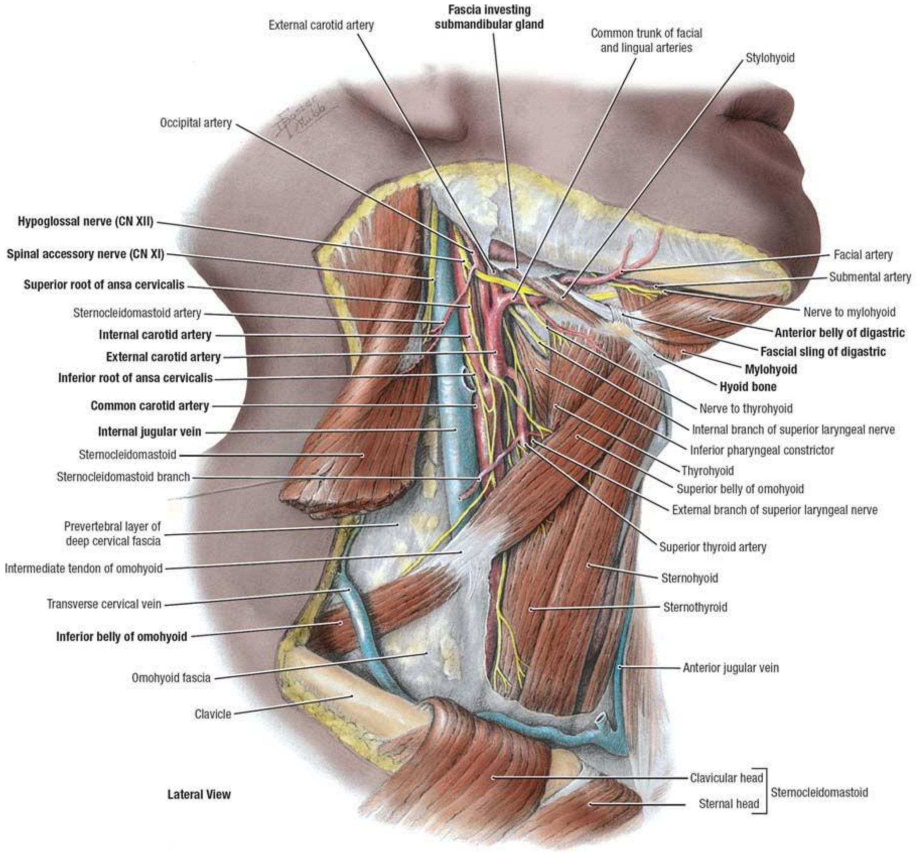

External carotid artery
Fascia investing submandibular gland
Common trunk of facial and lingual arteries
Stylohyoid
Occipital artery
Hypoglossal nerve (CN XII)
Spinal accessory nerve (CN XI)
Superior root of ansa cervicalis
Sternocleidomastoid artery
Internal carotid artery
External carotid artery
Inferior root of ansa cervicalis
Common carotid artery
Internal jugular vein
Sternocleidomastoid
Sternocleidomastoid branch
Prevertebral layer of deep cervical fascia
Intermediate tendon of omohyoid
Transverse cervical vein
Inferior belly of omohyoid
Omohyoid fascia
Clavicle
Lateral View
Facial artery
Submental artery
Nerve to mylohyoid
Anterior belly of digastric
Fascial sling of digastric
Mylohyoid
Hyoid bone
Nerve to thyrohyoid
Internal branch of superior laryngeal nerve
Inferior pharyngeal constrictor
Thyrohyoid
Superior belly of omohyoid
External branch of superior laryngeal nerve
Superior thyroid artery
Sternohyoid
Sternothyroid
Anterior jugular vein
Clavicular head
Sternal head
Sternocleidomastoid

Figure 3: Vessels and nerves of the neck, right lateral view [12];

2.2. Arterial Stiffness

Nowadays, arterial stiffness and wave reflections are well accepted markers of cardiovascular (CV) risk, being the most important parameters of increasing systolic and pulse pressure and thus as the cause of CV complications and events [13, 14, 15].

Arterial stiffness measures the rigidity of the arterial wall, in other words, it defines the arteries capacity to expand and contract during the cardiac cycle. Structural components of the arterial wall, vascular smooth muscle tone and transmural distending pressure determine arterial stiffness [13, 14].

Arterial stiffness increases the velocity at which the pulse wave travels, causing an early return of reflected waves in late systole and hence, suboptimal ventricular-arterial interaction, increasing central pulse pressure (PP) and thus systolic blood pressure (SBP), which increases the load on the left ventricle, increasing myocardial oxygen demand [14, 15].

In addition, arterial stiffness is associated with left ventricular hypertrophy, a known risk factor for coronary events in normotensive and hypertensive patients. The increase in central PP and the decrease in diastolic BP may directly cause myocardial ischemia and increases the pressure-induced damage on coronary and cerebral arteries. The measurement of aortic stiffness may also reflect parallel lesions present at the site of the coronary arteries [14, 15].

Although not synonymous, compliance, distensibility and elasticity are interrelated aspects of arterial stiffness. Compliance is used to define the change in volume for a given pressure change, reflecting the change in artery diameter caused by left ventricular ejection. Distensibility defines compliance relative to the initial volume or diameter of an artery. Reduced arterial compliance and distensibility is a result of a loss of arterial elasticity. When pressure increases, a point is eventually reached with less distensibility occurring at higher pressures as a consequence of the elastic properties of the arterial media. At low pressures elastin fibres stands pressure, while at higher pressures the tension is absorbed by the rigid collagen fibres resulting in a decrease of the compliance [14].

2.2.1. Proximal and distal arterial stiffness

The elastic properties of conduit arteries vary along the arterial tree, with more elastic proximal arteries and stiffer distal arteries as a result of different molecular, cellular, and histological structure of the arterial wall [16].

Along a viscoelastic tube without reflection sites, a pressure wave is progressively attenuated, with an exponential decay along the tube, whilst a pressure wave which propagates along a viscoelastic tube with numerous branches is progressively amplified due to wave reflections, from central to distal conduit arteries. The result is that the amplitude of the pressure wave is higher in peripheral arteries than in central arteries. Thus, it is not accurate to use brachial pulse pressure as a surrogate for aortic or carotid pulse pressure, particularly in young subjects [16].

2.2.2. Factors that affect Arterial Stiffness

A large number of pathophysiological conditions are associated with increased arterial stiffness, where age and blood pressure play the leading roles when evaluating the degree of arterial stiffness. Apart from these dominant conditions, several others are reported in the table bellow.

Table 2: Factors that affect arterial stiffness [16];

Ageing	**CV risk factors**	**CV diseases**
Other physiological conditions	-Obesity	-Coronary heart disease
-Low birth weight	-Smoking	-Congestive heart failure
-Menopausal status	-Hypertension	-Fatal stroke
-Lack of physical activity	-Hypercholesterolaemia	**Primarily non-CV factors**
Genetic Background	-Impaired glucose tolerance	-ESDR
-Parental history of hypertension	-Metabolic syndrome	-Moderate chronic kidney -disease
-Parental history of diabetes	-Type 1 diabetes	-Rheumatoid arthritis
-Parental history of myocardial infarction	-Type 2 diabetes	-Systemic vasculitis
-Genetic polymorphisms	-Hyperhomocyteinaemia	-Systemic lupus erythematosus
	-High CRP level	

2.2.2.1. Age

Stiffening of large arteries is a consequence of the normal aging process and age is the most important determinant of arterial stiffness. The most consistent and well-reported changes are luminal enlargement with wall thickening (remodeling) and a reduction of elastic properties (stiffening) at the level of large elastic arteries. However, this aging process in the arterial tree is heterogeneous and while the large central arteries stiffen progressively with age, the elastic properties of the smaller muscular arteries change little with age [14, 17].

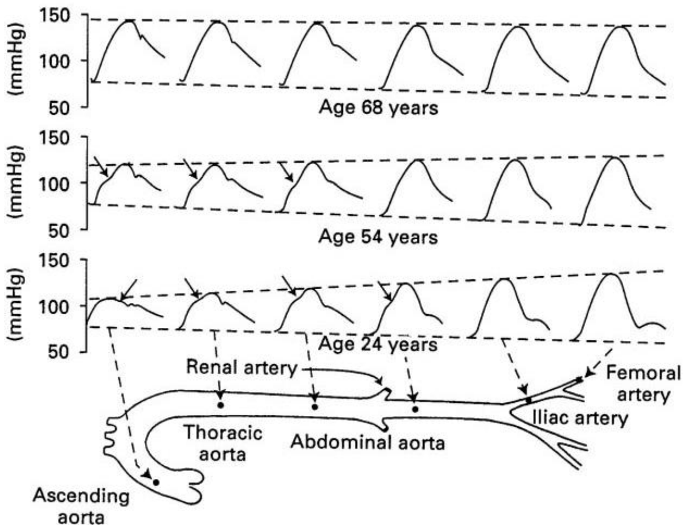


Figure 4: APW variation at different ages and locations in the arterial tree [2];

Large arteries are mainly composed by vascular smooth muscle cells and elastic and collagen fibers. With aging, medial degeneration takes place, which leads to progressive stiffening of the large elastic arteries. Accumulation of advanced glycation endproducts (AGE) on the structural matrix proteins alters their physical properties and causes stiffness of the fibers (figure 5). Another major change in the arterial wall is the increasing of calcium deposition, which might also contribute to the loss of arterial distensibility [17].

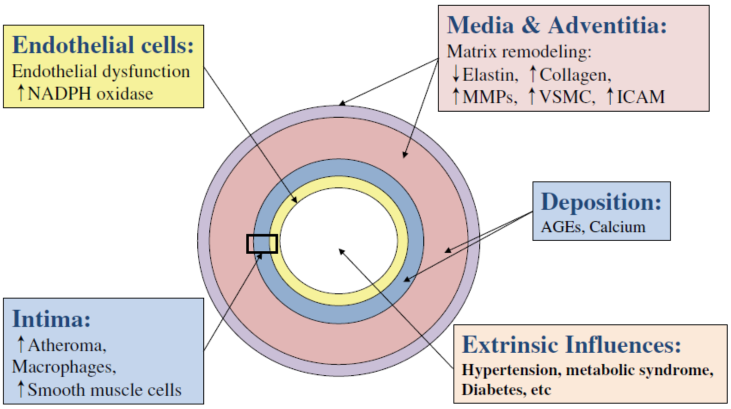


Figure 5: Causes of arterial aging [17];

2.2.2.2. Hypertension

Although large artery stiffening is a strongly age-related process, it is also markedly accelerated by the presence of hypertension. With aging, degeneration of compliant elastin fibers, and deposition of stiffer collagen, is considered a key cause of arterial stiffening. Moreover, blood pressure remodels vessel wall structure to compensate changes in wall stress. One mechanism of vessel wall reshuffle is through matrix metalloproteinases (MMPs), which modulate extracellular matrix proteins by enhancing collagen degradation. This way, the intrinsic distensibility of elastic arteries is improved and, thus, blunts any blood pressure rise. Therefore this compensatory mechanism increases stiffness [14, 18].

Table 3: International classification of hypertension according to blood pressure (BP) level [2];

	Systolic Blood Pressure (mm Hg)		Diastolic Blood Pressure (mm Hg)
Normotensive	< 140	and	< 90
Mild hypertension	140-180	and/or	90-105
Subgroup: Borderline hypertension	140-160	and/or	90-95
Moderate and severe hypertension	>180	and/or	>105
Isolated systolic hypertension (ISH)	>160	and	<90
Subgroup: borderline ISH	140-160	and	<90

2.2.3. Non-invasive determination of arterial stiffness

Several non-invasive methods are currently used to assess vascular stiffness. Unlike systemic arterial stiffness, which can only be estimated from models of the circulation, regional and local arterial stiffness can be directly determined noninvasively, at various sites along the arterial tree. Thus, regional and local evaluations of arterial stiffness allow direct measurements of parameters strongly linked to wall stiffness [16].

There are several noninvasive methods available for determination of local, regional and systemic arterial stiffness. According to Laurent *et al.* (2006) [16], the main features of the currently available methods are described in the following table:

Table 4: Available methods for arterial stiffness measures and wave reflections [16];

	Device	Methods	Measuring site
Regional Stiffness	Complior®	Mechanotransducer	Aortic PWV[a]
	Sphygmocor®	Tonometer	Aortic PWV[a]
	WallTrack®	Echotracking	Aortic PWV[a]
	Artlab®	Echotracking	Aortic PWV[a]
	Ultrasound Systems	Doppler probes	Aortic PWV[a]
Local Stiffness	WallTrack®	Echotracking	CCA[b], CFA, BA
	NIUS®	Echotracking	RA
	Artlab®	Echotracking	CCA[b], CFA, BA
	Various vascular ultrasound systems	Echotracking	CCA[b], CFA, BA
	MRI device	Cine-MRI	Ao
Systemic Stiffness (waveform shape analysis)	Area method	Diastolic decay	
	HDI PW CR-2000®	Modif. Windkessel	
	SV/PP	Stroke volume and pulse pressure	
Wave Reflections	Sphygmocor®	AIx	All superficial artery
	Pulse Trace ®	Finger photoplethysmography	Finger

Ao, aorta; CCA, Common carotid artery; BA, Brachial artery; RA, Radial artery; SV/PP, Stroke volume/pulse pressure.

[a] Aorta, carotid-femoral, also carotid radial and femoro-tibial PWV.

[b] All superficial arteries.

2.3. Arterial Pressure Waveform

Since the first pressure waveform recording in the 19th century using a sphygmograph, numerous methods for arterial waveform analysis have been used including invasive methods like catheterization and, more recently, non invasive methods such as applanation tonometry [19].

APW should be analysed at the central level, i.e. the ascending aorta, since it surrogates the true load imposed to the left ventricle and central large artery walls. Unlike radial or brachial arteries, the measurement of APW at carotid arteries doesn’t use a transfer function because carotid and central arteries have similar waveforms, although it requires a higher degree of technical expertise [16, 20].

2.3.1. Morphology of APW

The arterial pressure waveform is composed by a forward travelling incident wave, caused by left ventricular contraction, and a reflected wave returning to the heart, caused by arterial tree branch points or sites of impedance mismatch [2, 19, 21].

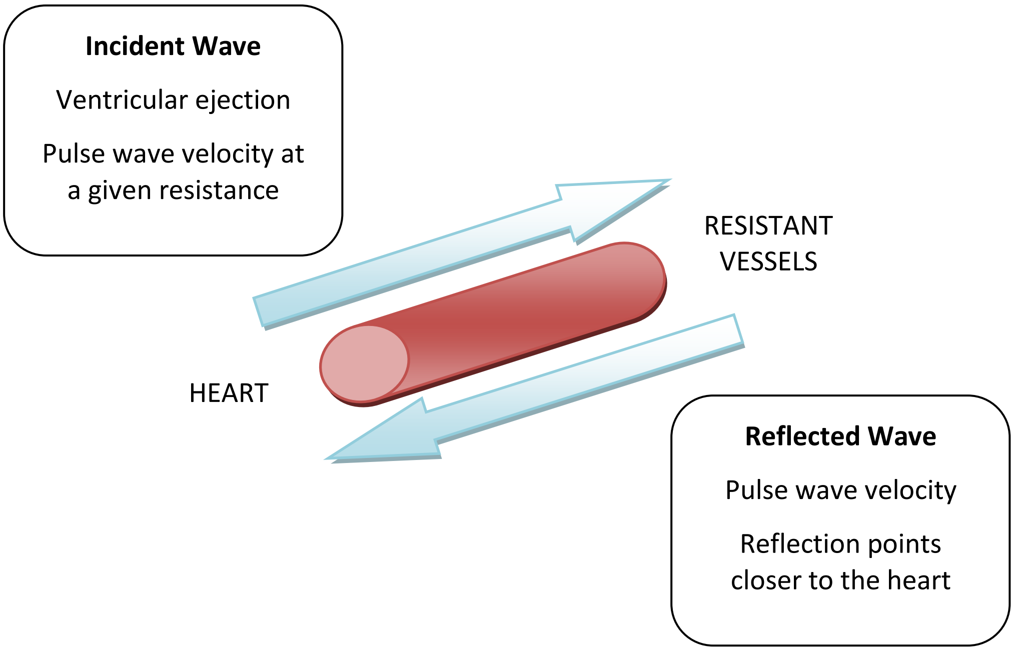


Figure 6: The arterial pressure waveform results of the summation between the incident and the reflected pressure wave [adapted from 2];

Based on the morphology of the APW, a classification was proposed by Murgo (1980) [22], where the determinant criterion for wave classification is the location of the reflected wave [20, 22].

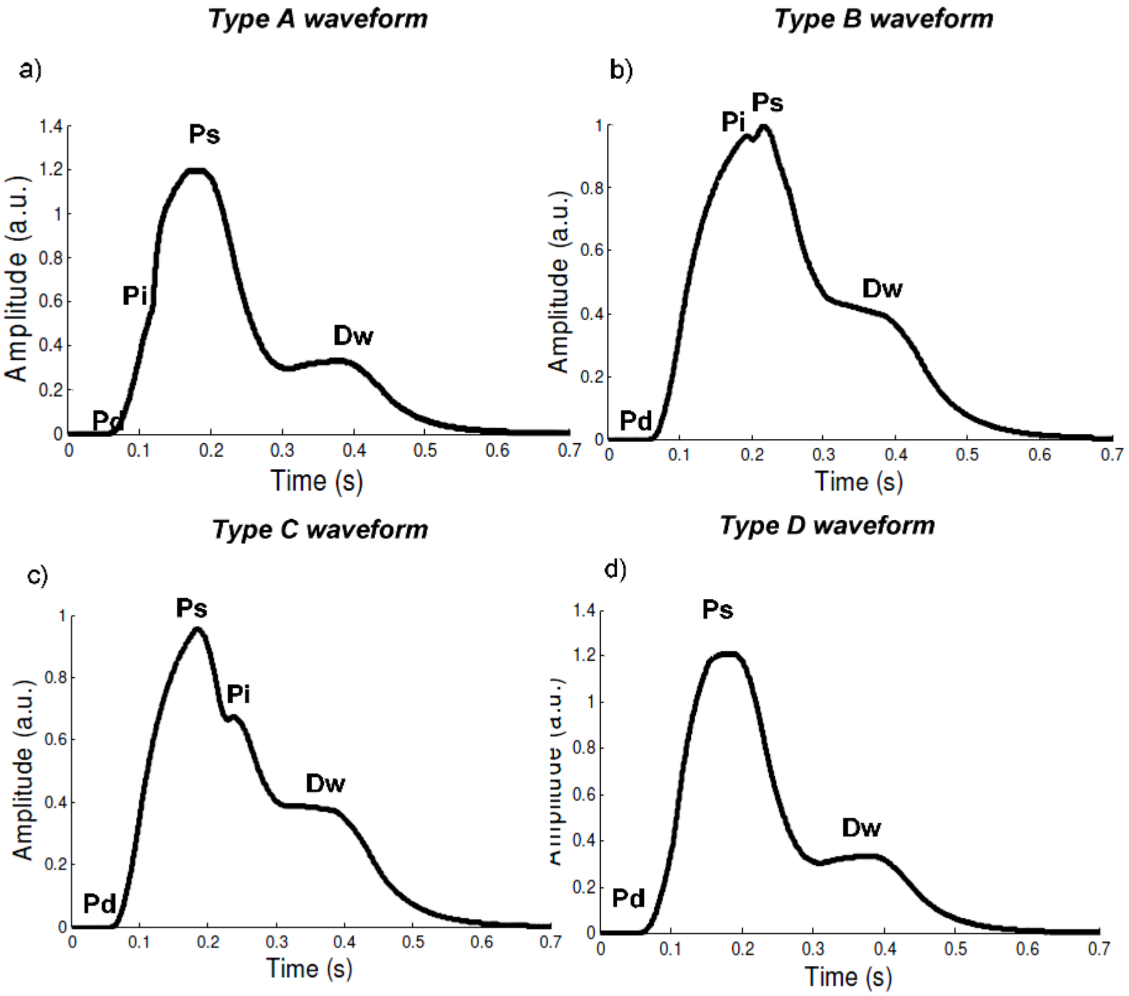


Figure 7: Classification of typical APW according to Murgo, where Pd is the diastolic pressure, Pi is the inflection point, Ps is the systolic pressure and Dw is the dicrotic wave [20];

2.3.1.1. Incident pressure wave

In the aorta, the incident wave occurs due to the capacitive (storage) effects of the ascending aorta segment [23]. The characteristics of the forward wave depend on the left ventricular ejection and stiffening of the aorta, not being influenced by wave reflections [2, 19, 21].

After the closure of the aortic valve, an increase in the aortic pressure along the ascending aorta takes place. This phenomenon is known as incisura and results as a reaction of the aortic pressure to the closure of the aortic valve and can be used to obtain systolic duration [20, 23].

2.3.1.2. Reflected wave

The characteristics of the backward reflected wave depend on the value of reflection coefficients, elastic properties of the arterial tree and the site of reflection points [2, 21].

For given values of reflection coefficients, increased pulse wave velocity and reflection sites closer to the heart produce more pronounced aortic backward wave, with a more substantial summation of forward and backward waves, higher pulse pressure and higher systolic peak [2].

2.4. Hemodynamic Parameters

In this section the most used hemodynamic parameters used in clinical practice will be discussed: pulse pressure, pulse wave velocity (PWV), augmentation index (AI), distensibility and compliance obtained from ultrasonography.

2.4.1. Pulse Pressure

Pulse pressure constitutes a surrogate marker for arterial stiffness assessment as it is determined by cardiac output, aortic and large artery stiffness, and pulse wave reflection. It is understood as the difference between systolic and diastolic blood pressure and is strongly influenced by the properties of the arterial tree [14].

Although arterial stiffness can be easily measured by pulse pressure using a sphygmomanometer, it can be quite inaccurate due to pulse wave amplification from the aorta to the peripheral arteries. Whereas pulse wave amplification decreases with age, the usefulness of brachial pulse pressure as a marker of arterial stiffness is poor in the young but increases with age. It should also be mentioned that it is the central blood pressure that contributes most to the development of the early stages of CVDs [14].

2.4.2. Pulse Wave Velocity

Pulse wave velocity (PWV) is the speed at which the pressure wave generated by cardiac contraction travels from the aorta through the arterial tree. Although several different measurement sites can be found in the literature, the carotid-femoral pulse wave velocity is the most commonly used in the evaluation of regional stiffness. However carotid-radial PWV is also commonly used when artery stiffness is examined, it mainly measures the stiffness in the brachial artery [14].

Studies show that PWV is an independent predictor of cardiovascular disease and mortality in both hypertensive patients, in patients with end-stage renal disease, in diabetic and elderly population samples [14].

According to Moens and Korteweg (1878) [24, 25], the relationship between arterial stiffness and PWV can be described by the following equation:

$$PWV = \sqrt{\frac{E \cdot h}{2r\rho}} \tag{1}$$

where E is the elastic modulus of the vessel wall, h is the wall thickness, r is the vessel radius and ρ the blood density and it is assumed that there is no, or insignificant, change in vessel area and there is no, or insignificant, change in wall thickness [26].

In 1922, Bramwell and Hill, cited the Moens-Kortweg equation and proposed a series of substitutions relevant to observable haemodynamic measures, that relates PWV to arterial distensibility:

$$PWV = \sqrt{\frac{dP \cdot V}{dV \cdot \rho}} = \sqrt{\frac{1}{\rho D}} \tag{2}$$

where P is the pressure, V is the volume, ρ is the blood density, $dP \cdot V/dV$ represents volume elasticity and D the volume distensibility of the arterial segment [26, 27].

2.4.3. Distensibility and Compliance

By ultrasound examination of large arteries (brachial, femoral and carotid arteries), images of the arterial walls are taken, allowing to register the maximum and minimum arterial diameter. Thus, it is possible to calculate the distensibility and compliance:

$$D = \frac{\Delta A}{A_d \Delta p} \tag{3}$$

$$C = \frac{\Delta A}{\Delta p} \tag{4}$$

with D as the distensibility, Δp as the pulse pressure and $\Delta p = (p_s - p_d)$, where p_s is the systolic pressure and p_d the diastolic pressure, ΔA the pulse cross-sectional area and $\Delta A = (A_s - A_d)$ where A_s is the systolic cross-section area and A_d the diastolic cross-section area [27].

Ultrasound has the advantage of being non-invasive, but the equipment is expensive and the mastering of this technique requires plenty of time and effort [2].

2.4.4. Augmentation Index

The AI attempts to measure the strength of the reflected wave relative to the total pressure waveform, which represents the proportion of central pulse pressure that contributes to a late systolic pressure increase due to overlap between forward and reflected pressure waves [20, 28, 29]. This index has been proposed as a surrogate of arterial stiffness [20, 28].

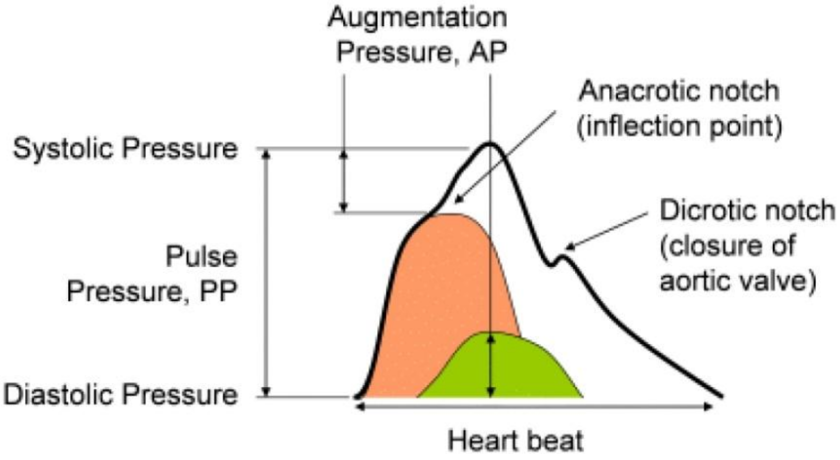

Augmentation Pressure, AP
Anacrotic notch (inflection point)
Systolic Pressure
Dicrotic notch (closure of aortic valve)
Pulse Pressure, PP
Diastolic Pressure
Heart beat

Figure 8: Augmentation pressure as the difference between the systolic and the inflection point pressure [30];

In elastic vessels, due to low PWV, reflected wave tends to arrive back at the aortic root during diastole. In other hand, in stiff arteries, PWV rises and the reflected wave arrives earlier to the central arteries, adding to the forward wave and increasing AI [16].

Table 5: Classification of the different APW according to the inflection point position and AIx calculus, where P_S is the systolic pressure, P_d the diastolic pressure and P_i the pressure in the inflection point (adapted from [20]);

APW type	APW property	Augmentation Index calculus
A	The inflection point occurs before the systolic peak. The value of AIx is positive representing larger stiffness artery.	$AI = \frac{P_s - P_i}{P_s - P_d}$
B	The inflection point occurs shortly before the systolic peak, indicating smaller arterial stiffness	$AI = \frac{P_s - P_i}{P_s - P_d}$
C	The inflection point occurs after the systolic peak. The value of AIx is negative representing that the artery is relatively elastic and healthier.	$AI = \frac{P_i - P_s}{P_s - P_d}$
D	The inflection point can't be visually detected because reflected wave arrives early in systole and merge with the incident wave.	$AI = \frac{P_s - P_i}{P_s - P_d}$

2.5. Electrocardiography

The electrocardiogram (ECG) allows the record of electrical phenomena that take place during the cardiac cycle. The electrocardiograph is a galvanometer that measures the electrical potential difference between two electrodes arranged in certain parts of the human body. This simple and non-invasive biomedical tool provides deep insights into the health status of an individual [8, 31].

2.5.1. Cellular electrophysiology

The cardiac cells, maintain a negative resting membrane potential with respect to their exterior, through a complex change of ionic concentration across the cell membranes. Through ionic pumps on the cellular membrane, ions (Na^+, K^+, Ca^+, Cl^-) are pumped in and out in order to maintain the electronegativity of the interior (resting state) [8, 31].

Due to the opening of ionic channels on the cellular membrane of cardiac cells that allow the charged ions to move along their gradient, an extracellular potential field is established which then excites neighboring cells, and a cell-to-cell propagation of electrical events occurs, resulting in a depolarization wave at the macroscopic level. When the ionic channels close down, the ionic flow is interrupted resulting in the repolarization of the cardiac cells, which means that the membrane potential returns to its resting state. Each spontaneous depolarization of these cells causes the beginning of one complete cardiac cycle [8, 31].

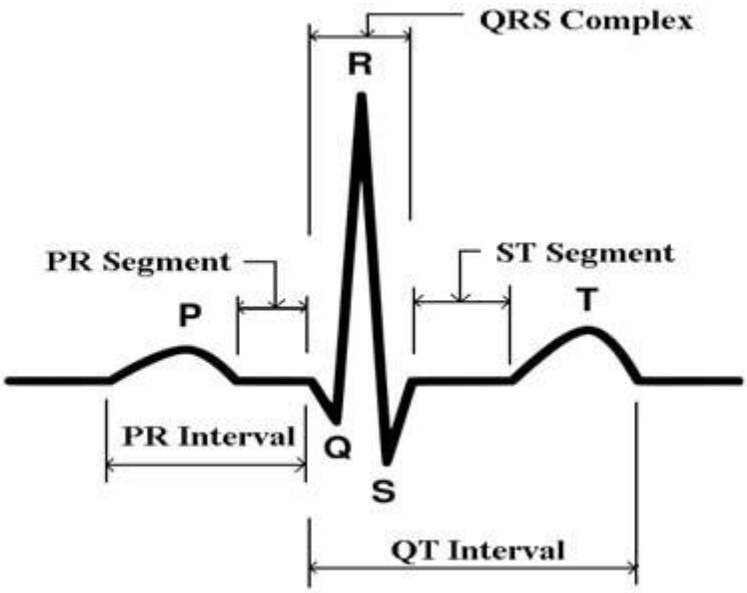


Figure 9: One cycle of a typical ECG signal showing P, Q, R, S and T waves, with segments and intervals [32];

The P wave represents the activation of the atria. Conduction of the cardiac impulse proceeds from the atria through the A-V node and the His-Purkinje system. There is a short, relatively isoelectric segment following the P wave. Once the large muscle mass of the ventricles is excited, causing them to contract and providing the main force for the blood to flow through the body. This ventricle contraction forms the QRS complex. The initial downward deflection is the Q wave, the initial upward deflection is the R wave, and the terminal downward deflection is the S wave. After this complex, appears another isoelectric segment, followed by ventricular repolarization, which results in a low frequency signal, known as T wave [31].

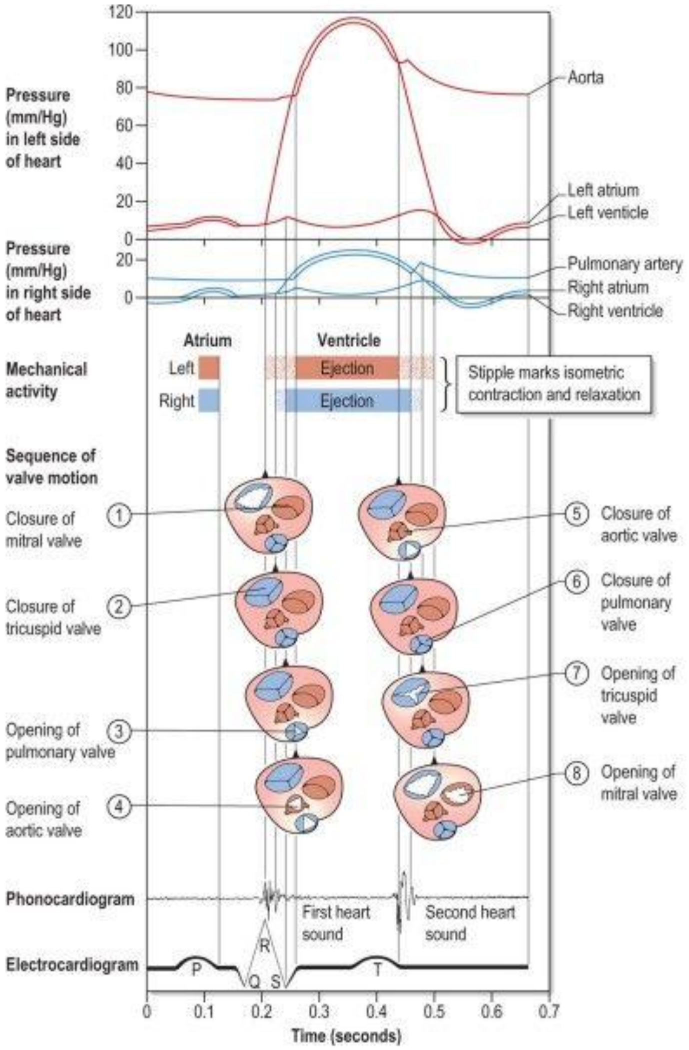

Pressure (mm/Hg) in left side of heart
120
100
80
60
40
20
0
Aorta
Left atrium
Left venticle
Pressure (mm/Hg) in right side of heart
20
0
Pulmonary artery
Right atrium
Right ventricle
Mechanical activity
Atrium
Ventricle
Left
Right
Ejection
Ejection
Stipple marks isometric contraction and relaxation
Sequence of valve motion
1 Closure of mitral valve
2 Closure of tricuspid valve
3 Opening of pulmonary valve
4 Opening of aortic valve
5 Closure of aortic valve
6 Closure of pulmonary valve
7 Opening of tricuspid valve
8 Opening of mitral valve
Phonocardiogram
First heart sound
Second heart sound
Electrocardiogram
P
Q
R
S
T
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Time (seconds)

Figure 10: The cardiac cycle. Comparison between physiological events, typical ECG waveform and APW (top). Duration of systole can be estimated by QT-interval duration [12];

3. Data Mining

Data mining is a process of learning in a practical, nontheoretical sense, searching for patterns in complex data. Besides discovering and describing structural patterns in data, the interest in data mining is to explain those patterns and make predictions from it [33]. The patterns discovered must be meaningful, leading to some advantage and allowing knowledge extraction from large amounts of data [33, 34].

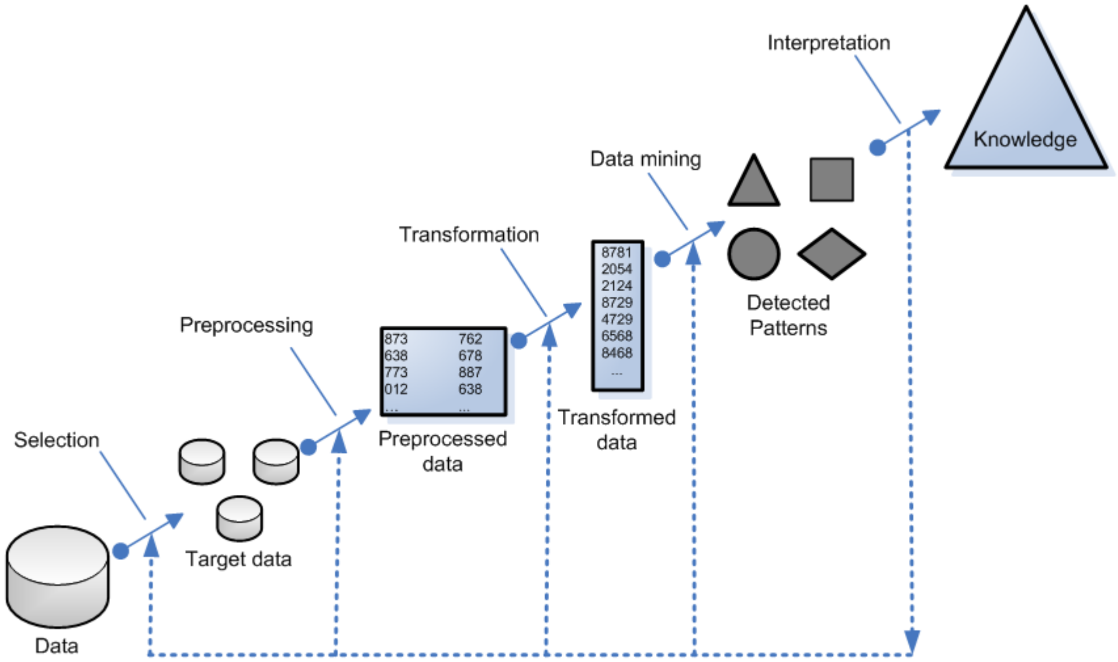


Figure 11: Steps of knowledge discovery;

3.1. Classification and prediction

Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data, consisting in several tuples described by attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes [33]. As the class label of each training tuple is already known, this form of data analysis is also known as supervised learning [33, 34].

Typically, the derived model is represented in the form of classification rules, decision trees, mathematical formulae or neural networks. After obtaining the model, the predictive accuracy is estimated using a test set of class labeled samples (different from the used training set) and if the accuracy is considered acceptable, the model can be used to classify new data tuples which the class is not labeled [34].

Prediction can be viewed as the construction and use of a model to assess the class of an unlabeled object, or to assess the value or value ranges of an attribute that a given object is more likely to have.

Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends [34].

3.1.1. Preparing the data for classification and prediction

Before applying any data mining algorithm to an available data set, a few processes need to take place in order to prepare the data in an effective manner. Although the data preparation may be one of the most time consuming step in the whole data mining process, it improves the accuracy, efficiency and scalability of the classification or prediction process [34, 35].

3.1.1.1. Data cleaning

Data cleaning refers to the preprocessing of data in order that noisy or inconsistent instances are detected, corrected or removed from the dataset. Although most classification algorithms have some mechanisms for handling noisy or missing data, they are not always robust and this step helps reducing confusion during learning. Therefore, a useful preprocessing step is to run your data through some data cleaning routines [34, 35].

3.1.1.2. Relevance analysis

Another step in the data preparation stage is feature selection. Depending on the aim of the application, some of the attributes in the data may be irrelevant to the classification or prediction task and they can even interfere with the learning mechanism of the data mining algorithm applied. Furthermore, other attributes may be redundant. Hence, relevance analysis may be performed on the data with the aim of removing any irrelevant or redundant attributes from the learning process, improving the data mining algorithm application in terms of efficiency, accuracy and generalization power [34, 35].

3.1.1.3. Data Transformation

Data transformation is particularly useful for continuous-valued attributes, allowing to discrete numerical and continuous values into ranges. This will compress the original data, decreasing the number of input/output operations that are involved in learning [34].

Another important data transformation filter is normalization. This transformation scales all values for a given attribute into a small range of values, being very useful when distance measurements are applied [34].

3.2. Decision tree

A decision tree is a flow-chart-like tree structure or model of decisions, where each internal node denotes a test on an attribute, each branch represents an outcome of the test that leads to a leaf node, representing classes or class distributions. The topmost node in a tree is the root node [34].

Decision trees are constructed in a top-down recursive divide-and-conquer manner. Starting with a training set of tuples and their associated class labels, the training set is recursively partitioned into smaller subsets as the tree is being built [33].

Nevertheless, not all branches are seen in a decision tree. Tree pruning attempts to identify and remove branches that may reflect noise or outliers, with the goal of improving classification accuracy [34].

The following algorithm explains how a decision tree is generated from the training tuples of a data partition (D):

Input: Data set of training tuples and their associated class labels (D); the set of candidate attribute (*attribute list*);

Output: Decision tree;

Method:

```
(1) create a node N;
(2) if tuples in D are all of the same class, C then
            return N as a leaf node labeled with the class C;
(3) if attribute list is empty then
            return N as a leaf node labeled with the majority class in D;
(4) select the attribute among attribute list with the highest information gain (test
    attribute);
(5) label node N with test attribute;
```

```
(6) for each known value $a_i$ of test attribute
        grow a branch from node N for the condition test attribute = $a_i$;
        let $s_i$ be the set of samples in samples for which test attribute = $a_i$;
        if $s_i$ is empty then
            attach a leaf labeled with the most common class in samples;
        else attach the node returned by Generate decision tree($s_i$, attribute
        list, test attribute) to node N;
(7) Return N;
```

3.2.1. Attribute selection measures

An attribute selection measure is a heuristic for selecting the splitting criterion that “best” separates a given data partition (D) of class-labeled training tuples, determining how the tuples at a given node are to be split. The attribute selection measure provides a ranking for each attribute describing the given training tuples, choosing the attribute having the best score as the splitting attribute for the given tuples [34].

3.2.1.1. Information gain

The information gain measure is used to select the test attribute at each node in the tree that better splits the data partition. For a given node, it is chosen as test attribute, the attribute with the highest information gain, which at the same time has the greatest entropy reduction. This attribute minimizes the information needed to classify the samples and the number of tests needed to classify an object [34].

The expected information needed to classify a tuple in D is given by

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{5}$$

where *p_i* is the probability that an arbitrary tuple in *D* belongs to the class *C_i* and is estimated by$\left|C_{i,D}\right|/|D|$. *Info(D)* represents the average amount of information needed to identify the class label of a tuple in *D*, based on the proportions of tuples of each class [33].

If the tuples in *D* were to be divided on any attribute *A* having *v* distinct values, {a_1, a_2, . . . , a_v}, splitting D into *v* pure partitions, it would mean that the attribute A should be used as a *splitting criterion*. However, most of the times, partitioning don’t produce exact classifications on the tuples [34]. The entropy or the amount of information needed in order to produce an exact classification is measured by

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j). \tag{6}$$

The smaller the entropy is, he greater the purity of the subset partitions.

Information gain is defined by the difference between the expected information necessary to classify *D* before partitioning on *A* (based on the proportion of the classes) and the actual needed information obtained after partitioning on *A* [36]. The encoding information that would be gained by branching on A is given by the information gain:

$$Gain\ (A) = Info(D) - Info_A(D). \tag{7}$$

3.2.1.2. Gain ratio

C4.5 (a decision tree classifier) uses gain ratio which tends to favor attributes that have a large number of values. This ratio is similar to information gain, although it applies a kind of normalization to information gain using a "split information" value defined as

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2\left(\frac{|D_j|}{|D|}\right). \tag{8}$$

This value, *SplitInfo$_A$(D)*, is the information due to the split of the training data set on the basis of the value of the categorical attribute A. The gain ratio is defined as

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \tag{9}$$

The splitting attribute is then selected as the attribute with maximum gain ratio [36].

3.2.2. Pruning decision trees

After a decision tree is produced by the divide and conquer algorithm, C4.5 prunes it in a single bottom-up pass. Tree pruning methods use statistical measures to remove branches that reflect anomalies in the training data due to noise or outliers. The resulting pruned tree is usually smaller and less complex and tends to be faster and better at classifying test data [34, 36].

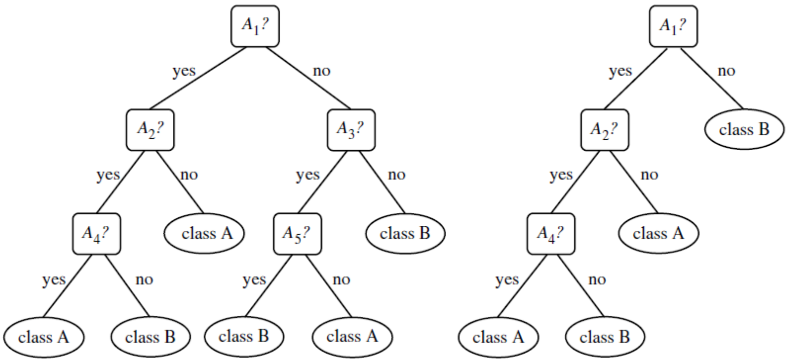


Figure 12: A version of a decision tree before prunning (left) and the prunned version of it (right) [34];

There are two mechanisms of tree pruning, prepruning (the tree is pruned by halting its construction early) and postpruning (which removes subtrees from a fully grown tree).

In prepruning, upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset tuples or the probability distribution of those tuples. In postpruning, the subtree is pruned by replacing its branches with a leaf. The leaf is labeled with the most frequent class among the subtree being replaced [34, 36].

3.2.3. C4.5 (J48)

During the first years of the 1980's, Quinlan developed a decision tree algorithm known as ID3. Later, he presented C4.5 algorithm [37] as the successor of ID3, becoming a benchmark to which newer supervised learning algorithms are often compared [34, 37]. C4.5 was later implemented in java, resulting in J48 algorithm available in Weka data mining tool [33].

Although both C4.5 and ID3 build decision trees using the concept of information entropy, some improvements were made in C4.5 allowing handling both continuous and discrete attributes, handling attributes with missing values and pruning the decision trees after creation [34, 37].

3.2.4. Random Forest

Random forest is a decision tree type machine learning algorithm first proposed by Tin Kam Ho in 1995 [38] and later improved by Leo Breiman in 2001 [39].

According to Breiman (2001) [39], random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large [39].

This algorithm combines Breiman idea of bagging (Breiman, 1996) [40], where to grow each tree a random selection is made from the tuples in the training set, and the selection of a training set from a random set of weights in the training set (Ho, 1998) [41] and the random selection of attributes for the best split at each node (Amit and German, 1997) [42], constructing a collection of decision trees with controlled variation [43].

3.3. Bayesian Classification

Bayesian classifiers are statistical classifiers based on Baye's theorem that predict the probability of a tuple to belong to a certain class. Similarly to decision trees and selected neural network classifiers, when applied to large databases, Bayesian classifiers (as the Naïve Bayesian and Bayesian Networks) show high accuracy and speed [34].

Baye's theorem gives the *a posteriori* probability of an event H conditioned by X, P(H|X). It requires the *a priori* probability of H, P(H), the posterior probability of X conditioned on H, P(X|H) and the prior probability of X, P(X).

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{10}$$

Naïve Bayesian classifiers assume class conditional independence, meaning that the effect of an attribute value on a given class is independent of the values of the other attributes.

3.3.1. Naïve Bayesian Classification

For a data set where each tuple is an n-dimensional vector, *X = (x$_1$, x$_2$, . . . , x$_n$)*, representing, respectively, n attributes, *A$_1$, A$_2$, . . . , A$_n$*, having C as the m-dimensional class vector, C= C$_1$, C$_2$, . . . , C$_m$, the Naïve Bayesian classifier works as follows:

(1) For any tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X.

If, and only if

$$P(C_i|X) > P(C_j|X) \quad \text{for } 1 \leq j \leq m, j \neq i \tag{11}$$

naïve Bayesian classifier will predict that X belongs to the class C_i

Thus, P(C$_i$|X) is maximized and by Bayes' theorem,

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{12}$$

(2) Only $P(X|C_i)P(C_i)$ need to be maximized because P(X) is constant for all classes, and if the class prior probabilities are unknown, the classes are assumed to be equally likely, therefore, $P(X|C_i)$ would be maximized.

(3) It's made an assumption of **class conditional independence**, reducing computation in $P(X|C_i)$ .Therefore, the values of the attributes are presumed to be conditionally independent of one another, given the class label of the tuple.

(4) For class label prediction, $P(X|C_i)P(C_i)$ is evaluated for each class C_i. In order that the class label of tuple X is predicted as class Ci for which $P(X|C_i)P(C_i)$ maximum.

$$P(X|C_i)P(C_i) > P\left(X\middle|C_j\right)P\left(C_j\right) \quad \text{for } 1 \leq j \ \leq m, j \neq i \tag{13}$$

3.3.2. Bayesian Network

The Naïve Bayes classifier produces a probability estimate, rather than hard classifications. For each class value, it estimates the probability of a given tuple to belong to that class [33]. Futhermore, for a given class of a tuple, it assumes that the attributes are conditionally independent of each other, simplifying the computing [34].

Developed by Pearl (1995) [43], bayesian networks (BN), also known as Bayes nets, are a statistical based alternative belonging to the family of probabilistic graphical models that represent a set of random variables in the nodes and their conditioned dependencies in the edges between the nodes, combining principles from graphical theory, probability theory, computer sciences and statistics [44]. BN specify joint conditional probability distributions that allow class conditional independencies to be defined between subsets of variables [34].

As the Naïve Bayesian algorithm, BN also use Bayesian statistical methods, offering an efficient and principled approach for avoiding the overfitting of data [43].

3.4. Associative classification

Associative classification mining is an approach that makes use of association rules discovery techniques to construct classification systems [45].

Associative algorithms search for frequent patterns and their corresponding association or correlation rules that characterize interesting relationships between attribute conditions and class labels [34].

In a first step, the algorithm seeks frequent itemsets, searching for patterns of attribute-value pairs that occur repeatedly in the data set, where each attribute-value pair is considered an *item*, forming a group of *frequent itemsets*. In a second step, the *frequent itemsets* are analyzed in order to generate association rules [34].

3.4.1. RIPPER (JRip)

RIPPER (Repeated Incremental Pruning to Produce Error Reduction) was first purposed by Cohen (1995) [46] as an optimized version of IREP (Incremental Reduced Error Pruning) of Fürnkranz and Widmer (1994) [47].

In order to obtain decision rules from a dataset, RIPPER splits the training data, grows a single rule using one subset of the data using an heuristic method and then prunes the rule using one subset of the data. After producing a rule set for the class, each rule is reconsidered using reduced-error pruning, ensuring that each rule accurate, before proceeding to generate rules for the next class [33, 47, 48]. Weka uses a java implementation of the rule learner RIPPER, known as JRip.

3.5. Cluster analysis

Clustering is the process of grouping the data into classes or clusters according to their similarities, so that objects within a cluster are very similarities between them but very dissimilar to objects in other clusters [34, 35].

Cluster analysis can be used as a data mining function tool to understand the characteristics of each cluster, to serve as a preprocessing step for other algorithms (characterization, attribute selection) and to focus on a particular set of clusters for further analysis [34, 35].

3.5.1. Partitioning Methods

For a given data set of n objects, D, and being k the number of clusters to form, a partitioning algorithm organizes the objects into k partitions (clusters). A partition criterion is used to form the clusters, being *k-means* the most well-known method [33, 34].

3.5.1.1. k-Means Method

The k-means algorithm partitions a set of n objects into k clusters so that the resulting intracluster similarity is high (objects in the same cluster) but the intercluster similarity is low (objects in the different clusters). The cluster's is the mean value of the objects in a cluster, featuring each cluster.

The k-means algorithm proceeds as follows:

```
(1) randomly selects k objects from D as the initial clusters centers;
(2) repeat
(3)     (re)assign each object to the cluster to which it is most similar, based on the
        distance between the object and the cluster mean;
(4)     Updates the cluster means for each cluster;
(5) Until no change (criterion function converges).
```

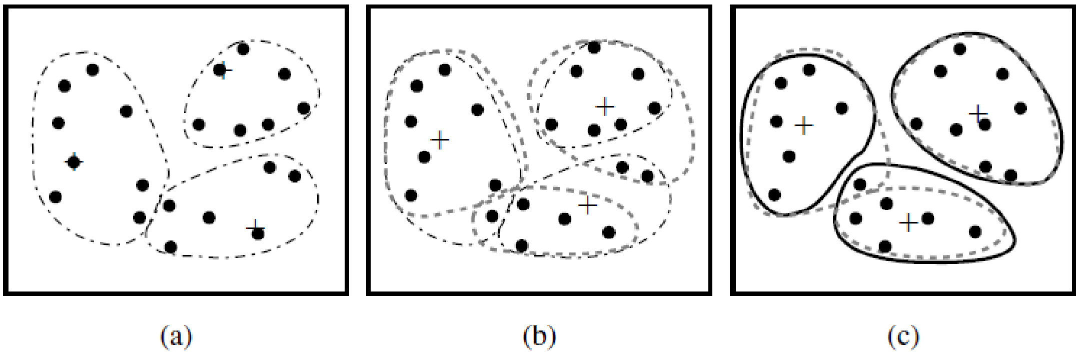


Figure 13: Clustering of objects based on k-means method, where the "+" represent the mean of each cluster [34];

The square-error criterion is used to calculate the square error for all objects in the dataset, and is defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{14}$$

where p is the point representing an object and m_i is the mean of cluster C_i. The algorithm attempts to determine k partitions that minimize the square-error function [34].

4. Hardware

In this chapter, the developed hardware systems will be described. The hardware development was the first stage of this project. The prototype consists of two modules, one for APW acquisition and other for ECG recording. To improve operator independence for APW recording, a multi probe was developed, consisting of four piezoelectric (PZ) transducers.

It is expected from this non-invasive device to be able to improve diagnosis quality by allying a powerful and historical technique such as ECG to a technique like APW which is recent and yet to fulfill all its potential.

4.1. Acquisition System

The developed acquisition system incorporates two main blocks, one for APW reproduction and another for ECG recording. First, the APW module will be addressed by explaining the PZ transducer multiprobe purpose and development, and later the signal conditioning system will be approached. Then, the ECG module development will be discussed.

The following figure represents an overview of the entire signal acquisition methodology:

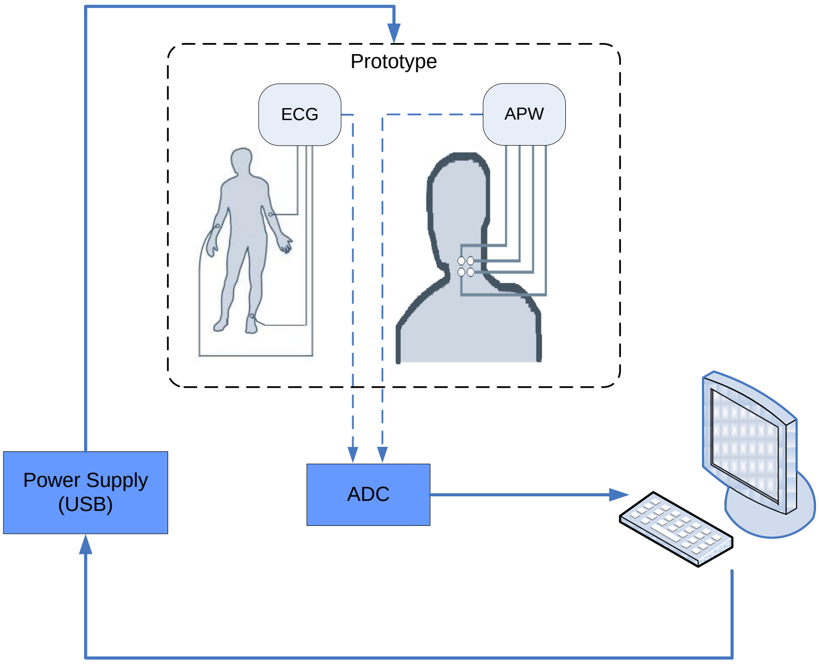


Figure 14: Schematic overview of the acquisition system;

4.1.1. APW module

The APW module allows waveform recording and storage in a digital format, enabling future data processing. This module consists of a multi piezoelectric transducers probe and a signal conditioning circuit. The probe is placed over the common carotid, in contact with the patient, while the conditioning circuit is responsible by the filtering and amplifying tasks.

4.1.1.1. Multi probe

In recent years, piezoelectric based probes have been widely used in APW signal measurements due to their characteristics. PZ transducers present high sensitivity, high signal-to-noise ratio (SNR) and are available at a low price, which made these sensors desirable to be used in this prototype.

The sensor design consists of four piezoelectric transducers bonded to a plastic block that supports bending under normal use. In order to provide the best conditions for APW signal recording, a mushroom interface was assembled over each piezoelectric element (figure 15).

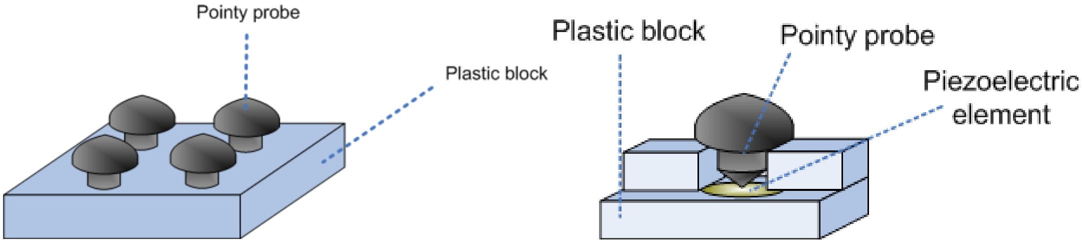


Figure 15: Housing characteristics of the constructed probe;

As suggested by Almeida *et al.* (2011) [49] where a single probe was tested, a mushroom-shaped PVC (polyvinyl chloride) pointy interface was used. The pointy probe's base, which contacts directly with the sensor, exhibits the best performance in reproducing the waveform with low RMSE variance [49].

In hemodynamic studies, the signal's bandwidth is bellow the resonance peak frequency, hence, Murata's piezoelectric transducer 7BB-27-4 was used, which has a resonance frequency of 4.6 ± 0.5 kHz, a maximum resonance impedance of 200 Ω and a capacitance of 20.0 ± 30% [1 kHz] nF [50].

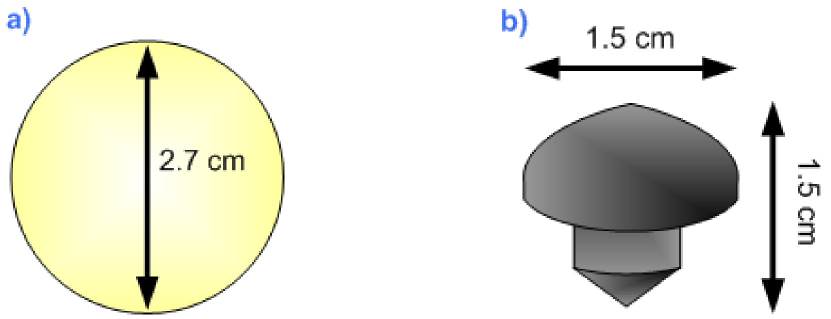


Figure 16: a) Used piezoelectric disc sensor dimensions, b) mushroom-shaped PVC interface dimensions;

The probe was designed to be placed over one's common carotid, eliminating the need of an experienced operator and leading to greater independence. Placing the probe over the common carotid artery, it is expected that at least one of the PZ sensors is capable of APW measurement by the transmission of mechanical energy to the PZ sensor, caused by tissue surface displacement due to carotid's blood flow.

Figure 17 shows an example of how the probe should be used: PZ 1 is placed over the common carotid artery, providing a good APW signal; PZ 2 and PZ 4 are placed in the patient's neck, not reaching the carotid artery, hence, the obtained signal will be mainly noise; although a part of PZ 3 is placed over the carotid, the generated signal by this sensor will be corrupted by noise. Therefore, the better signal will be the one obtained from PZ 1, with a great SNR, while the signal generated by the other sensors will be corrupted by noise, hence, a lower SNR.

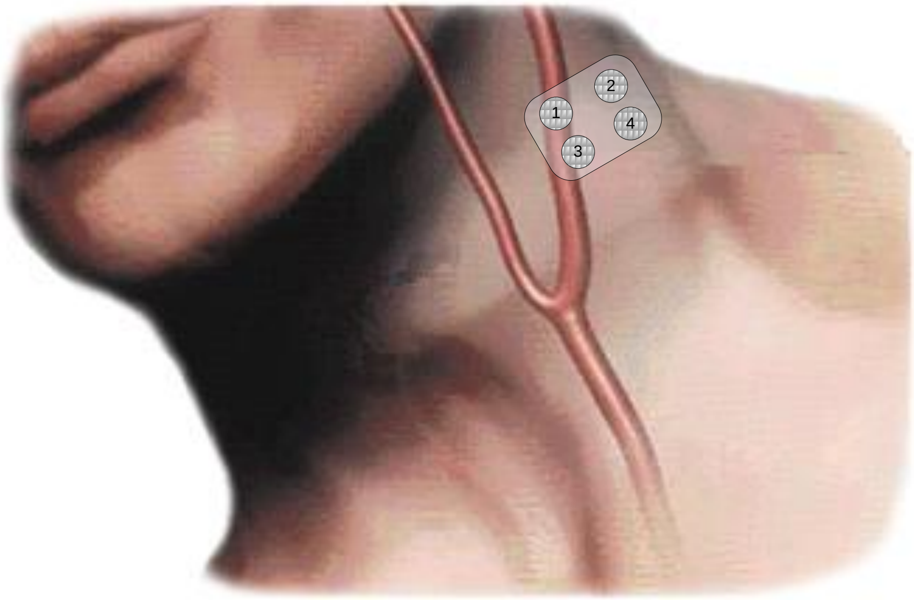


Figure 17: Example of the multiprobe placement over the common carotid artery (adapted from [51]);

4.1.1.2. Signal conditioning

Figure 18 shows the architecture of the APW signal conditioning. This module is divided in three parts: the power supply module, an amplifying stage and a processing module. The power is supplied via an USB cable connected to a personal computersupplying the necessary voltage to the other two modules. In the first amplifying stage, the raw signal obtained from the probe is amplified using an active differentiator mode (ADM) amplifier, purposed by Almeida et al. (2011) [49] with a gain of 1000. After this first stage, the best signal needs to be selected and it's chosen to be processed in the third stage. However, this step hasn't been completed and the reasons will be exposed ahead. In the processing module a peak detector with a timer is used to extract the reference time signal associated with the signal peak [49]. Since the signal obtained with the sensor is a time derivative of the physiological signal (due to of the sensor) [49], the signal needs to be integrated. This would be made by a microcontroller module, a previous task developed by Almeida, *et al.* (2011) [52].

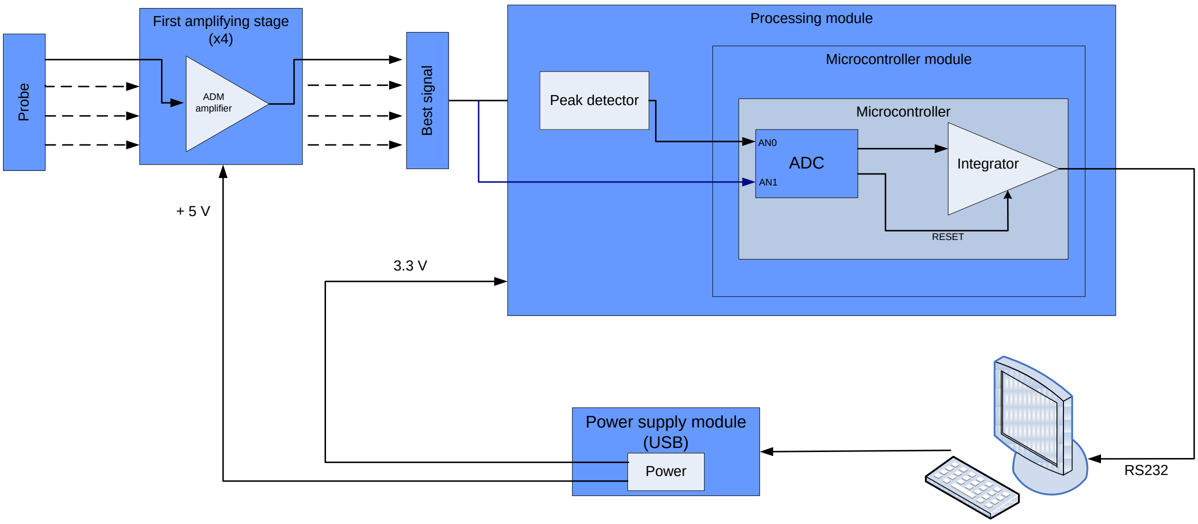


Figure 18: Signal conditioning architecture of the APW module;

This stage of this work hasn't been completed due to the difficulties encountered in finding an algorithm capable of distinguish the signal with better quality among the four signals generated from the PZ probe.

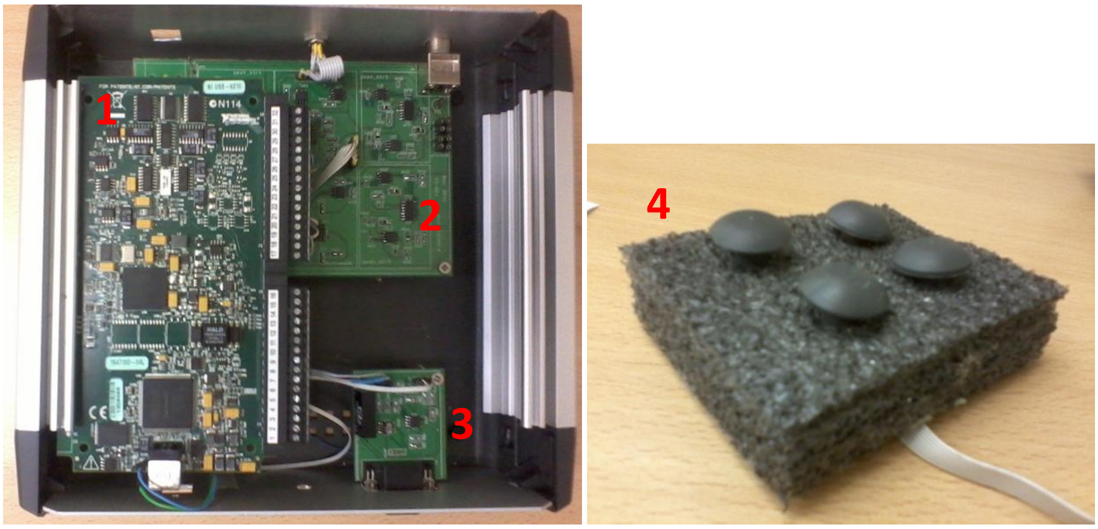


Figure 19: Images of the developed prototype for non-invasive cardiovascular studies. (1) ADC; (2) APW signal conditioning module; (3) ECG signal conditioning module; (4) multipiezoelectric probe;

Figure 20 shows an example of APWs before integration that were acquired with the developed prototype where each piezoelectric generated a signal. Although it's not perceptible in the figure, the signal is corrupted with baseline noise. Furthermore by analyzing the figure, an experienced user can easily verify that only one signal is optimal (piezoelectric 4), while the other signals are usually corrupted by noise. In this case, it can be concluded that piezoelectric 4 was the better placed sensor over the carotid.

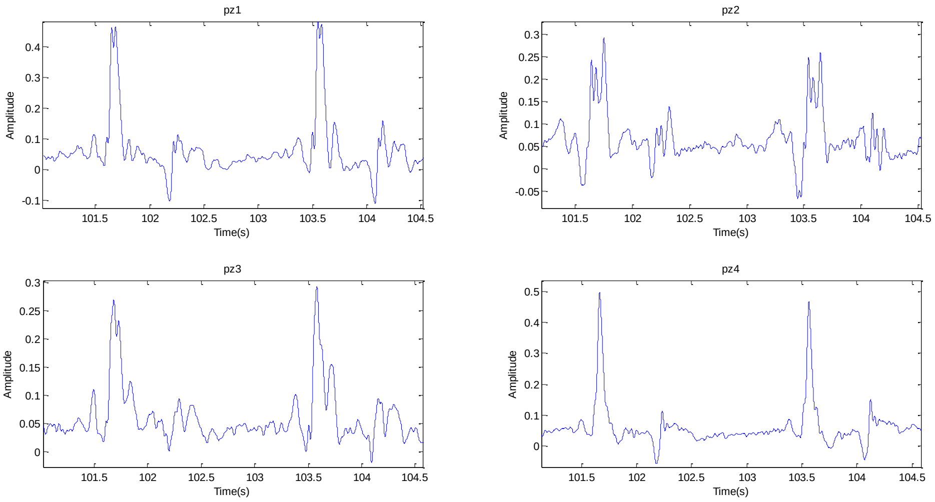


Figure 20: APW signal before integration acquired with the developed prototype;

Despite there is usually only one optimal signal, the methods tested for best signal identification weren't very effective. Methods using a threshold for peak detection allowing a temporal analysis failed because noisy signals present wave periods very similar to the best signal. Also, a method consisting on correlating each signal to itself was unsuccessfully tried.

Frequency analysis via Fourier transform for noise estimation also didn't prove to be very effective.

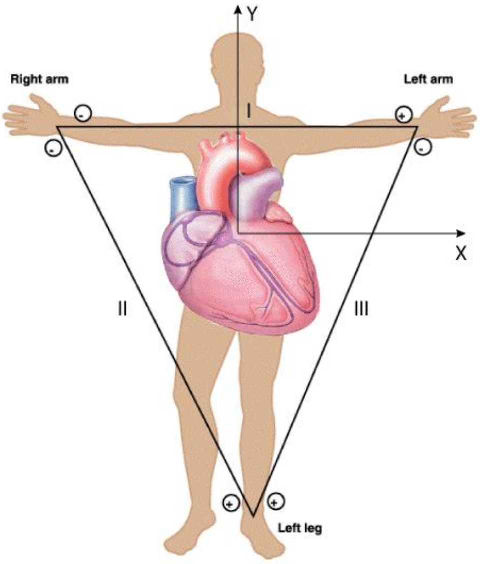
This way, the prototype wasn't successfully finished due to the difficulties found in the selection of the better signal.

4.1.2. ECG module

Just as the module presented before, the ECG module allows ECG recording and storage in a digital format, enabling future data processing. In order to access ECG information, a module consisting of an ECG signal conditioning circuit was built.

4.1.2.1. Leads

In this ECG configuration, three electrodes are placed in one's body, one on each arm and one on the left leg, forming the points known as Einthoven's triangle. The ECG record will be obtained from the voltage difference from any two sites. Therefore, as ECG is a differential record between two points, three different leads will be obtained. Lead I is the voltage between the left arm (positive) and the right arm electrode. Lead II is voltage between the left leg (positive) and the right arm electrode. Lead III is the voltage between the left leg (positive) and the left arm electrode [8].



$$I = V_{LA} - V_{RA}$$

$$II = V_{LL} - V_{RA}$$

$$III = V_{LL} - V_{LA}$$

Figure 21: Proper placement of the limb electrodes for a 3-lead ECG [53];

For ECG recording, the Biopac® disposable Ag-AgCl electrodes (EL503) were used in the tests. These pre-gelled electrodes have a circular contact area diameter of 1 cm, a backing diameter of 35 mm and are high chloride (7%), being suitable for short-term recordings [54]. These electrodes are responsible for converting ionic current flow of the body to the electron flow of the metallic wire.

Figure 22: Biopac® disposable Ag-AgCl electrodes used [54];

4.1.2.2. Signal conditioning

Several artifacts can corrupt raw ECG recorded signal and can have a physiological or non-physiological origin. The most common artifact is caused by power-line interference, appearing as a sinusoidal wave with a frequency of 50 Hz. Other artifacts can be caused by patient movement and respiration (movement artifacts, > 0.5 Hz), heart muscle contraction (electromyography interference, 20 to 1000 Hz), bad coupling of the electrodes to the patient (electrode contact noise) and baseline wander (typically between 0.15 and 0.3 Hz) [8].

In order to remove ECG noise's a bandpass filter with a frequency range of 0.03 Hz ~ 50 Hz was used. The filter is implemented by cascading a low-pass filter and a high-pass filter. In general, components of the signal of interest will reside in the 0.67 to 40-Hz bandwidth for standard ECGs [55].

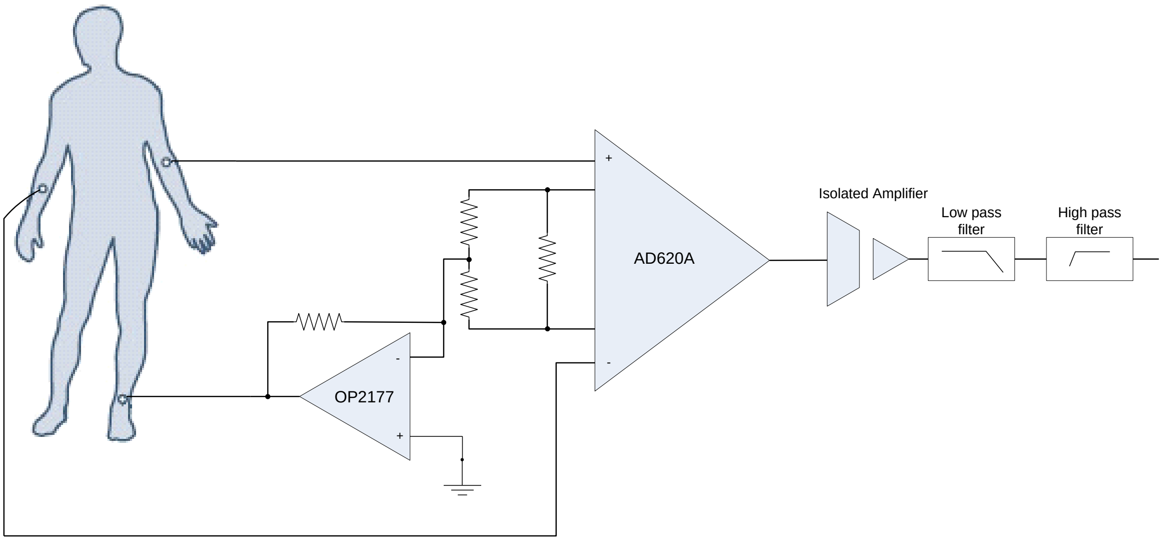


Figure 23: Scheme of the developed ECG platform, filtering the range of frequencies between 0.03 Hz and 50 Hz;

The figure bellow shows an example of the ECG signal acquired with the developed module.

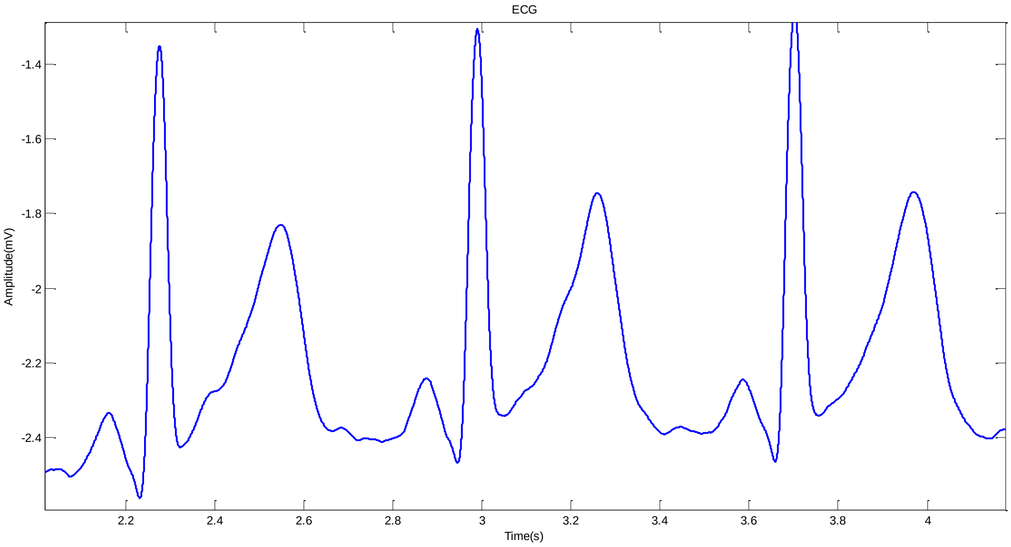


Figure 24: Example of a signal acquired with the developed ECG module;

5. Methodology

In this chapter, the algorithms for APW and ECG features detection will be addressed. The developed database will also be present as an essential platform for future data mining analysis, which will allow exporting stored data for Weka® software. In order to ease the interaction between the user and the database, an interface was created, making it simple to visualize data stored in the database.

The algorithms for APW and ECG analysis were developed using Matlab® 2009a. For the database construction, PostgreSQL, a free and open-source Object-Relational database management system (ORDBMS) was used. Yet, a protocol was created in Matlab®, allowing Matlab® connection with PostgreSQL in order to insert in the database the required information for future data mining analysis. As for the database interface, a Java desktop application was built using NetBeans IDE 7.0, a free and open-source integrated development environment. The software used for data mining tasks was Weka 3.6.3 (Waikato Environment for Knowledge Analysis), which is a machine learning software written in Java, free and available under GNU General Public Licence.

5.1. APW

In order to obtain the desired parameters from the APW signal, several steps need to be performed previously. In this subchapter, the stages for data processing will be discussed, including signal pre-processing, pulse segmentation and spatial feature extraction.

Since the developed prototype wasn't fully operational, the system developed in a previous work by Almeida, et al. (2011) [49] was used for APW recording and analysis.

5.1.1. APW onset calculation

According to Li *et al.* (2010) [56], the onset of an APW is related to a zero-crossing point before a maximal inflection of its derivative [56]. In this paper, the APW onset was determined using the same methodology purposed by Li [56].

First, a third order low pass Bessel filter was applied to the signal, with a cutoff frequency of 30 Hz to remove noises and artifacts that are common in raw signals. The value for the cutoff frequency was based in the frequency response of a typical APW signal. The frequency analysis through Fourier Transform (figure 25) shows that a frequency range until 30

Hz is sufficient for an adequate reconstruction of the blood pressure curve, and therefore, frequencies from 30 Hz onwards can be removed.

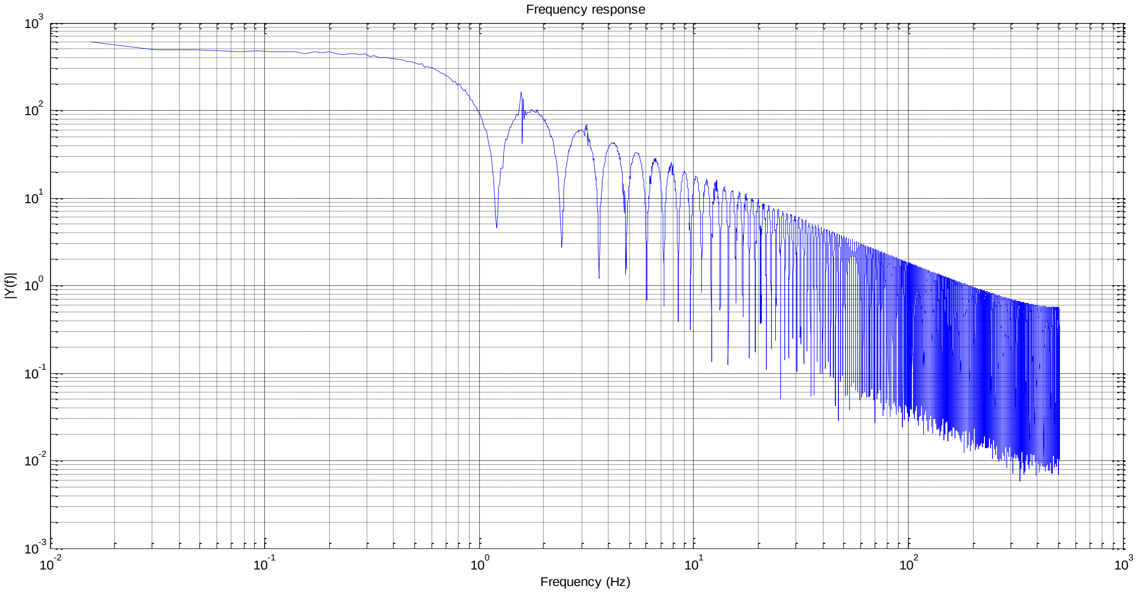


Figure 25: Frequency response of an ABP waveform signal showing interferences after 30 Hz;

Then, the differentials are calculated and the local extreme corresponding to the maximal inflection point of each pulse is determined, by applying a magnitude threshold. Having the maximal inflection points, the first zero crossing before each maximum is determined, matching the onset for each ABP waveform pulse. The figure bellow resumes the mechanism for the onset determination.

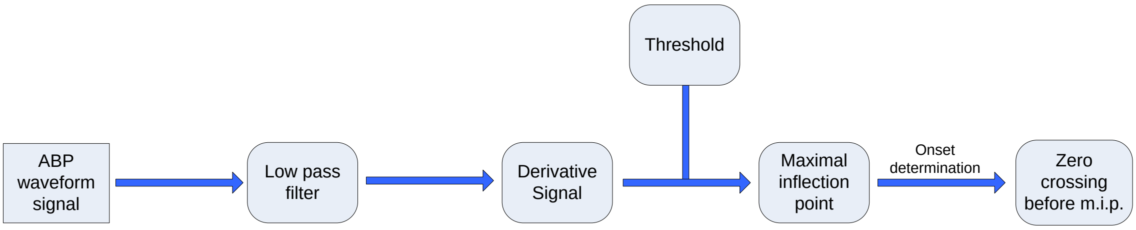


Figure 26: Mechanism for onset determination, where m.i.p. is the maximal inflection point;

5.1.2. Pre-processing

Any operation to retrieve information from APW, rely on a clean signal, without noise, artifacts or irregular waveforms, otherwise it can lead to strange results. Therefore, before applying any algorithm to extract clinical relevant information from an APW, it is necessary to remove anomalous beats from the signal.

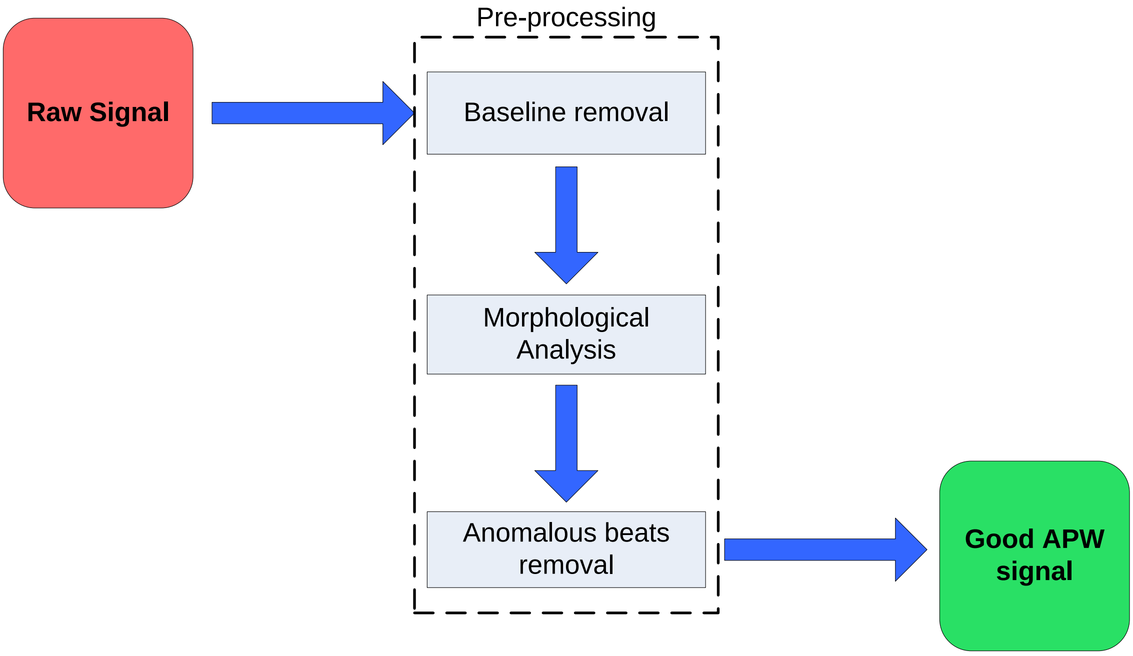


Figure 27: Diagram of the basic stages for APW "cleaning";

5.1.2.1. Baseline shift elimination

The first step in signal pre-processing consists on baseline noise removal from the raw signal. The baseline noise can be caused by electrical signal fluctuations, slow motion of the piezoelectric (PZ) probe attached to the patient neck, motion artifacts, or due patient breathing during signal acquisition procedure. Baseline modulation can lead to misleading and inaccurate determination of characteristic points.

In spite of PZ probe enables real time baseline elimination based on a reliable baseline restorer based (BLR), Almeida *et al.* (2011), an essential process that manages real time visualization [49], in this work, the baseline wander was removed from the raw signal by applying a baseline fit, ensuring that the APW is precise and valid.

First, the baseline index points are accurately determined. These points will match the APW pulses onset and offset. Then, the baseline fit is interpolated from the baseline index points, correcting the signal by vertically adjusting each sample between two successive baseline point indexes and matching the baseline point index to zero amplitude.

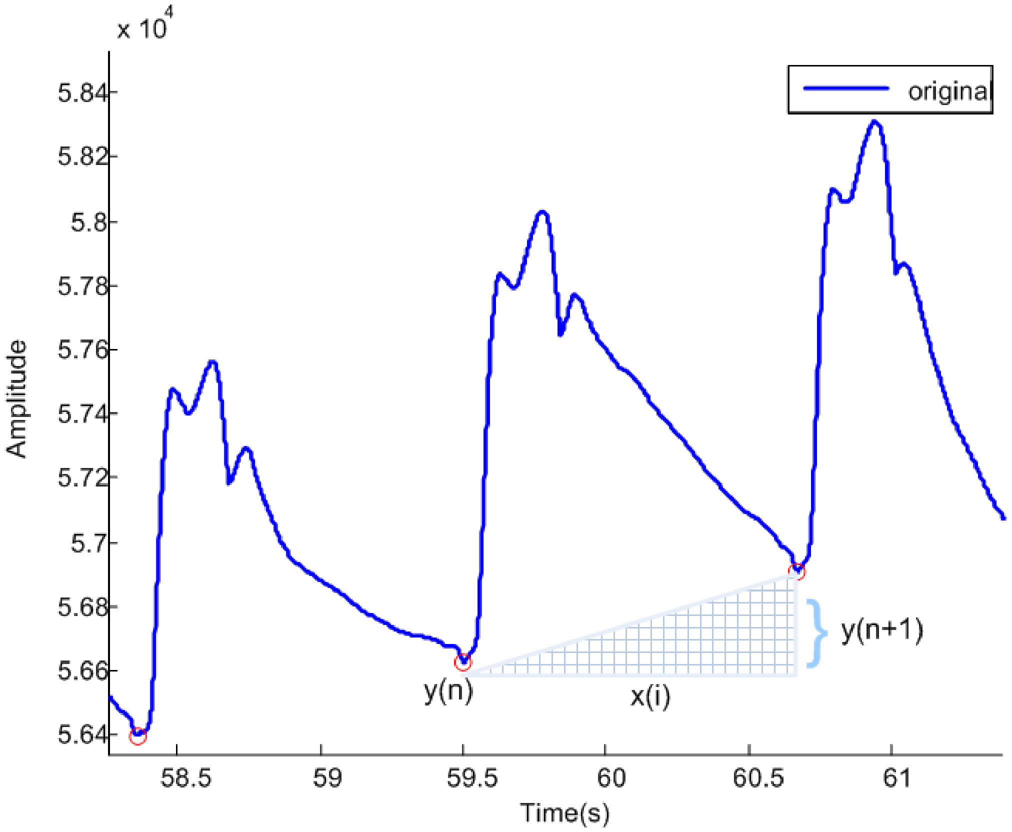


Figure 28: Method adopted for baseline correction where y(n) and y(n+1) are adjacent APW onset heights;

5.1.2.2. Morphological analysis

The morphological quality of APW pulses is another important aspect that must be taken into account to achieve an accurate and reliable analysis. During acquisitions motion artifact originated from voluntary or involuntary subject movement causes volume changes of the APW in a manner that is not associated with normal resting conditions.

After baseline removal, each APW beat is analyzed and several morphological features are evaluated by a series of constraints in order to check for abnormal beats. First, each APW pulse onset is marked, allowing the extraction of individual pulse features, which will be later compared with the abnormality criteria.

These abnormality criteria used are based in physiological ranges and the root mean square error (RMSE) between an average pulse and every pulse (individually). If any abnormality criterion is matched by an APW pulse feature, the pulse is flagged and removed.

The flagging criterions were based in trial-to-trial setups, where APW observation and inspection is made by a human expert engineer (HEE). The table below presents the criteria:

Table 6: APW features and abnormality criteria for pulse flagging;

Feature	Description	Abnormality Criteria
A$_S$	Systolic amplitude	$A_S > 3 \times A_m$
A$_m$	Mean systolic amplitude	$A_S < \left({}^{A_m}/_2\right)$
A$_D$	Diastolic amplitude	A$_D$ < 0
T **T$_m$**	Pulse width Mean width of all pulses	$T > 1.5 \times T_m$

Furthermore, a comparison between APW pulses is another criterion for pulse flagging. In this case, a mean APW pulse is calculated and the RMSE between the mean APW pulse and each pulse is computed. If it turns out that a particular pulse is very distinct from the mean APW pulse, the pulse is flagged.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{1,i} - x_{2,i})^2}{n}} \qquad (15)$$

Afterwards the morphological analysis and anomalous pulse flagging, the flagged pulses are removed, giving rise to a clean APW signal.

5.1.3. Pulse segmentation and normalization

After detecting the time intervals of anomalous beats, the APW signal is segmented into individual pulses and the abnormal pulses are erased. As the acquired signal is not calibrated, only the morphological characteristics of the APW are analyzed. The normalization will be used to scale heterogeneous beats, so that their morphology can be compared relevantly.

5.1.4. Spatial feature extraction

Several prominent points can be identified based in the identification of the most important time and amplitude coordinates in the APW. This information can be improved as ratios results.

The APW propagates along the arterial tree and, as its branches change in diameter and stiffness, reflections of a portion of the original energy is sent back towards the heart.

The waves reflected from the periphery, superimpose to the forward wave originating a visible change in the APW profile which determines the reflection point (RP). The dicrotic incisures is originated when the aortic valve closures occurring a moving back to the left ventricle of a small portion of the ejected blood.

The spatial feature extraction was based in the proposed method by Almeida *et al.* (2010) [36]. With this method several APW fiducial points were determined, enabling to classify the type of APW (see section 2.3.5.1) and to establish other parameters to be used in the future. All the used parameters are represented in the following table:

Table 7: Through the determination of several prominent points, several other parameters are obtained;

Category	Features	Description
Prominent Point	SP	Systolic peak
	RP	Reflection point
	DN	Dicrotic notch
	DP	Dicrotic peak
Time duration	T_{ba}	Ascent time of the forward pressure
	T_{cb}	Descent time of the forward pressure
	T_{dc}	Ascent time of dicrotic wave
	$T_{a'b}$	Descent time of waveform
Spatial parameters	$T_{a'b}/T_{ba}$	- Ratio of descent time to ascent time
	h_c/h_b	- Ratio of amplitude of DN to forward wave
	h_d/h_b	- Ratio of amplitude of DP to forward wave
	$\begin{cases} h_b - h_r & if\ t_r < t_b \\ h_r - h_b & if\ t_r > t_b \end{cases}$	- Difference between the amplitude of the forward wave and the reflected wave
	$\begin{cases} {h_r}/{h_b} & if\ t_r < t_b \\ -\left({h_r}/{h_b}\right) & if\ t_r > t_b \end{cases}$	- Ratio of the amplitude of the reflected wave to the systolic peak amplitude
	AI	- Augmentation index: Ratio that measures wave reflection and arterial stiffness (see table 5)

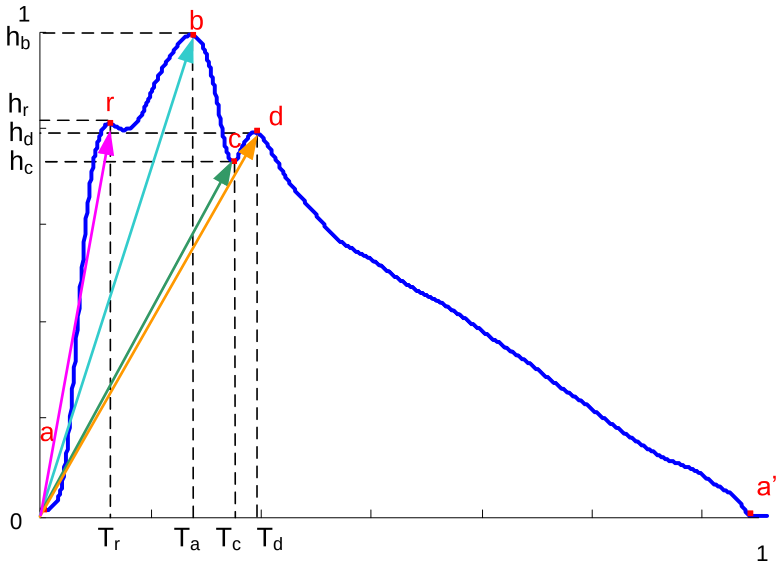


Figure 29: Fiducial points identification on an APW;

5.1.5. Heart rate

Heart rate can be determined as the number of beats per minute that occur in one's heart. The heart rate can be computed by:

$$HR = \frac{n_S}{\Delta T} \times 60 \tag{16}$$

where HR is the heart rate in beats per minute (b.p.m.), n_S is the number of systolic peaks detected in the APW signal and ΔT is the time interval in seconds. If the time interval corresponds to one pulse, then, only one systolic peak is detected, yielding

$$HR = \frac{1}{\Delta T} \times 60 \tag{17}$$

which results in a heart rate measurement based on each APW pulse.

Since it calculates the heart rate for each APW pulse, this method will allow heart period variability determination. Subsequently, an average heart rate is determined through the following expression

$$\overline{HR} = \frac{\sum_{i=1}^{n}\left(\frac{1}{\Delta T_i} \times 60\right)}{n} \tag{18}$$

where n is the number of pulses in the signal.

5.2. ECG

In APW analysis, ECG's P wave doesn't seem to be of great interest, therefore, the algorithm for ECG analysis was built taking into account only the QRS and T wave detection [12, 57]. Since P wave results from an atria depolarization, it doesn't represent any mechanical event from a physiological point of view, therefore, it doesn't influence the APW.

The developed algorithm is dependent of R and S wave detection. The first step of the algorithm is QRS detection. Firstly, positive and negative thresholds are applied in order to identify R and S wave (respectively), and consequently, identify Q wave as the first minimum before the R wave.

After S wave detection, a continuous wavelet transform (CWT) is applied to the ECG signal in order to detect the T wave. The first minimum and first maximum in CWT after S wave will identify T onset and T offset, respectively, while T maximum can be identified by CWT zero-crossing after S wave. Figure 30 shows a flow diagram of the algorithm.

Since the signals from the developed circuit for ECG acquisition didn't had the best quality, mostly due to baseline noise and sometimes a T wave with greater amplitude than R wave, MP36R from Biopac® acquisition box was used instead.

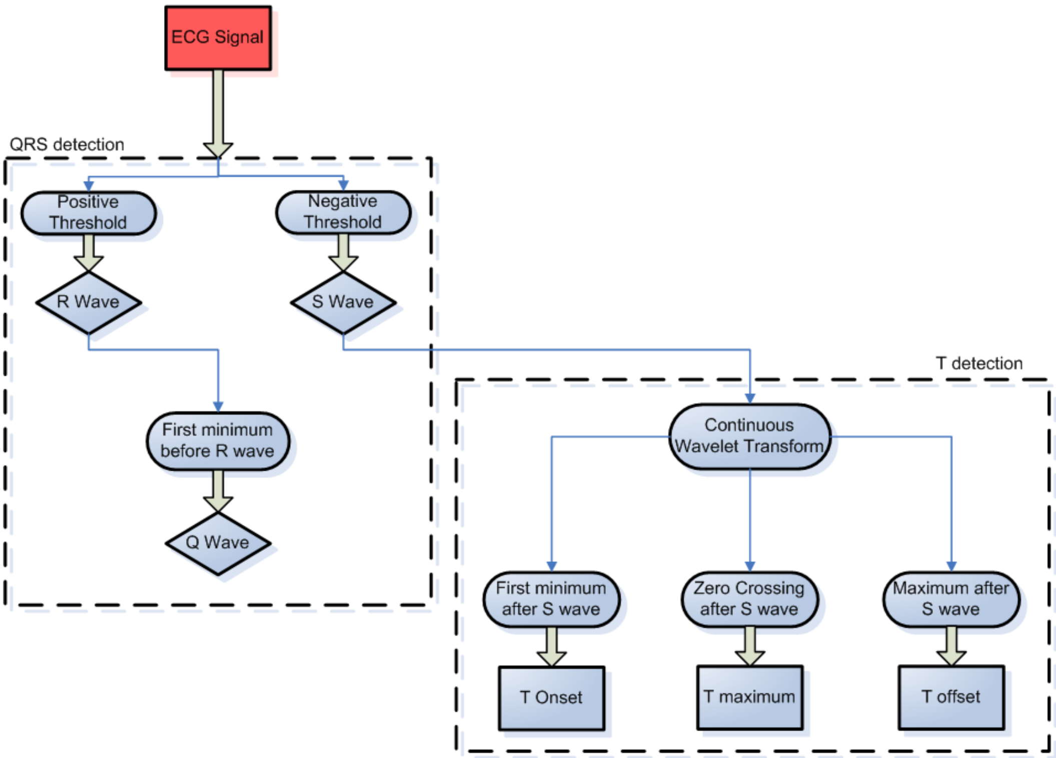


Figure 30: Algorithm for QRS complex and T wave detection from ECG signal;

In figure 31, the method for QRS complex and T wave identification is illustrated, showing the CWT of an ECG signal (top) and its corresponding ECG signal (bottom). The R and S waves are identified using a positive and negative threshold (respectively) in the ECG signal. The grey vertical bar marked with 1, shows the minimum in the CWT plot after the S wave, which will match T wave onset. The next grey bar shows the zero-crossing of ECG's CWT, which will match T maximum. The third bar marks the first maximum after S wave in the CWT plot and will match T offset in the ECG plot.

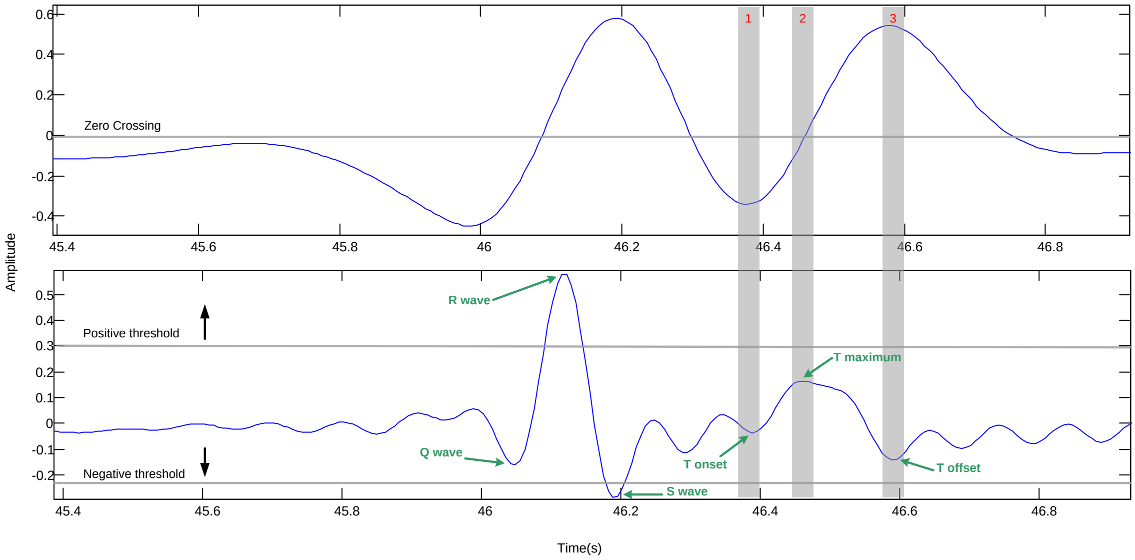


Figure 31: ECG signal (bottom) and its continuous wavelet transform (top), illustrating the method for QRS complex and T wave detection;

5.3. Database

A database system is a collection of programs that manages the database structure, allows the storage of data and controls the access to it. In healthcare, if clinical trials are to progress efficiently according to a plan, the use of patient databases is essential. The use of health information technology facilitates the organisation and management of clinical trials [58].

This part of the work is divided in two steps. The first step involves the creation of both entity-relationship and physical diagrams in order to obtain the SQL code for the database "skeleton" building. In the second step a graphical interface in Java was developed, facilitating data management and visualization.

5.3.1. Entity-relationship and physical diagrams

The purpose of this database is to store several important information related to people registered in this study. Besides allowing the storage of personal information, it allows to follow up the medical history of each patient, store information related with APW signal post-processing and can be useful in for future studies.

In this part of the work, the Sybase® Power Designer was used for designing the entity-relationship diagram and thus the physical diagram. Entity-relationship diagrams are a form of conceptual models used for illustrating relationships between entities in a database, being useful for describing data requirements and assumptions. A physical data model is a representation of a database design which takes into account the facilities and constraints of a given database management system, showing all table structures, including column name, column data type, column constraints, primary key, foreign key, and relationships between tables.

For obtaining the entity-relationship, an entity “person” was first created, allowing future storage of personal information of all people in the database. This entity was associated to other entities through a heritage, enabling to identify the type of person registered in the database (“patient”, “medical doctor”, “technician” and “administrator”).

In order to monitor patient’s clinical condition, the “medical history” entity was created and connected to “patient” entity. To allow data storage of each patient’s examination, the entity “Exam” was created, with several attributes from APW data post processing. The entity “Type of sensor” will allow having information about the type of sensor that was used in each exam.

The figures bellow show the final entity-relationship and physical diagram used for database creation.

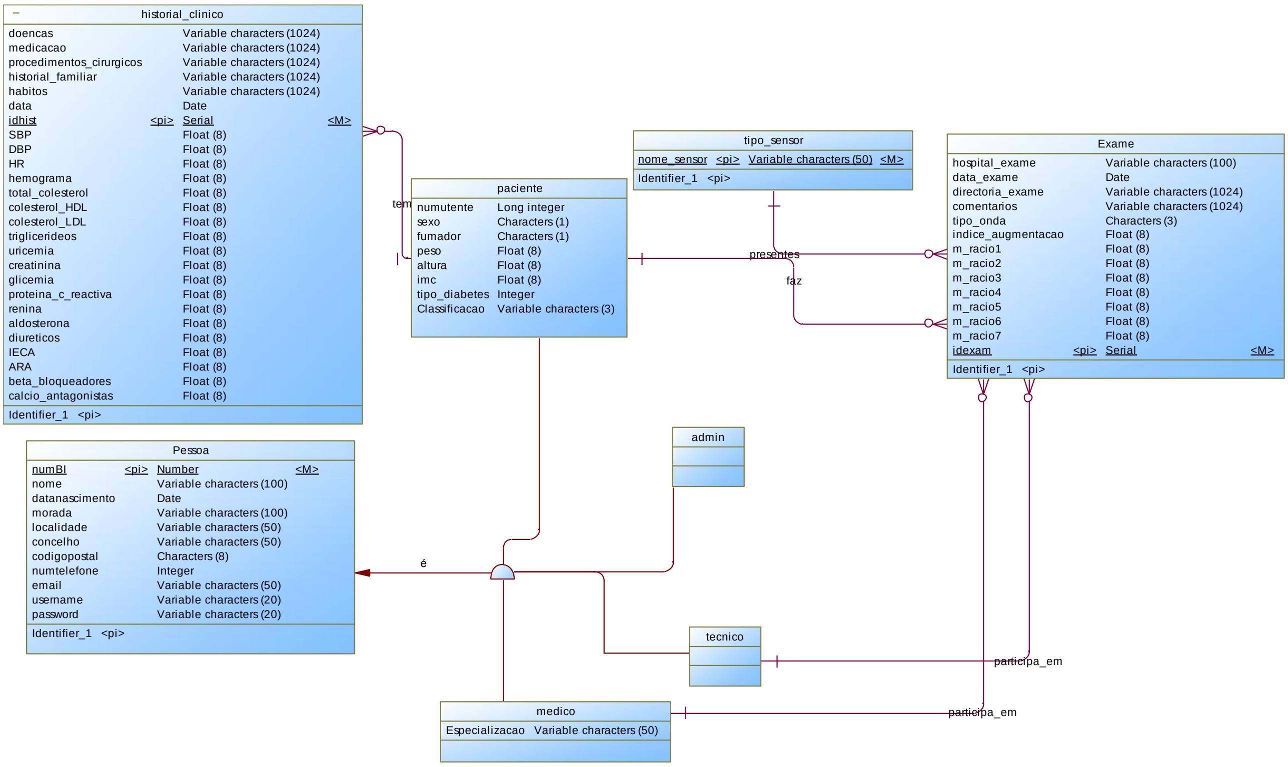


Figure 32: Extended entity-relationship metamodel of the constructed database;

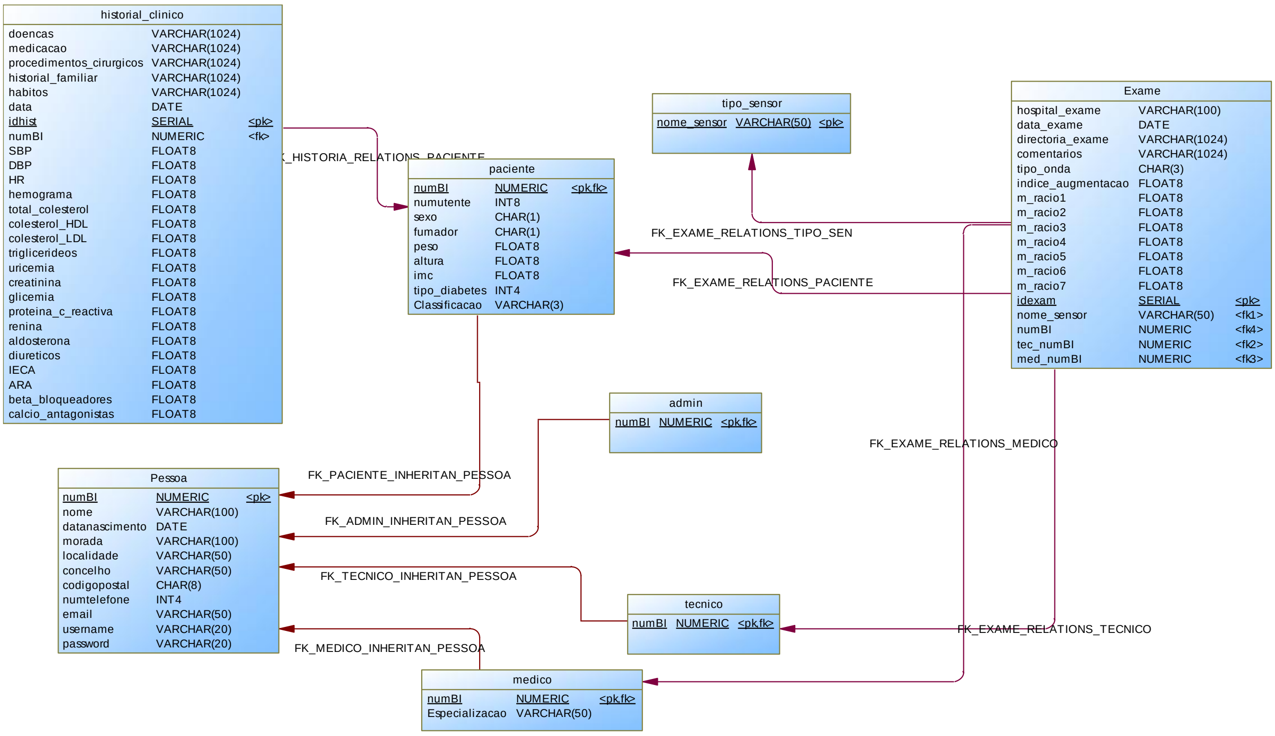


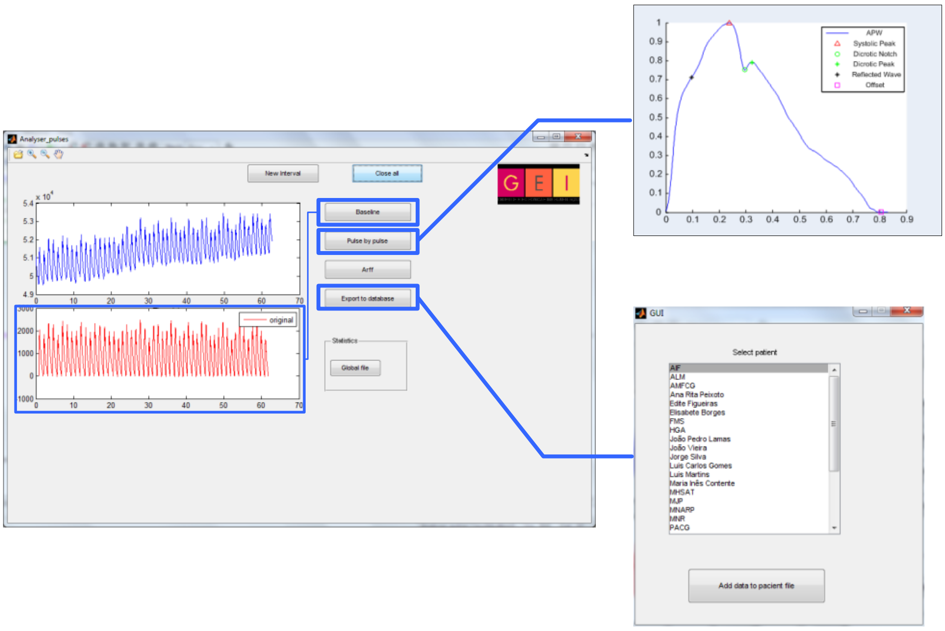
Figure 33: Physical diagram the constructed database;

5.3.2. Graphical user interface

In this step, two graphical user interfaces (GUI) will be presented. The first one was made using Matlab® and allows the user to retrieve important information from analyzed APW signals. Using the methods presented before, it allows removing the baseline from the signal, segment and analyze each APW pulse and to export several parameters to a patient's file in the database.

The second GUI was built in Java, using NetBeans, with the purpose of serving has a bridge between the database and the user. Using this friendly user interface, it's possible to insert new patients in the database, modify its personal information, add a medical history to their files, visualize the information that was exported from Matlab® among several other things.

In figure 34, the Matlab® GUI is presented showing some signal processing images and a window that allows information exporting to the database, by selecting the person whose signal was analyzed.

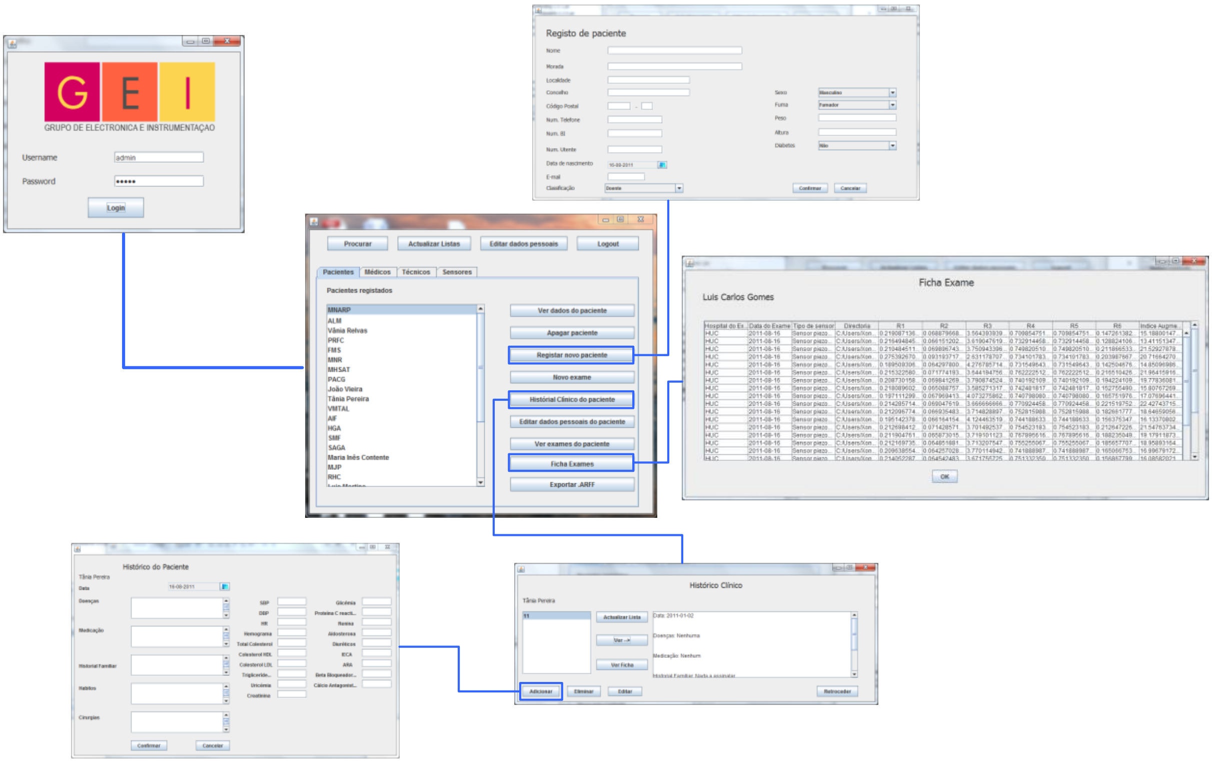


Select patient

AF
ALM
AMFCG
Ana Rita Peixoto
Edite Figueiras
Elisabete Borges
FMS
HGA
João Pedro Lamas
João Vieira
Jorge Silva
Luis Carlos Gomes
Luis Martins
Maria Inês Contente
MHSAT
MJP
MNARP
MNR
PACG

Figure 34: Matlab® GUI for APW signal analysis that allows exporting post-processing information to the database by selecting the person whose signal was analyzed;

Figure 35 shows some of the windows of the Java GUI. After user login, the main application window is presented. Here the user can choose between several available options. This figure presents the windows for patient registration, APW signal parameters visualization, and medical history visualization and insertion.



Username: admin
Password: •••••

Data de nascimento: 16-08-2011
Classificação: Doente
Sexo: Masculino
Fuma: Fumador
Diabetes: Não

MNARP
ALM
Vânia Relvas
PRFC
FMS
MNR
MHSAT
PACG
João Vieira
Tânia Pereira
VMTAL
AIF
HGA
SMF
SAGA
Maria Inês Contente
MJP
RHC
Luis Martins

Luis Carlos Gomes

Hospital do Ex...	Data do Exame	Tipo de sensor	Directoria	R1	R2	R3	R4	R5	R6	Indice Augme...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.219087136...	0.068879668...	3.564393939...	0.709854751...	0.709854751...	0.147261382...	15.18800147...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.216494845...	0.066151202...	3.619047619...	0.732914458...	0.732914458...	0.128824106...	13.41151347...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.210484511...	0.069896743...	3.750943396...	0.749820510...	0.749820510...	0.211866533...	21.52927878...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.275392670...	0.093193717...	2.631178707...	0.734101783...	0.734101783...	0.203987667...	20.71664270...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.189508306...	0.064297800...	4.276785714...	0.731549643...	0.731549643...	0.142504676...	14.85096986...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.215322580...	0.071774193...	3.644194756...	0.762222512...	0.762222512...	0.216510426...	21.96415916...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.208730158...	0.069841269...	3.790874524...	0.740192109...	0.740192109...	0.194224109...	19.77836081...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.218089602...	0.065088757...	3.585271317...	0.742481817...	0.742481817...	0.152755490...	15.80767269...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.197111299...	0.067969413...	4.073275862...	0.740798080...	0.740798080...	0.165751976...	17.07696441...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.214285714...	0.069047619...	3.666666666...	0.770924458...	0.770924458...	0.221519752...	22.42743715...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.212096774...	0.066935483...	3.714828897...	0.752815988...	0.752815988...	0.182661777...	18.64659056...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.195142378...	0.066164154...	4.124463519...	0.744188633...	0.744188633...	0.156375347...	16.13370802...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.212698412...	0.071428571...	3.701492537...	0.754523183...	0.754523183...	0.212647226...	21.54763734...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.211904761...	0.065873015...	3.719101123...	0.767895616...	0.767895616...	0.188235048...	19.17911873...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.212169735...	0.064851881...	3.713207547...	0.755255067...	0.755255067...	0.185657707...	18.95893164...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.209638554...	0.064257028...	3.770114942...	0.741888987...	0.741888987...	0.165066753...	16.99679172...
HUC	2011-08-16	Sensor piezo...	C:/Users/Xon...	0.214052287...	0.064542483...	3.671755725...	0.751332350...	0.751332350...	0.156857798...	16.08582021...

Tânia Pereira
Data: 16-08-2011

Tânia Pereira
11
Data: 2011-01-02
Doenças: Nenhuma
Medicação: Nenhum
Historial Familiar: Nada a assinalar

Figure 35: Main windows of the Java GUI built that allows to visualize, insert and modify information in the database;

It may be referred that this database and Java GUI were already made with some parameters regarding future experiments. Some parameters from patient’s blood analysis may be identified from the images, although they’re not used in this particular study.

5.4. Data mining

The goal of the data mining process is to create a model for automatic classification and detection of cardiovascular pathologies, to identify and understand the profile of a characteristic APW in the presence of arterial stiffness.

From the analysis of each wave in the APW signal, a dataset was created as an ARFF (Attribute-Relation File Format) file, which is an ASCII text file that describes a list of instances sharing a set of attributes [60]. Subsequently, the data mining framework Weka is used for classification and clustering analysis purposes. In figure 36, a flowchart describes the main steps in the data mining process.

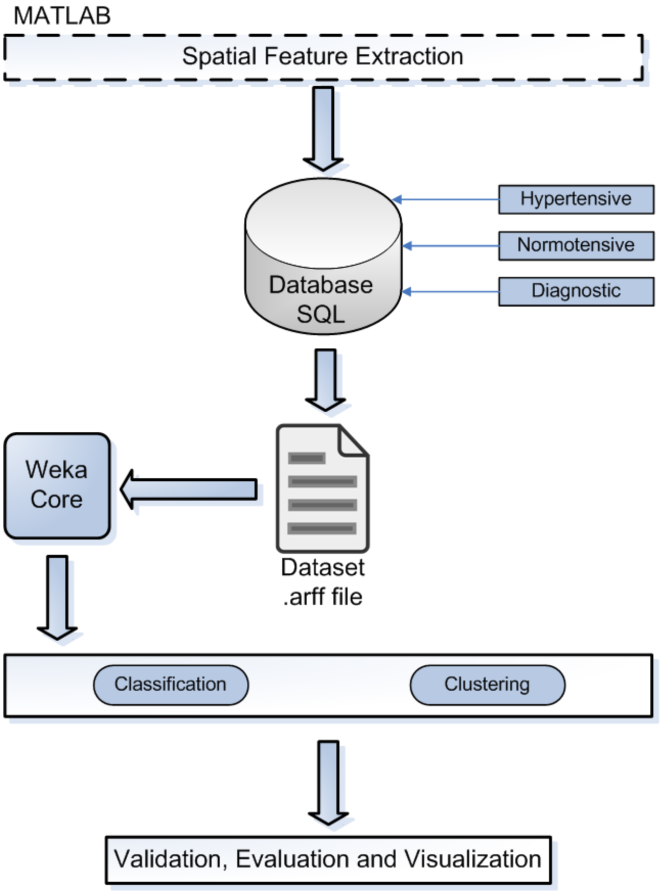


Figure 36: Flow diagram of the data mining process;

5.4.1. Dataset construction

As seen before in chapter 2.2.2, arterial stiffness is greatly influenced by age and hypertension. It is well known that the increasing of arterial stiffness in large vessels is a natural process of aging. Furthermore, arterial stiffness is also accelerated by the presence of hypertension.

The first step in the data mining process is the dataset construction. In order to identify the profile of a characteristic APW in the presence of arterial stiffness, a dataset consisting of elderly hypertensive individuals (group 1) and young normotensive individuals (group 2) was firstly built. Later, a group of people for model testing and early prediction of cardiovascular diseases predisposition was added to a dataset (group 3). For this last group, a set of people with a family history of cardiovascular diseases and with suspicious forms of APW was selected.

In this dataset, each wave in an APW signal is individually analyzed in order to extract the spatial parameters explained in table 7 (section 5.1.4) and each of these parameters will correspond to an attribute. For each wave in the APW signal, a tuple (instance) will be added to the dataset. Besides the spatial parameters, the data*set* also includes the parameters presented in table 8. The class for each tuple is assign according to the group that each individual belongs.

The dataset is composed by a group of 11 hypertensive people and a group of 9 healthy volunteers without CV diseases. An additional group of 6 volunteers without clinical diagnose was studied.

The Ethics Committees of the Academic Hospital of the University of Coimbra (HUC) in Portugal, where the measurements were performed, approved the study. Informed consent was obtained from the subjects before the recordings were made.

Table 8: Description of the general clinical parameters and population used for dataset construction. Group 1 consists of an elderly and hypertensive population; Group 2 consists of a young and normotensive population; Group 3 is the diagnostic group for early prediction of arterial stiffness predisposition;

Parameter	Group 1 Hypertensive, n= 11	Group 2 Normotensive, n= 9	Group 3 Diagnostic, n= 6
Age (years)	55 ± 12.23	23.53 ± 1.77	23.6 ± 1.8
Gender (M/F)	6/5	4/5	1/5
Smoker (Y/N)	3/8	2/7	0/6
Weight (Kg)	68.45 ± 5.65	65.33 ± 10.65	59.39 ± 6.10
Height (cm)	1.66 ± 0.04	1.70 ± 0.06	1.68 ± 0.07
BMI	24.82 ± 2.44	22.57 ± 2.75	20.95 ± 1.56
Systolic Blood Pressure (mm Hg)	171.6 ± 11.9	108.5 ± 13.4	101.0 ± 13.4
Diastolic Blood Pressure (mm Hg)	103 ± 6.9	69.4 ± 7.4	73.04 ± 9.72
Heart Rate (bpm)	66.6 ± 4.5	73.6 ± 8.9	63.0 ± 10.4

5.4.2. Pre-process

Pre-process will allow data transformation. For clustering purposes only, a supervised filter for attribute discretization is applied to the data set. This filter will consider the correlation between attributes and class attribute and will discretize the numeric attributes.

Attribute selection is an important tool that improves the accuracy of the built model, in a process where the unneeded information is identified due its low predicative value in the dataset.

For attribute selection, the unsupervised attribute filter “RemoveUseless” is used to remove attributes that do not vary or that vary too much. Furthermore, for classification purposes, only spatial features and class attributes are used, so that the diagnosis is only dependent on the APW morphology.

5.4.3. Classification

Classification uses supervised learning algorithms in order to train the classifier to distinguish tuples from different classes, based in the tuple's attributes. This will allow to build a model, which can later be applied to new measurements where the class is not known, giving a classification prediction over the new data.

For the model building, several different classifiers were tested and the criteria for the classifier choose will be presented in the next chapter. The training dataset used was constituted by hypertensive and normotensive, while the group for diagnostic purpose was used to predict the class attribute by applying the model.

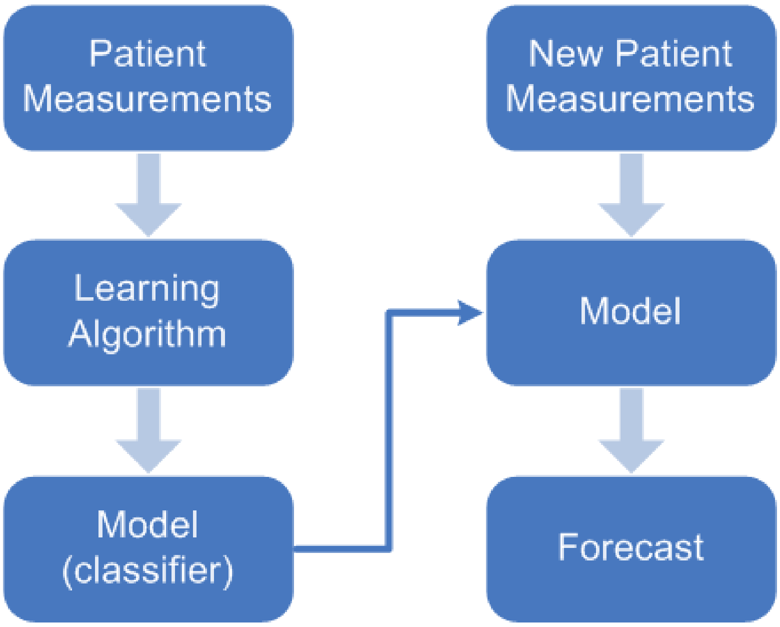


Figure 37: Flow diagram on how to use of classifiers for model construction and further diagnostic prediction;

5.4.4. Clustering

As an unsupervised learning method, clustering algorithms automatically group data according to their degree of similarity, not taking into account the class of each tuple.

The purpose of clustering is to determine the profile of an APW in the presence of arterial stiffness by analysing the trends of each APW spatial parameter in all dataset. It is expected that these parameters converge to a specific position depending on the health conditions contained in the dataset, allowing distinguishing different groups.

There are numerous available techniques to build multidimensional clusters. In this paper, the classic algorithm k-means was used for clustering purposes.

6. Results and discussion

In this chapter, some of the obtained results will be presented and discussed, particularly the algorithm for APW signal analysis, the ECG signal analysis, the obtained model for diagnosis prediction and clustering.

6.1. APW analysis

In order to obtain several parameters through the analysis of an APW signal analysis and to validate the developed algorithms, this section presents some of the most interesting results obtained. In the figure bellow, the major steps to obtain the spatial parameters presented before in the table 7 are illustrated.

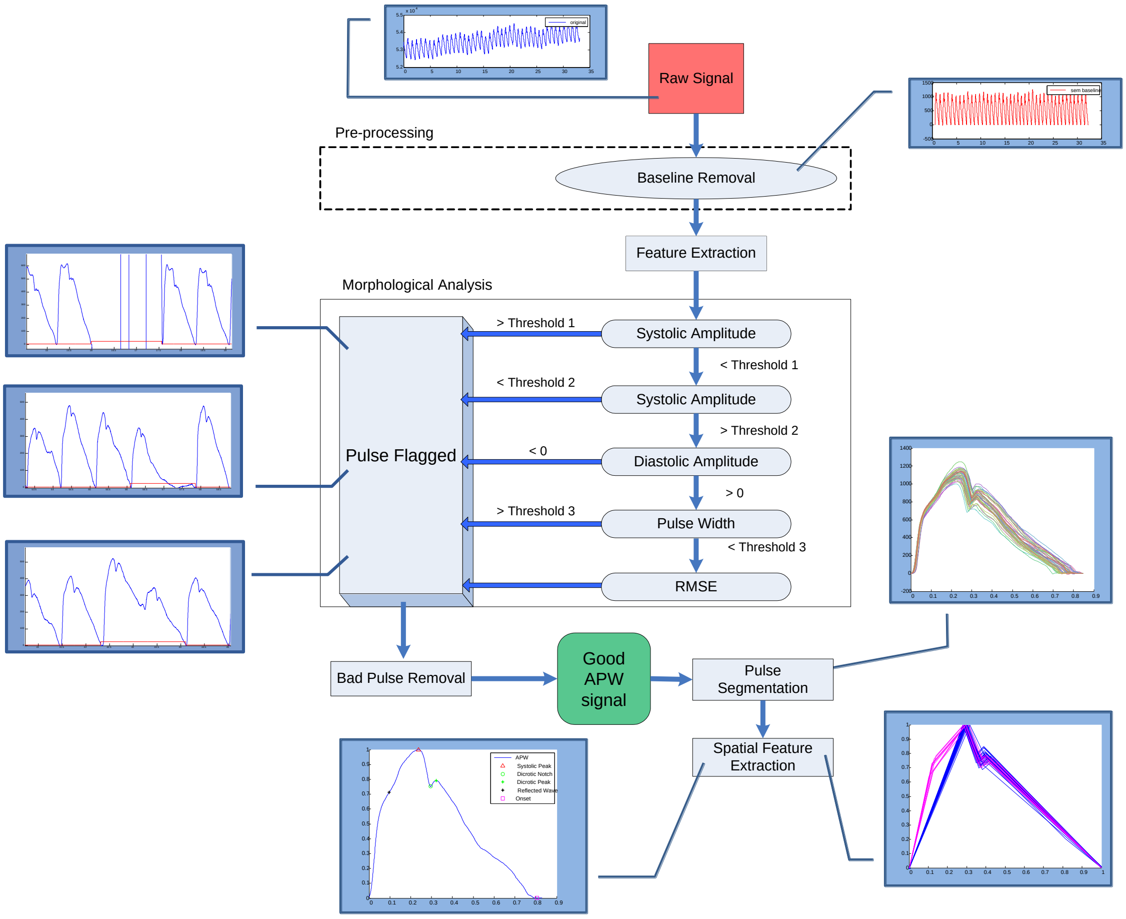


Figure 38: Flowchart of the global data processing stages for spatial feature extraction, where threshold 1 is given by $3 \times A_m$, threshold 2 is given by ${A_m}/_2$ and threshold 3 is given by $1.5 \times T_m$, where A$_m$ is the mean systolic amplitude and T$_m$ is the mean pulse width;

The developed APW signal processing algorithm was accurately tested using normal and abnormal data with several artifacts. The study protocol was approved by the ethical committee of the Hospital of University of Coimbra, Portugal. All subjects were volunteers and gave a written informed consent. The APWs were collected from healthy volunteers and unhealthy patients with hypertension.

6.1.1. Onset determination

In order to test the developed algorithm for onset determination, APW signal was acquired from a universe of 10 volunteers with ages between 22 and 55 years. The following figures show the APW onset marking on the acquired signal without previous pre-processing (figure 39).

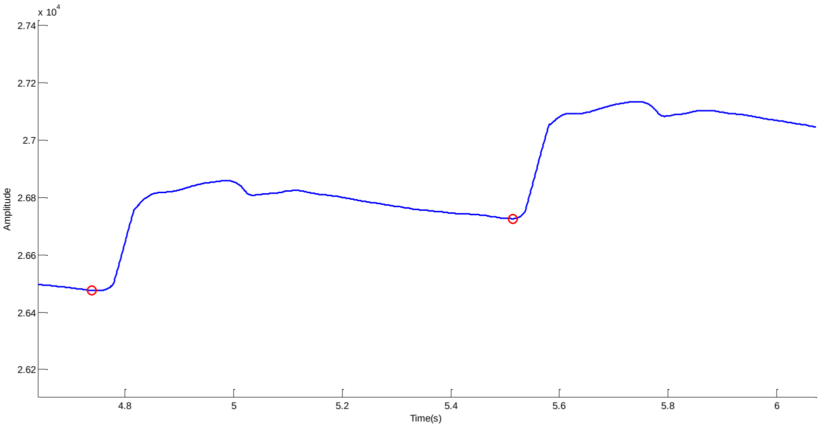


Figure 39: Example of APW signal and its identified onsets (red) for each pulse;

To evaluate the performance of the onset detection, the results obtained with the algorithm were visually inspected by a human expert engineer (HEE), being classified as true positive (TP) when the algorithm identifies correctly the onset, false negative (FN) when the algorithm doesn't identify the pulse onset and false positive (FP) when a pulse onset is identified although it shouldn't be. Sensitivity, positive predicted value, or precision rate (P$^+$) and error of the algorithm are presented in table 9, being computed by:

$$Sensitivity = \frac{TP}{TP + FN} \tag{19}$$

$$P^{+} = \frac{TP}{TP + FP} \tag{20}$$

$$error = \frac{FP + FN}{TP + FP} \tag{21}$$

Table 9: Evaluation of the developed algorithm performance for APW onset identification;

	N# of pulses	TP	FP	FN	Sensitivity (%)	P^+ (%)	Error (%)
Onset	542	518	15	9	98.3	97.3	4.5

6.1.2. Baseline correction

Baseline modulation correction is necessary for accurate extraction of fiducial points from the APW signal. The first step for the baseline correction is the APW onset identification by the algorithm presented before. Figure 40 shows an example of an APW raw signal with onset identification.

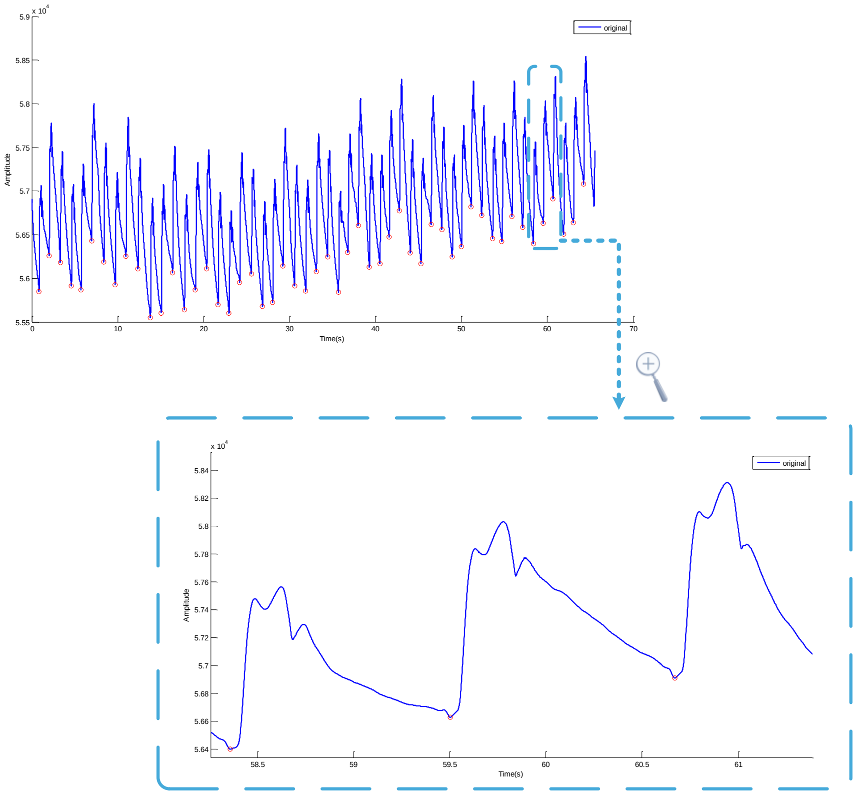


Figure 40: Example of an APW raw signal without baseline correction and a zoomed in segment, showing the onset identification;

By proportionally shifting the data points between two consecutive onsets in a vertical direction and getting the onsets in a horizontal line, the baseline is corrected. Figure 41 shows an example of an APW signal with baseline correction using the developed method.

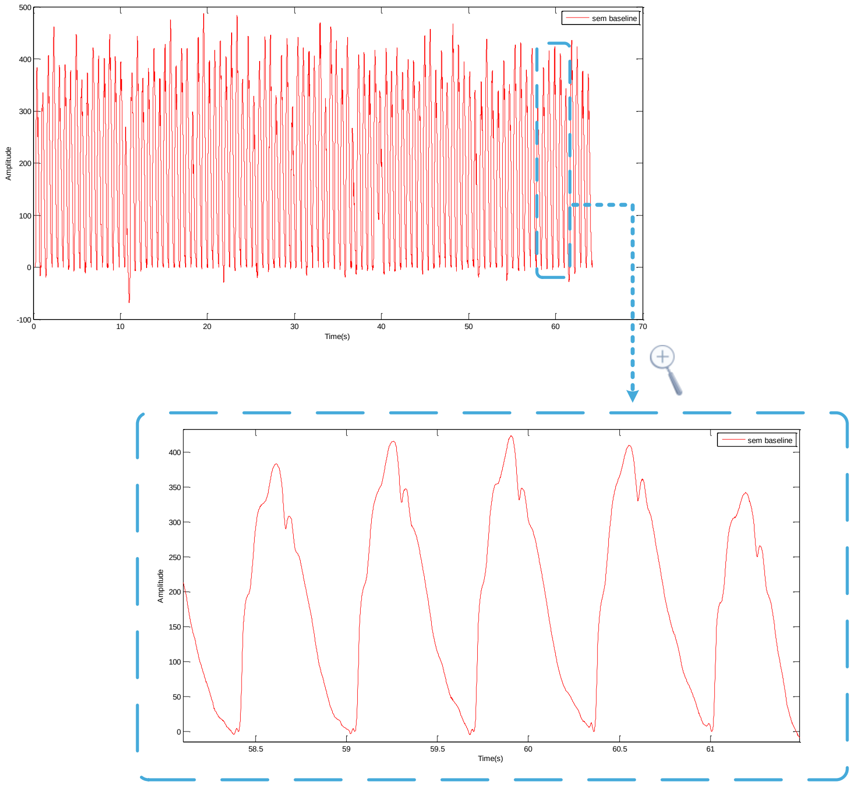


Figure 41: Example of an APW signal wit baseline correction and a zoomed in segment of the signal;

6.1.3. Abnormal pulse removal

To evaluate the performance of the abnormal pulse removal, the results obtained with the algorithm were visually inspected by a HEE, being classified as TP when the algorithm identifies correctly a normal pulse, FN when the algorithm identifies a pulse as abnormal however it is normal, FP when the pulse is classified as normal although it is abnormal and as true negative (TN) when the pulse is truly identified as abnormal. Sensitivity relates the algorithm ability to identify normal pulses, while specificity is related to the algorithm ability to identify abnormal pulses. Precision rate (P^+) gives a measure of the proportion of times normal pulses are correctly classified by the algorithm. In table 10 the performance of the developed algorithm is presented.

$$Specificity = \frac{TN}{TN + FP} \tag{22}$$

Table 10: Evaluation of the developed algorithm performance for APW good pulse identification;

	Sensitivity (%)	Specificity (%)	P^+ (%)	Error (%)
Good pulse identification	97.4	69.6	98.1	4.5

Abnormal pulse may arise from several reasons. The most common factors that affected the quality of APW recordings are related to patient's involuntary movement (such as breathing during signal acquisition), bad sensor coupling, signal saturation, not effective removal of the baseline or onset identification. In figure 42, some examples of anomalous pulse identification are shown.

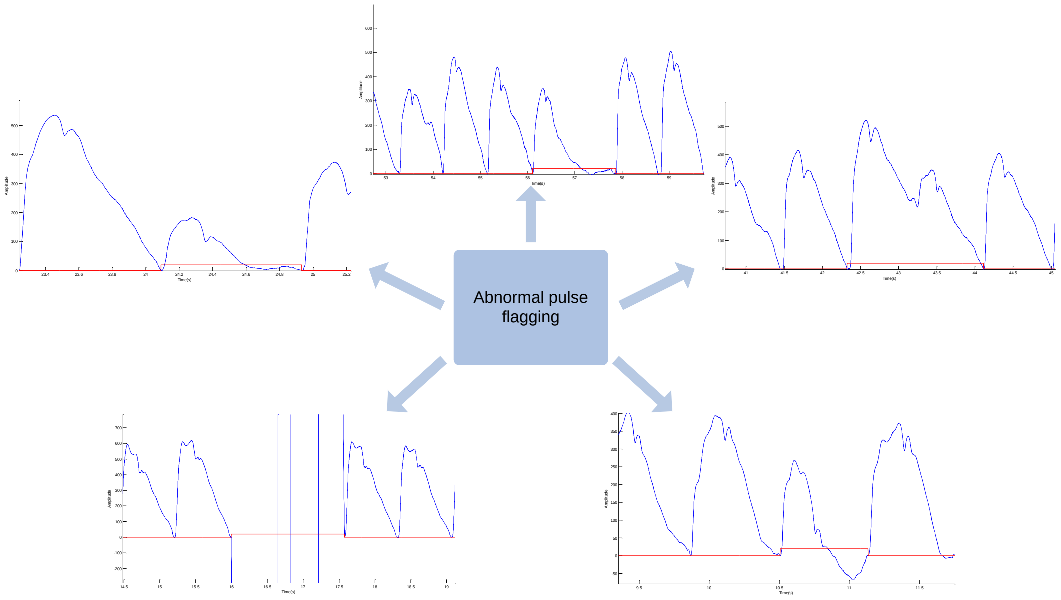


Figure 42: Examples of abnormal pulse flagging in APW signals with the developed algorithm;

6.1.4. Segmentation and normalization

In order to compare its morphological characteristics, APW pulses need to be segmented and afterwards normalized in both temporal and amplitude. Figure 43 shows APW pulses overlap before and after normalization, where each color corresponds to a different APW pulse.

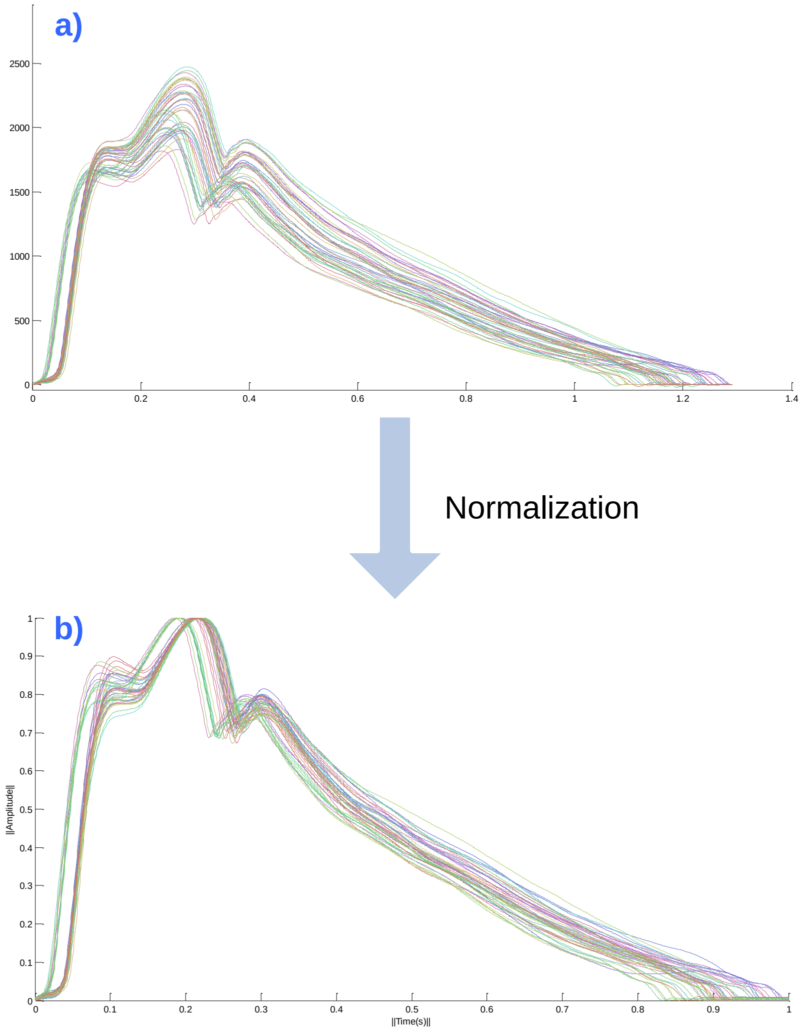


Figure 43: APW pulse overlap plot of segmented pulses before normalization (a) and after normalization (b);

6.1.5. Spatial feature extraction

For spatial feature extraction, a pulse by pulse analysis is performed using the algorithm proposed in a previous work [59]. Figure 44 shows a pulse detail where the prominent points (systolic peak (SP), reflected wave (RW), dicrotic notch (DN) and dicrotic peak (DP)) are identified.

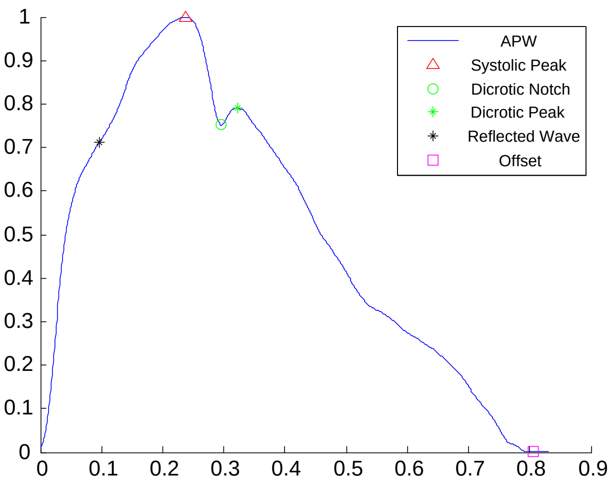


Figure 44: APW pulse before temporal normalization, showing the prominent points identified by the algorithm of [59];

In table 11, the time information that corresponds to these points is represented for all subject groups. SP and RP time coordinates splits the population in two groups of different time periods. Due to the subject groups in study, the dicrotic peak seem to present similar temporal values between the different classes of these populations and therefore will not be included in this study, once it doesn't present any relevant information, although it may be of interest to study in other pathologies. In other hand, the dicrotic notch time seem to vary between 200 and 400 ms.

Table 11: Temporal and amplitude analysis of APW characteristic points. Dicrotic peak doesn't seem to present variations between different types of populations;

		Time (ms)		Amplitude	
	Subject	Mean	STD	Mean	STD
SP	**Unhealthy**	221.540	35.831	0.995	0.001
	Healthy	149.779	32.359	0.991	0.006
	Diagnose	235.470	36.423	0.995	0.002
RP	**Unhealthy**	103.179	34.440	0.721	0.134
	Healthy	206.154	56.128	0.823	0.081
	Diagnose	133.815	32.988	0.798	0.094
DP	**Unhealthy**	329.969	46.661	0.785	0.054
	Healthy	338.732	59.049	0.712	0.129
	Diagnose	313.206	95.483	0.694	0.258

6.1.6. Heart rate

Figure 45 shows an example of heart rate variability representation and average beat rate (in red box).

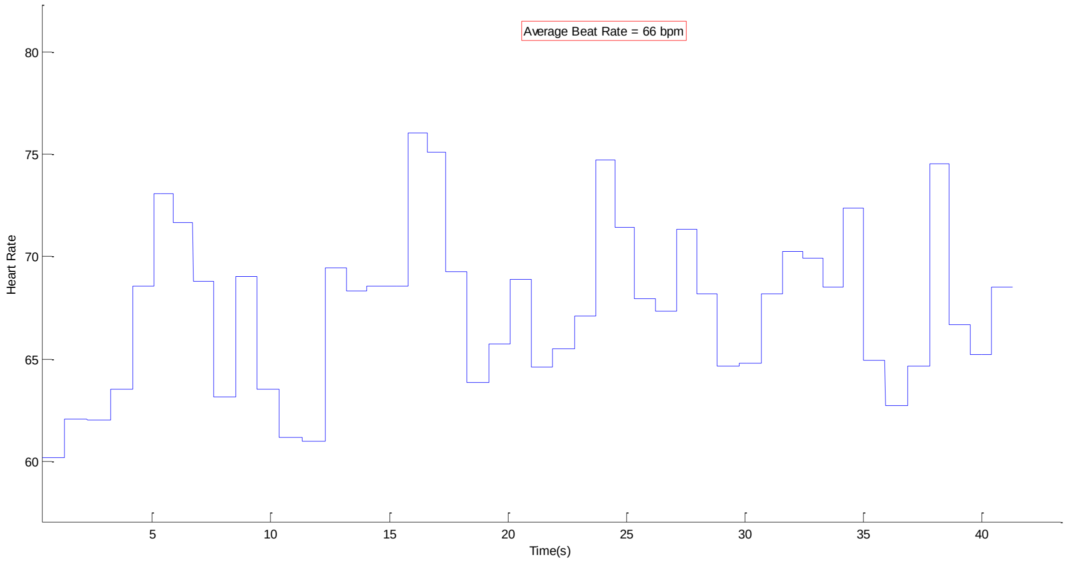


Figure 45: Example of heart rate variability plot and average beat rate (red box);

6.2. ECG analysis

As it has been before, since the signals from the developed circuit for ECG acquisition didn’t had the best quality, mostly due to baseline noise and sometimes a T wave with greater amplitude than R wave, MP36R from Biopac® acquisition box was used instead. All the analysed ECG signals were acquired using this platform. The figure bellow shows an ECG wave and the identified QRS complex and T wave with the developed algorithm presented in 5.2.

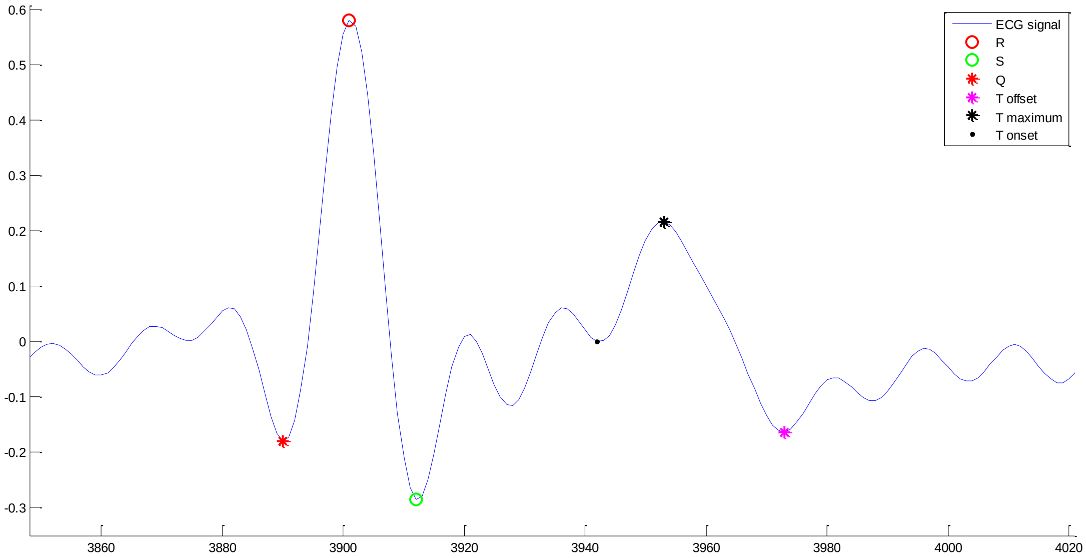


Figure 46: ECG signal acquired with MP36R with QRS complex and T wave identification using the developed algorithm;

In order to evaluate the performance of the developed algorithm, the results obtained with the algorithm were visually inspected by a HEE. The following table the presents the results obtained for performance evaluation.

Table 12: Performance of the developed algorithm for ECG analysis in 438 pulses from 4 different people;

	Sensitivity (%)	P^+ (%)	Error (%)
Q	99.5	100	0.5
R	100	100	0
S	99.7	99.5	0.7
T onset	92	93	1.4
T maximum	99.5	100	0.5
T offset	97.4	97.8	4.6

6.3. Data mining

The data mining process relies on several alternative experiments in order to find the better solution for the problem. In this subchapter the experiments final results and reasons behind the selection of patient measurements to be evaluated and the use of particular algorithms will be presented and discussed.

Table 13 presents the dataset parameters, their units and data type. Some spatial parameters will be named by the letter R and a number (from 1 to 6).

Table 13: Dataset attributes, units and data type;

Attribute #	Feature	Units	Attribute data type
1	Age	Years	Numeric
2	Sex	Male (1), Female (2)	Nominal
3	Smoking	Non-smoker (1), Smoker (2)	Nominal
4	Weight	Kg	Numeric
5	Height	meters	Numeric
6	BMI	Kg/m^2	Numeric
7	Diabetes	Negative (1), Type I (2), Type II (3)	Nominal
8	SBP	mmHg	Numeric
9	DBP	mmHg	Numeric
10	HR	bpm	Numeric
11	T_{ba} (R1)		Numeric
12	T_{cb} (R2)		Numeric
13	$T_{a'b}/T_{ba}$ (R3)		Numeric
14	h_c/h_b (R4)		Numeric
15	$\begin{cases} h_b - h_r & if\ t_r < t_b \\ h_r - h_b & if\ t_r > t_b \end{cases}$ (R5)		Numeric
16	$\begin{cases} {h_r}/{h_b} & if\ t_r < t_b \\ -\left({h_r}/{h_b}\right) & if\ t_r > t_b \end{cases}$ (R6)		Numeric
17	AI	Percentage (%)	Numeric
18	Class	Unhealthy (1), Healthy (2)	Nominal

Key: BMI= Body mass index; SBP= Systolic blood pressure; DBP= Diastolic blood pressure; HR= Heart Rate; bpm= beats per minute; AI= Augmentation index;

Information Gain algorithm with a 10-fold cross validation was used to discover which spatial parameters are the most important. Table bellow displays the attributes with its relative average rank and average merit.

Table 14: Spatial attribute ranking using Info Gain attribute evaluator using 10-fold cross-validation;

Attribute	Average rank	Average merit
R5	1.1 ± 0.3	0.543 ± 0.01
R6	2.2 ± 0.6	0.536 ± 0.01
AI	2.7 ± 0.46	0.535 ± 0.011
R3	4.3 ± 0.46	0.468 ± 0.009
R1	4.7 ± 0.46	0.468 ± 0.009
R2	6.0 ± 0.00	0.444 ± 0.008
R4	7.0 ± 0.00	0.216 ± 0.008

Average merit indicates how important one attribute is (higher values correspond to better parameters) and average rank tells the average ranking of each attribute throughout the 10 folds. It can be conclude that h_b-h_r is the best attribute for data splitting and h_c/h_b is the worst. In general, these parameters show to be capable of accurate data splitting. It is important to note that both h_b-h_r and h_r/h_b had a higher ranking than AI.

6.3.1. Data visualization

Data visualization using histograms is a step that may be useful as a first step to distinguish the attributes which better can be used to classify instances as healthy and unhealthy.

An approximate amount of instances of different class was used (figure 47) to make sure that the verified relations between attributes are credible.

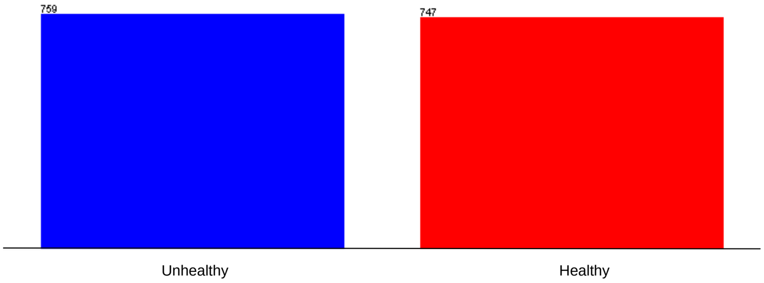


Figure 47: The amount of instances in the dataset classified as unhealthy (blue) and healthy (red) is very similar;

Since the amount of instances from diabetic patients classified as unhealthy is few and between the instances of non-diabetic patients an equal distribution of instances classified as healthy and unhealthy is verified, this parameter doesn't seem to be of great importance in this dataset and will not be taken into account, being removed during data pre-processing. In the following figure a distribution of instances according to diabetes status is shown.

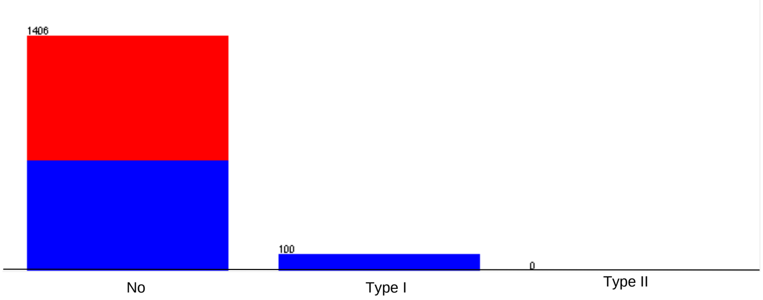


Figure 48: Distribution of dataset instance's according to patients diabetes status and classification (blue for unhealthy and red for healthy). In this dataset, 1406 instances are from patients that don't suffer from diabetes, 100 instances are identified as type I diabetes (all unhealthy) and none instance is identified as type II;

Among smokers and non-smokers, the amount of healthy and unhealthy instances for both groups is very similar, so, this attribute is not relevant (in this dataset) and therefore must be eliminated during pre-processing.

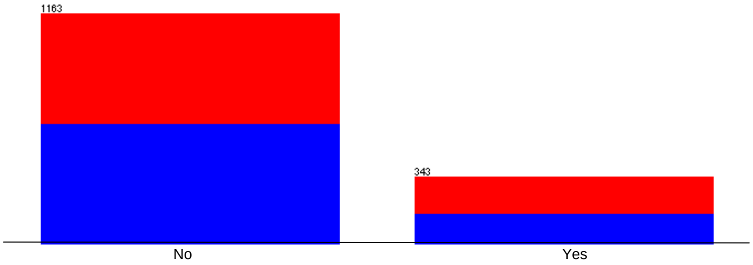


Figure 49: Distribution of instances classified as healthy (red) and unhealthy (blue) among smokers and non-smokers. 1163 instances are form individuals that don't smoke and 343 are from smokers. Each group has a similar amount of healthy and unhealthy instances;

Figure 50 shows the distribution of ages in the dataset. This parameter depends of available volunteers for the study and therefore can be misleading because unhealthy population has ages above 38 years and this would make the classification algorithm to choose this parameter as the one with the highest information gain and would lead the algorithm to classify data according to age. The group in blue consists in elderly and hypertensive people and therefore, present a characteristically profile of APW with arterial stiffness, which is what is being tried to be identified.

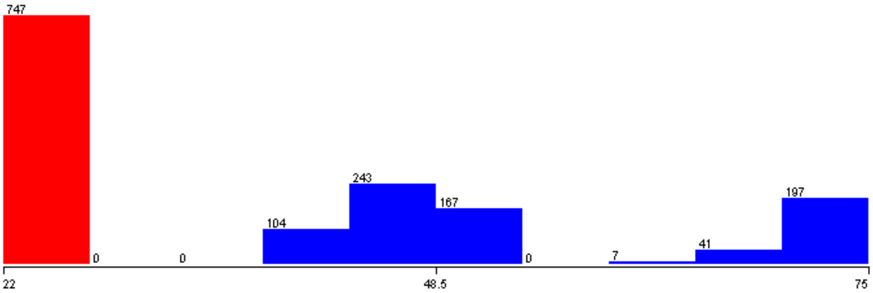


Figure 50: Instances distribution according to age and class. All instances classified as healthy (red) are from a young group (ages between 22 and 27 years) while unhealthy (blue) are from an older group (ages between 38 and 75 years);

By observing the body mass index (BMI) histogram in figure 51, it's easily assumed that this parameter is redundant. Both healthy and unhealthy instances are widely distributed and no visible relation can be established with this parameter.

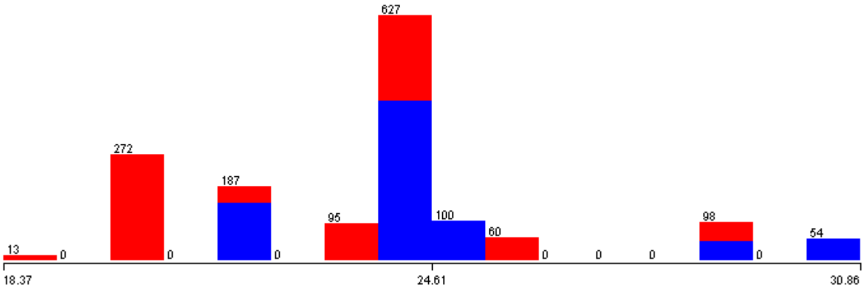


Figure 51: Body mass index histogram ;

By the way the dataset was constructed the distribution of health and unhealthy instances according to both SBP and DBP was expected (figures 52 and 53). It's very clear that all unhealthy people had SBP and DBP above 142 mmHg and 90 mmHg, respectively, while healthy people had SBP and DBP bellow the values aforementioned, which is consistent with the values on table 3.

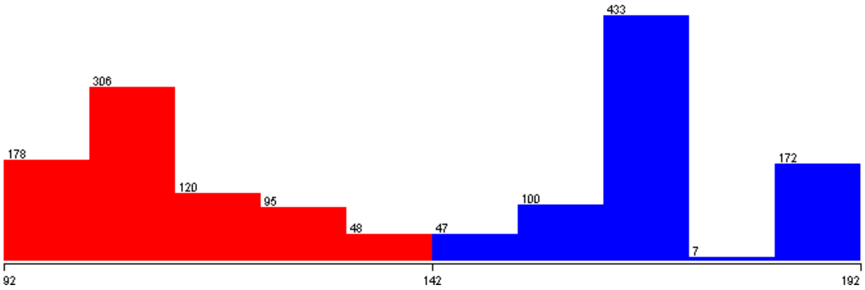


Figure 52: Systolic blood pressure (SBP) distribution. Instances classified as unhealthy have values of SBP above 142 mmHg while instances classified as healthy have SBP values below 142 mmHg;

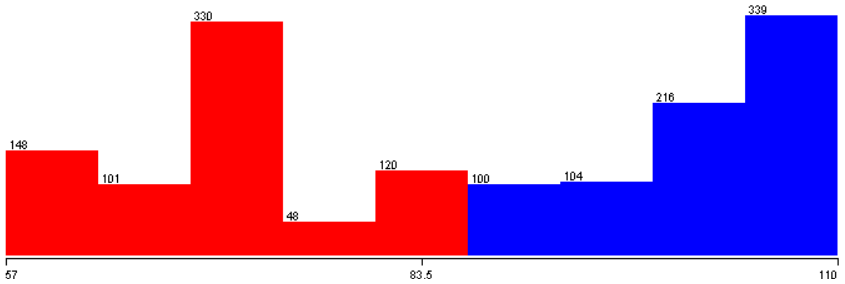


Figure 53: Diastolic blood pressure (DBP) distribution. Instances classified as unhealthy have values of DBP above 90 mmHg while instances classified as healthy have DBP values below 90 mmHg;

From the analysis of the histogram of the spatial parameter labeled as R1 (figure 54), there seems to be a prevalence of unhealthy measurements over healthy measurements when R1 takes values above 0.18, while healthy measurements seem to prevail when R1 takes values below 0.16.

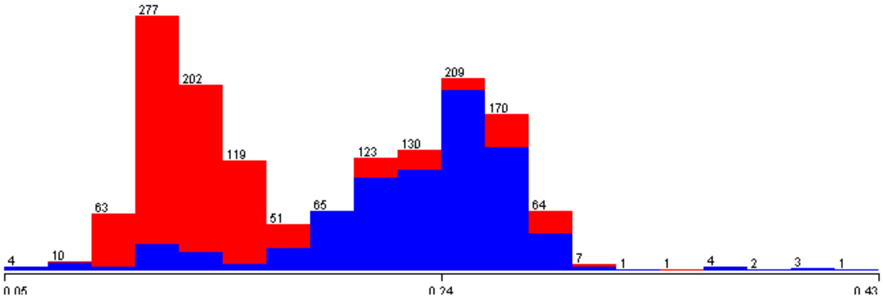


Figure 54: R1 histogram showing a prevalence of unhealthy for R1 greater than 0.18 and a prevalence of healthy instances for R1 smaller than 0.16;

From the analysis of the histogram of the spatial parameter labeled as R2 (figure 55), there seems to be a prevalence of unhealthy measurements over healthy measurements when R2 is smaller than 0.16, while healthy measurements seem to prevail when R2 is greater than 0.18.

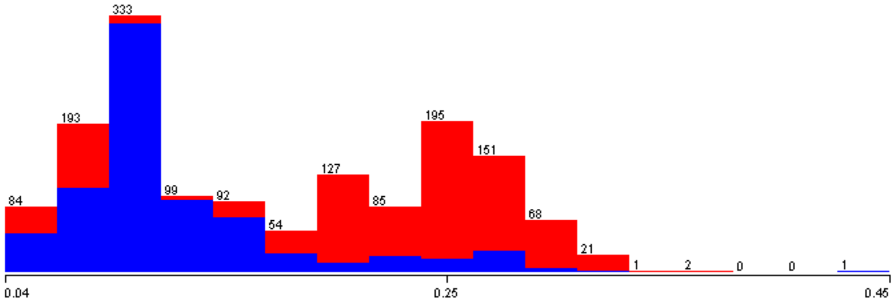


Figure 55: R2 histogram showing a prevalence of unhealthy for R2 smaller than 0.16 and a prevalence of healthy instances for R2 greater than 0.18;

From the analysis of the histogram of the spatial parameter labeled as R3 (figure 56), there seems to be a prevalence of unhealthy instances over healthy instances when R3 takes values lower than 4.8, while healthy instances seem to prevail when R3 takes values higher than 4.8.

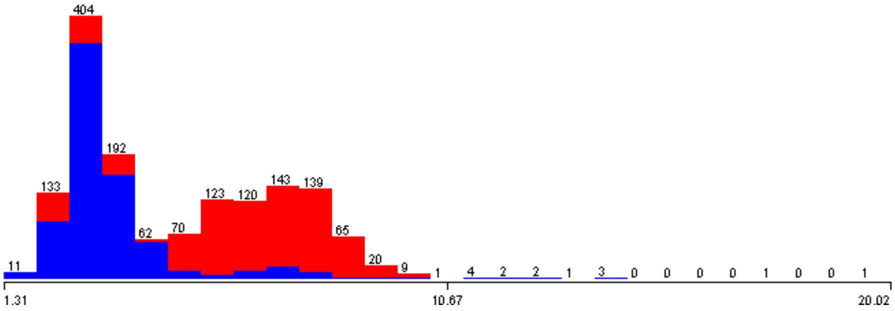


Figure 56: R3 histogram showing a prevalence of unhealthy for R3 smaller than 4.8 and a prevalence of healthy instances for R3 greater than 4.8;

From the analysis of the histogram of the spatial parameter labeled as R4 (figure 57), there seems to be a prevalence of unhealthy instances over healthy instances when R4 takes values between 0.7 and 0.86, while healthy measurements seem to have an uniform distribution along the plot.

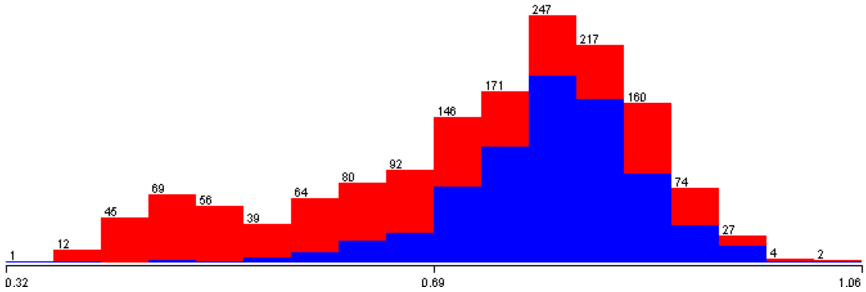


Figure 57: R4 histogram suggesting a prevalence of unhealthy instances when 0.7 < R4 < 0.86, while healthy instances seem to have almost an uniform distribution;

From the analysis of the histogram of the spatial parameter labeled as R5 (figure 58), there seems to be a prevalence of unhealthy instances over healthy instances when R5 takes positive values, while healthy measurements seem to prevail when R5 is negative.

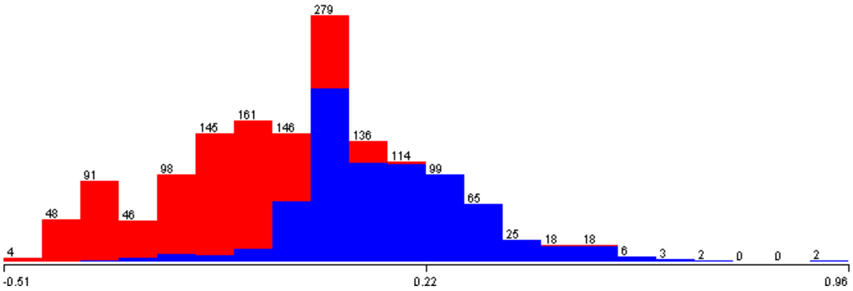


Figure 58: R5 histogram suggesting a prevalence of unhealthy instances when R5 takes positive values and a prevalence of healthy instances when R5 is negative;

From the analysis of the histogram of the spatial parameter labeled as R6 (figure 59), there seems to be a prevalence of unhealthy instances over healthy instances when R6 takes positive values, while healthy instances seem to prevail when R6 takes negative values.

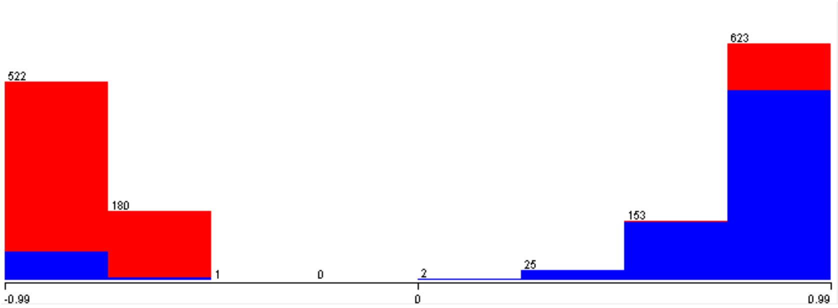


Figure 59: R6 histogram suggesting a prevalence of unhealthy instances when R6 takes positive values, specially between 0.25 and 0.99 and a prevalence of healthy instances when R6 is negative, specially between -0.5 and -0.99;

From the analysis of the histogram of AI (figure 60), there seems to be a prevalence of unhealthy instances over healthy instances when AI is positive, while healthy instances seem to prevail when AI is negative.

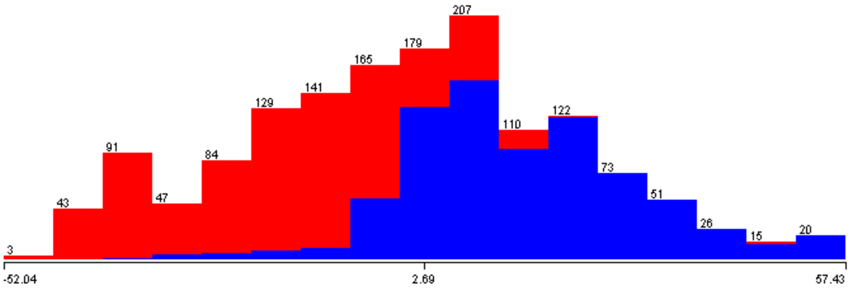


Figure 60: Augmentation Index (AI) histogram showing a prevalence of unhealthy instances for positive values of AI and a prevalence of healthy instances for negative AI;

Data visualization can be useful to visualize the geometric transformations and projections of the data to produce useful and insightful visualizations. Geometric projections can be helpful for finding interesting multiparameter datasets.

Figure 61 shows a scatterplot-matrix that allows the visualization of data in rows and columns of cells with simple graphical depictions.

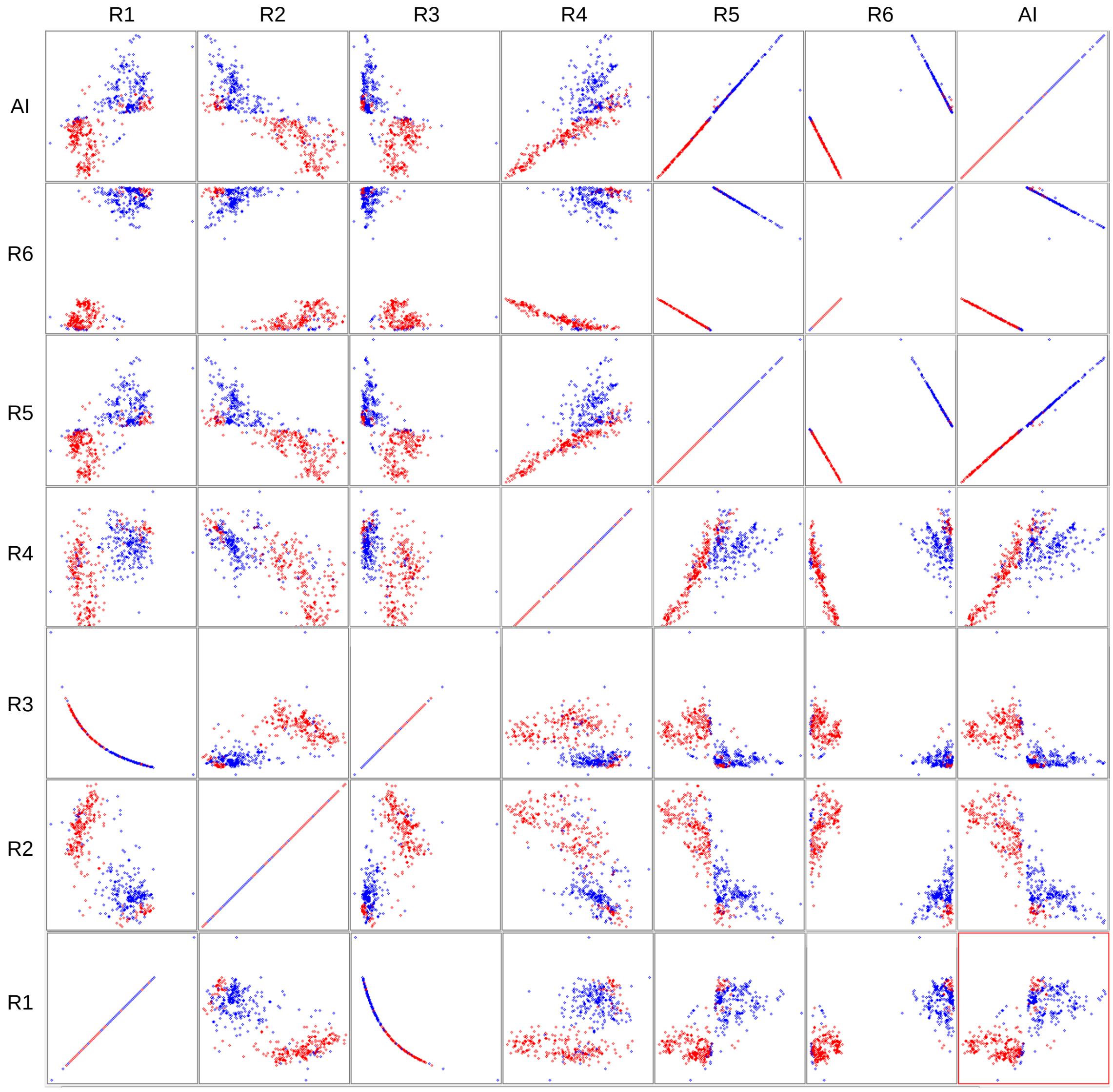


Figure 61: Scatterplot-Matrix for multiparameter data distribution visualization of healthy (red) and unhealthy (blue) instances. For each cell, the xx axis parameter is presented on the top of each column while the yy axis is on the left of each row;

By the inspection of the scatterplot-matrix in figure 61, it's possible to conclude that the chosen spatial parameters are very interesting, being possible to accurately distinguish healthy from unhealthy people. In every cell it's clearly visible the formation of two major groups, corresponding to healthy (red) and unhealthy (blue) people.

6.3.2. Data pre-processing

The feature subset selection process reduces the dimensionality of data to be analyzed, speeds up execution of learning algorithms, improves the performance of data mining algorithms and improves the comprehensibility of the output.

Since the Weka unsupervised filter for instance removal “RemoveWithValues” removes all instances bellow a certain threshold, it cannot be applied to this dataset. Instead, Matlab was used to remove instances containing outliers. Instances containing AI values of zero were eliminated from the dataset, because these instances corresponded to pulses where the algorithm for prominent point’s identification failed.

In order to better understand the distribution of healthy and unhealthy people according to spatial parameters values and to reduce data dimensions, in the clustering analysis an unsupervised discretization filter was applied to numerical attributes. The same filter was not applied during classification because it was experimentally proved this filter reduces classification accuracy.

Furthermore, for classification purposes, only the APW spatial attributes will be used, making the future model dependent of only these spatial attributes, allowing to produce a diagnose considering only the morphology of APW pulses.

6.3.3. Classification

After data pre-processing, several learning algorithms needed to be tested so that the algorithm with best performance is used for model construction. As pointed out by many authors [43, 44], *a priori*, no classification method is better than any other and each method has a target class for which they best fit. Therefore, experiments needed to be made in order to investigate which algorithm best suited this dataset.

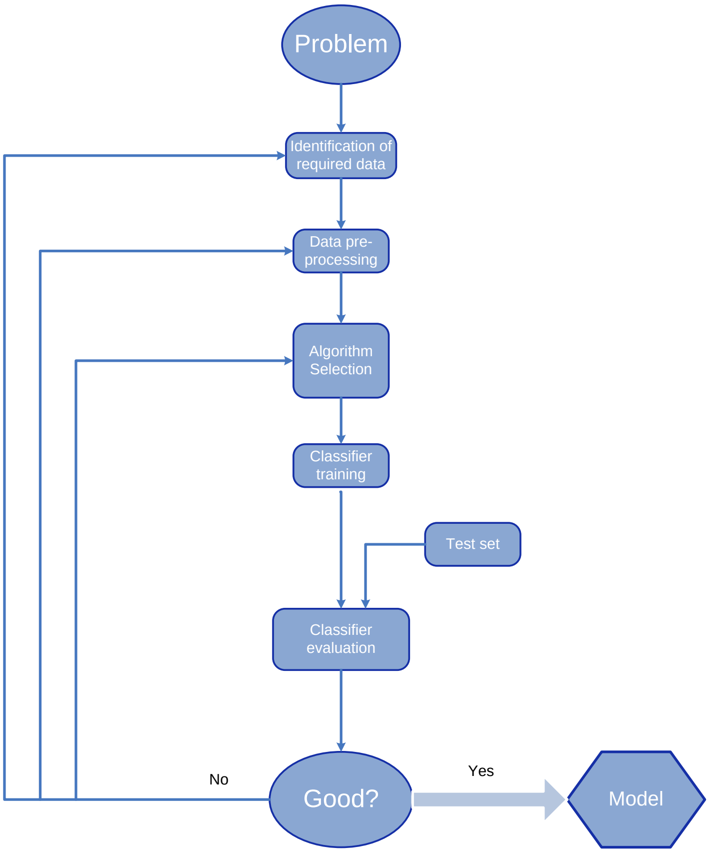


Figure 62: Process of supervised learning with classification algorithms [adapted from 60];

6.3.3.1. Classifier Selection

Popular machine learning techniques were chosen to be tested. Two statistical methods (Bayes-Net and Naïve Bayes), two decision trees (J48, the java version of C4.5 and Random forest) and two decision rules (JRIP and Decision table).

Tables 15 and 16 show the result of the evaluation of these algorithms. Table 15 shows a summary of the accuracy obtained for the different classifiers that were tested using the created dataset and with a 10-fold cross-validation. Cross-validation is a technique for assessing classifier accuracy, by randomly sample partitions of the given data. In this case, the initial dataset is partitioned in 10 exclusive subsets of the same size (D_1, . . ., D_{10}). For the iteration i, D_i functions as the test set and the remaining subsets are used to train the classifier. The classifier accuracy is then estimated by the overall number of correct classifications divided by the total number of instances in the initial dataset. The 10-fold cross-validation has been extensively tested and proved to provide accurate estimate error [62].

The Receiver Operating Characteristic (ROC) curve is a graphical plot of sensitivity given by true positive rate as a function of false positive rate, being a good tool for visualizing a classifier performance and to select a suitable decision threshold. The area under curve (AUC) of ROC is often used as a statistic for model comparison. The larger AUC is, the more accurate the classifier is. For example, an ideal classifier has an AUC of 1 while a poor one has an area of 0.5.

The accuracy of a classifier is the percentage of correctly classified instances in a test set, measuring how well the classifier recognizes instances of the various classes [41].

$$sensitivity = \frac{True\ Positive}{Positive} \tag{23}$$

$$specificity = \frac{True\ Negative}{Negative} \tag{24}$$

$$precision = \frac{True\ Positive}{(True\ Positive + Fake\ Positive)} \tag{25}$$

$$recall = \frac{True\ Positive}{(True\ Positive + Fake\ Negative)} \tag{26}$$

$$accuracy = sensitivity \frac{Positive}{(Positive + Negative)} + specificity \frac{Negative}{(Positive + Negative)} \tag{27}$$

Table 15: Several classifiers were tested, all with 10-fold cross-validation. J48 and Random forest had the best performances;

Classifier	Accuracy (%)	Relative absolute error (%)	Weighted average		
			Precision (%)	Recall (%)	ROC area
J48	96.28	10.00	96.3	96.3	0.97
Random forest	96.75	10.70	96.8	96.7	0.99
Bayes-Net	86.12	28.01	86.2	86.1	0.96
Naïve Bayes	86.45	27.40	86.6	86.5	0.90
JRIP	94.95	14.08	95.0	95.0	0.97
Decision table	92.96	23.10	93.0	93.0	0.97

A confusion matrix displays the amount of correct and incorrect classifications from each class. It appears on the form of a table showing the differences between the true and predicted classes for a set of labeled instances. Each column represents the predicted classes while each row represents the actual classes.

Table 16: Confusion matrices for the tested classifiers for the dataset with 1506 instances;

Classifier	Actual class	Predicted Class	
		Healthy	Unhealthy
J48	Healthy	719	28
	Unhealthy	28	731
Random forest	Healthy	719	28
	Unhealthy	21	738
Bayes-Net	Healthy	624	123
	Unhealthy	86	673
Naïve Bayes	Healthy	623	124
	Unhealthy	80	679
JRIP	Healthy	718	29
	Unhealthy	47	712
Decision table	Healthy	705	42
	Unhealthy	64	695

ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making. In figure 62, ROC plots (left side) and respective cost/benefit plots (right side) are represented. The chosen thresholds are displayed as well in the lower right corner of ROC plots and each blue cross represents the point associated with the selected threshold, indicating the expected true positive rate and false positive rate.

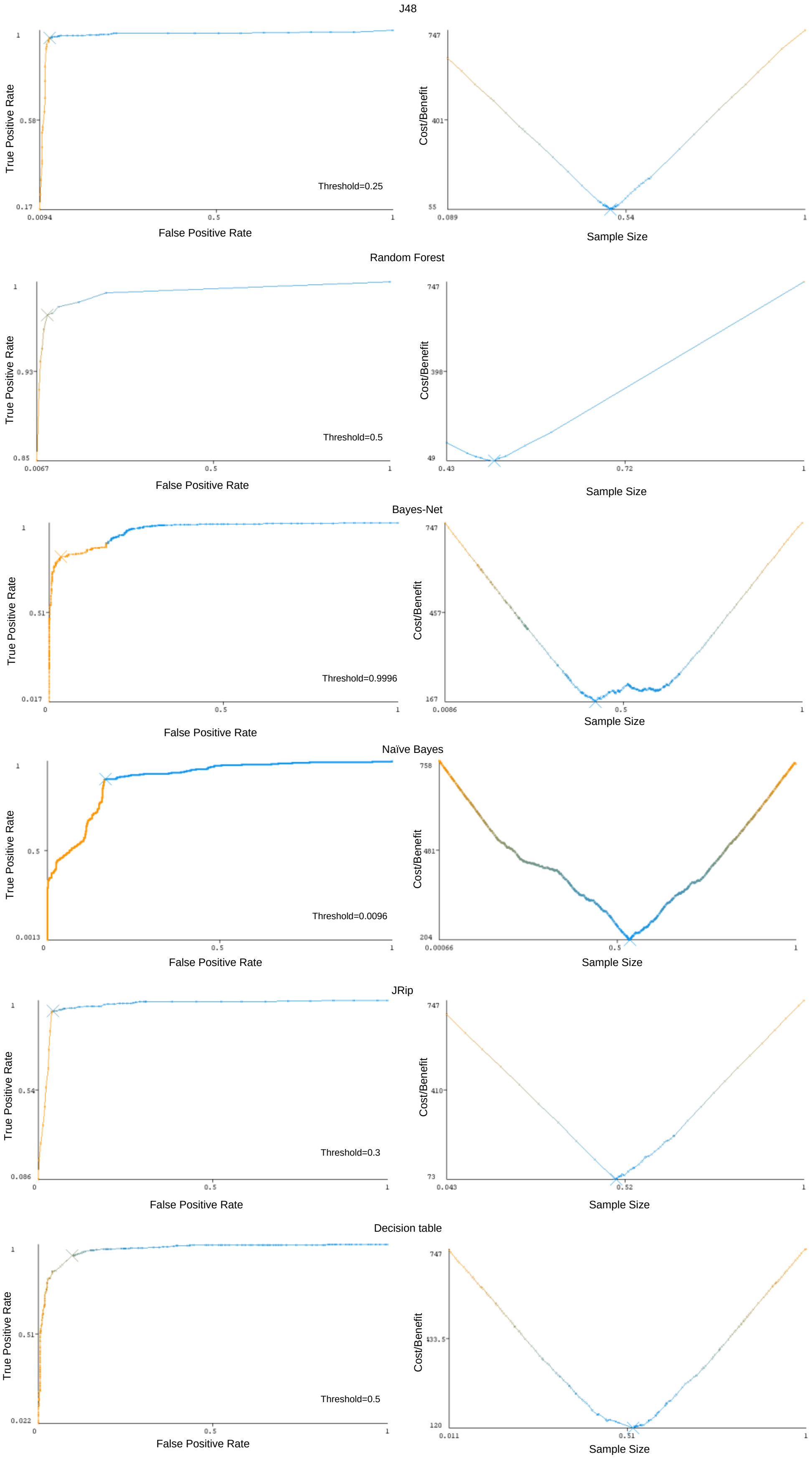


Figure 63: ROC curves and cost/benefit plot for the classification algorithms in analysis for arterial stiffness prediction (class 1). The value of the chosen threshold is displayed in the lower right corner for each ROC plot. These thresholds were chosen in order to minimize the cost/benefit;

From the analyzed information so far, the two decision trees stand out from the rest of the learning algorithms. Both, J48 and Random Forest show an overall accuracy of 96.28% and 96.75%, respectively and a ROC AUC of 0.97 and 0.99, respectively, which are the major parameters for model selection. Figure 64 shows an overlap of the ROC curves for these two algorithms. It’s visible that Random Forest ROC curve has a greater area and encompasses all J48 ROC curve. Therefore, Random Forest was chosen to be used for model construction due to its best performance.

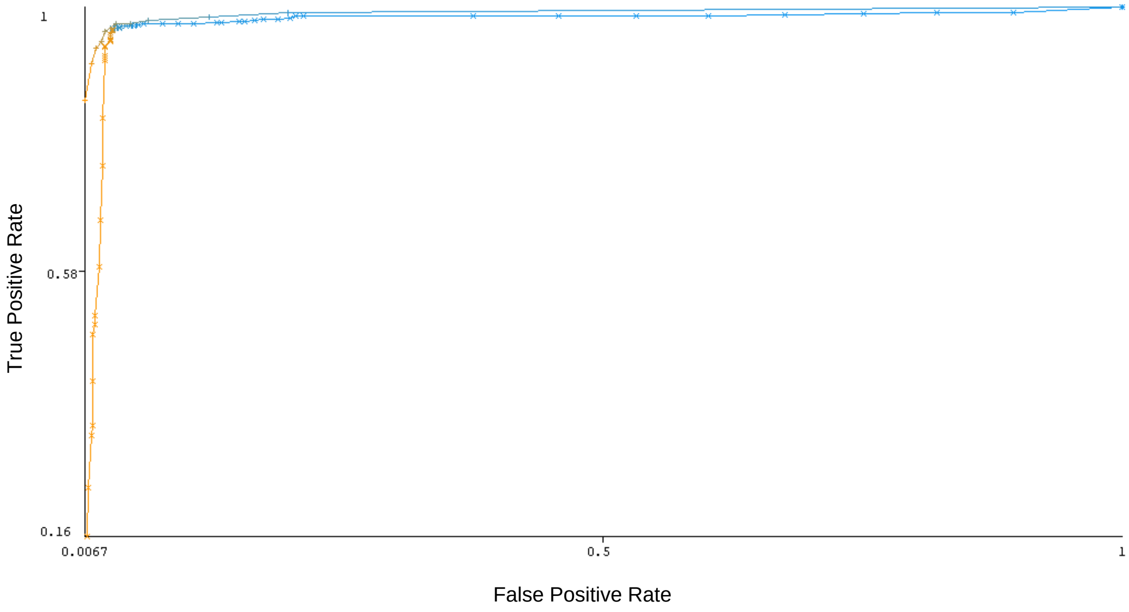


Figure 64: ROC curves for Random Forest (+) and J48 (x) classifiers. It’s visible that Random Forest dominates J48 in all areas of the chart, so it should be a better classifier to use for arterial stiffness prediction (class 1);

6.3.3.2. Prediction

After the model construction, as described in the previous section, this same model can be applied to predict instance’s class based in the spatial parameters presented before. It should be remembered that each instance is obtained from the analysis of a single APW.

For model testing, only the dataset for diagnose purpose was used. Since the class attribute in this dataset is unknown, the model will give a prediction. The group of individuals from whom the APW information was used to construct the dataset is a set of people with a family history of cardiovascular diseases and with suspicious forms of APW, although still no one suffers from cardiovascular diseases.

Table 17 shows the results obtained by testing the model in this dataset. Cardiovascular risk factor is given by the ratio between the number of "unhealthy" predictions and the total number of instances for each person. Each cardiovascular risk factor measure has an associated standard deviation of ±3.75%, due to the accuracy of Random Forest classifier upon model construction.

Table 17: The model was applied to data from people with unknown class. Diagnosis prediction is given by the cardiovascular risk factor parameter ±3.75%;

Person #	Number of instances	Number of "Unhealthy" predictions	Number of "Healthy" predictions	Cardiovascular risk factor (%)
1	128	125	3	97.66
2	89	75	14	84.27
3	95	89	6	93.68
4	101	101	0	100.00
5	100	99	1	99.00
6	126	19	107	15.08

A high cardiovascular risk factor may not mean that a person suffers already from a cardiovascular disease. Instead, it gives the probability of that person to suffer from any cardiovascular disease in the future. From the analysis of the predictions made for this dataset in table 18, it's expected that the first 5 people suffer from a cardiovascular disease in the future while person #6 has a low risk.

6.3.4. Clustering

K-means clustering algorithm is commonly used for unsupervised learning tasks being one of the most popular and efficient clustering methods (Hartigan and Wang, 1979; Lloyd, 1957; MacQueen, 1967), using one centroid to represent each cluster [64, 65, 66, 67].

As been said before, for cluster analysis all the attributes in the dataset will be used and continuous attributes will be discretized. Once again, only the dataset consisting of hypertensive and normotensive people is used. The table bellow presents the obtained centroids using K-means algorithm, with k=5 and discrete attributes.

Table 18: Clusters centroids obtained with k-means algorithm for k=5 for discrete attributes;

	Cluster #				
	1	2	3	4	5
Clustered instances	406 (27%)	577 (38%)	258 (17%)	167 (11%)	98 (7%)
Attribute					
Age	[48.5:53.8]	< 27.3	< 27.3	[43.2:48.5]	> 69.7
Sex	Female	Male	Male	Male	Female
Smoking	Non-smoker	Non-smoker	Non-smoker	Non-smoker	Smoker
Weight	[69:72.8]	[61.4:65.2]	[53.8:57.6]	[57.6:61.4]	[76.6:80.4]
Height	[1.68:1.704]	[1.73:1.75]	[1.63:1.66]	[1.63:1.66]	[1.63:1.66]
BMI	[23.36:24.61]	[23.36:24.61]	[19.62:20.86]	[20.86:22.11]	[28.36:29.61]
Diabetes	Negative	Negative	Negative	Negative	Negative
SBP	[162:172]	[102:112]	< 102	> 182	[162:172]
DBP	> 104.7	< 62.3	[67.6:72.9]	> 104.7	[99.4:104.7]
HR	[65.4:69]	[76.2:79.8]	[72.6:76.2]	[69:72.6]	[65.4:69]
T_{ba} (R1)	[0.24:0.28]	[0.09:0.13]	[0.20:0.24]	[0.24:0.28]	[0.24:0.28]
T_{cb} (R2)	[0.08:0.12]	[0.25:0.29]	< 0.08	[0.08:0.12]	< 0.08
$T_{a'b}/T_{ba}$ (R3)	< 3.18	[6.92:8.80]	[3.18:5.06]	< 3.18	< 3.18
h_c/h_b (R4)	[0.77:0.84]	[0.62:0.69]	[0.84:0.92]	[0.77:0.84]	[0.77:0.84]
$(\mathbf{R5})\begin{cases} h_b - h_r & if\ t_r < t_b \\ h_r - h_b & if\ t_r > t_b \end{cases}$	[0.22:0.37]	[-0.22:-0.07]	[-0.07:0.08]	[0.08:0.22]	[0.08:0.22]
$(\mathbf{R6})\begin{cases} {h_r}/{h_b} & if\ t_r < t_b \\ -\left({h_r}/{h_b}\right) & if\ t_r > t_b \end{cases}$	[0.59:0.79]	< - 0.79	> 0.79	>0.79	> 0.79
AI	[13.64:24.59]	[-19.20:-8.25]	[2.69:13.64]	[2.69:13.64]	[2.69:13.64]
Class	Unhealthy	Healthy	Healthy	Unhealthy	Unhealthy

Cluster centroids are the mean vectors for each cluster and therefore, the value of each centroid attribute represents the mean value for that attribute in the cluster. Thus, centroids can be used to characterize the clusters [67].

In a first view, from the five obtained clusters, three are mainly formed by unhealthy people (clusters 1, 4 and 5) and two are formed by healthy people (clusters 2 and 3). As expected, “unhealthy” clusters have elderly people while young people is majorly distributed in “healthy” clusters.

Due to the limitations of this dataset, as there was a lack of volunteers, no ascertained conclusions can be taken from some attributes. Nevertheless, some interesting patterns can be observed.

A visible pattern in these clusters associates the parameters AI, R5 and R6. These parameters are strongly related with the reflected wave position. The systolic peak amplitude

(h_b) is always greater than the reflected peak amplitude (h_r), therefore, positive values of AI, R5 and R6 indicate that the reflected wave appears before the systolic wave. As it is widely stated in the literature, this is associated with unhealthy pulses, which is also proved with clusters 1, 4 and 5. Cluster 3 also presents these characteristics and the reasons for this will be addressed later.

On other hand, negative values of AI, R5 and R6 are associated mean that the reflected wave appears after the systolic peak and is associated with healthy pulses.

Also, combining the suggested results on table 18 and the scatterplot-matrix on figure 61 (first line, columns 5 and 6), it can be concluded that AI and R5 are directly proportional (figure 65) while AI and R6 are inversely proportional.

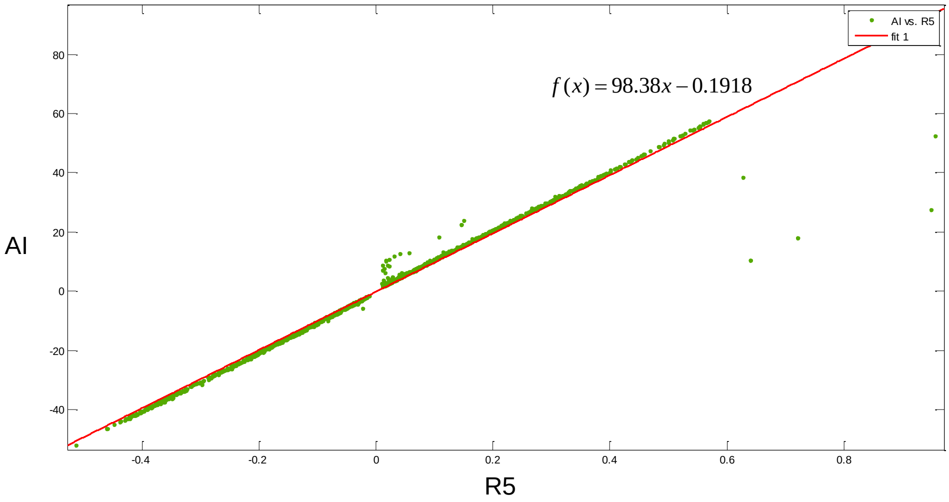


Figure 65: AI and R5 are directly related by$f(x) = 98.38x - 0.1918$;

As it was suggested in section 6.3.1, unhealthy patients seem to have higher values of R1 (mostly between 0.24 and 0.28), meaning that it takes longer to unhealthy APW to reach the systolic peak. Therefore, the gradient of the APW on the segment onset-systolic peak is smaller in unhealthy pulses, which could mean that the variation of arterial pressure during this segment is smaller when compared to healthy pulses. However, unhealthy pulses have usually a reflected wave before the systolic peak, as it can be seen by analyzing the parameters AI, R5 and R6, making what was stated before not as linear as it seems. In healthy pulses, the reflected wave tends to appear after the systolic peak, thus, this kind of waves have negative AI, negative R5 and negative R6 .

According to table 18, R3 takes values lower than 3.18 for unhealthy people and greater than 3.18 for healthy people. It can be conclude that the majority of the descent time of an unhealthy APW takes approximately less than 3 times the ascent time of the forward wave. On other hand, the descent time of a healthy APW takes between 3 to almost 9 times the time of the ascent time of the forward wave.

Another verified pattern relates R3 and R1 in an exponential way (figure 66). As the ascent time of the forward wave of an APW increases, the number of times that the descent forward wave is greater than the time of the forward wave, decreases exponentially. However, the descent time of the forward wave is always greater than its ascent time.

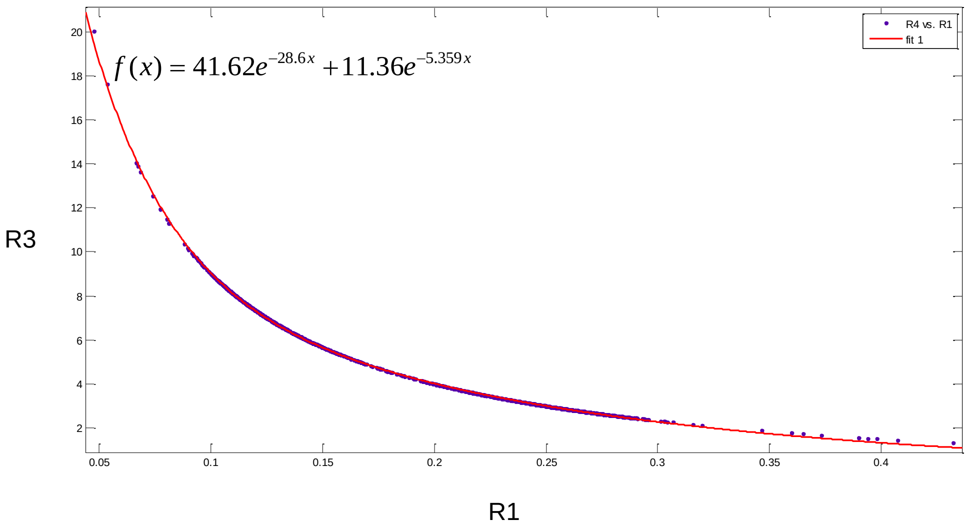


Figure 66: R3 and R1 plot with an exponential fitting curve. R4 and R1 are exponentially correlated by the equation$f(x) = 41.62e^{-28.6x} + 11.36e^{-5.359x}$;

R2 seems to take smaller values than 0.12 in unhealthy people while in a healthy population seem to have greater values than 0.25. However, ascertained conclusions cannot be taken from this parameter due to cluster 3. Nevertheless, this would mean that the relative time between the systolic peak and the dicrotic notch is smaller in unhealthy populations.

One possible error during dataset construction may have occurred, resulting in the misclassification of some individuals. Individuals in cluster 3, which are supposedly young and healthy, have a positive AI, a positive R6 and thus a positive R5, meaning that the reflected wave appear earlier than the systolic peak, suggesting a wave characteristic of unhealthy people. This fact has a great influence in other spatial and temporal parameters, being possible to find some similarities in waveform parameters between cluster 3 and unhealthy clusters (1, 4 and 5). This way, people whose pulses present the characteristics of cluster 3, though young and with normal SBP and DBP, may probably develop cardiovascular diseases in the future if a healthy lifestyle is not adopted.

Based in the obtained results, it’s possible to draw characteristically healthy and unhealthy APW, based on the built dataset. Figures 67 and 68 represent unhealthy and healthy pulses, respectively, based in the identified patterns.

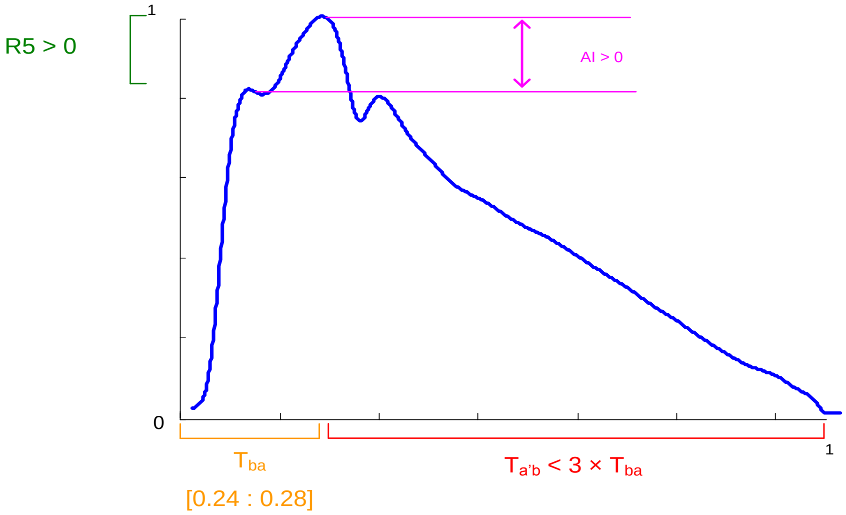


Figure 67: Patterns identified on characteristically APW of unhealthy people;

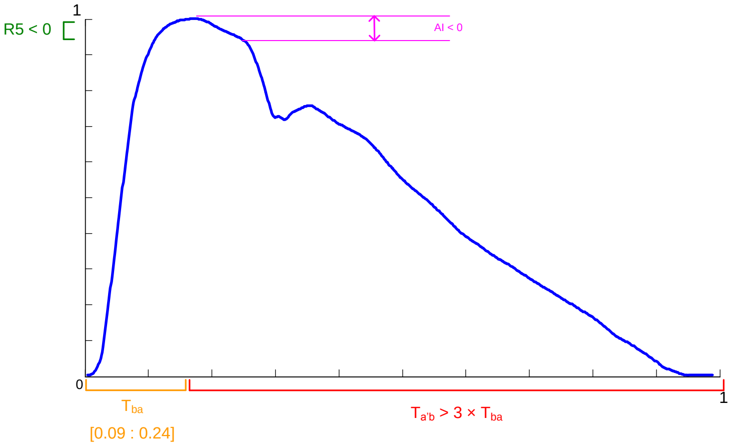


Figure 68: Patterns identified on characteristically APW of healthy people;

7. Conclusion

The major goal of this thesis was the development of a system capable of produce a clinical diagnosis by applying a prediction model based on the analysis of a physiological signal such as APW. Algorithms for APW signal processing have been developed, allowing to obtain clinical relevant information from a raw signal acquired from a previously built prototype. Furthermore, a database was built, allowing the storage of information that may be used in other steps along this work. Also, an algorithm for ECG analysis has been developed, although it proved to be ineffective in the analysis of some signals since it relies on R peak detection by a threshold.

A prototype composed by multiple PZ sensors for APW signal acquisition is being developed in order to achieve operator independence and to produce a signal free of artifacts by applying an adaptive noise cancellation algorithm.

A module for ECG signal acquisition has been built enabling future comparisons with APW signals in order to achieve a more complete diagnosis.

7.1. General Results

The collected signals and the studies carried out along this work have proved the potentialities of medical instruments in this field. The developed prototype has proved to be effective in the acquisition of APW signals pre-integration, although future improvements need to be made, especially in the development of an algorithm capable of choosing the best signal, which has been proved to be a complicated task since the signals acquired at the same time present some similarities.

The simultaneous analysis of APW and ECG signals is expected to open new challenges in the cardiovascular system diagnosis. The capabilities of such a classical technique as ECG in the cardiovascular system analysis have already been widely explored, but not simultaneously with APW. The developed system for ECG acquisition still needs to be improved, mostly due to the presence of baseline noise. The algorithm for ECG analysis also needs to be improved mostly in the detection of R peak. A method based on frequency analysis may be the best answer for this problem.

For APW signal analysis, the developed method consisting of several steps has proven to be very effective. The first step consisted on the construction of an algorithm for onset calculation and baseline removal, which proved to be very effective. The method for anomalous beats detection proved to be useful and effective, allowing to achieve a clean signal. Heart rate variability estimation using the signal obtained from the previous steps seems legit, though it needs to be compared with products already established in the market.

From the clean APW signals, a series of spatial parameters were obtained, being stored in a database, along with other patient's information. Using these parameters from several of the analyzed signals, a classification model has been constructed by using the supervised learning algorithm Random Forest, obtaining a good accuracy in the model construction. The results obtained by the application of this model to are good, although future comparisons need to be made. Also, a dataset with a more diversified population needs to be analyzed in order to give credibility to this method and to increase its accuracy.

As for the cluster analysis, it has proven to be effective in distinguishing different groups of people based mostly on APW parameters analysis. A key limitation of k-means model is based on its tendency to form spherical clusters. The clusters are expected to be of similar size, so that the assignment to the nearest cluster centre is the correct assignment.

8. References

[1] - http://www.who.int/cardiovascular_diseases/en/ (accessed in 16/08/2011);

[2] - Safar, Michel, Arteries in Clinical Hypertension, Lippincott-Raven Publishers, 1996;

[3] - Vermeersch, S. J., et al., Determining carotid artery pressure from scaled diameter waveforms: comparison and validation of calibration techniques in 2026 subjects, Physiol. Mea., 29:1267-1280, 2008;

[4] - Pereira, Tânia, et al., Signal analysis is a new optical pulse waveform profiler for cardiovascular applications, Acta press, 2011;

[5] - McLaughlin, J., McNeill, M., McCormack P. D., Piezoelectric sensor determination of arterial pulse wave velocity, Physiol. Meas. ,24:693-702,2003;

[6] - Clement, F., Arpaia, P., Cimmino, P., A piezo-film-based measurement system for global haemodynamic assessment, Physiol. Meas, 31:697-714, 2010;

[7] - Ting, S. L., Data mining in biomedicine: current applications and further directions for research, Software Engineering & Applications, 2:150-159, 2009;

[8] - Bronzino, J.D., Biomedical Engineering Handbook, CRC Press LLC, 2000;

[9] - Rogers, Kara, *et al.*, The Cardiovascular System, Britannica, 2011;

[10] - Stranding, Susan, Gray's Anatomy – The Anatomical Basis of Clinical Practice, Elsevier, 2008;

[11] - Purves, William K., *et al.*, Life: The Science of Biology Seventh Edition , Sinauer, 2004;

[12] - Agur, Anne M. R., Dalley, Arthur F., Grant's Atlas of Anatomy, Lippincott Williams & Wilkins, 2009;

[13] - Cheung, Yiu-Fai, Arterial Stiffness in the Young: Assessment, Determinants and Implications, The Korean Society of Cardiology, 2010;

[14] - Rönnback, Mats, Arterial stiffness and cardiovascular risk factors, Helsinki 2007;

[15] - Laurent, Stéphane, Boutouyrie, Pierre, Arterial stiffness: a new surrogate end point for cardiovascular disease?, J Nephrol 2007; 20 (suppl 20): S45-S50;

[16] - Laurent, Stéphane, *et al.*, Expert consensus document on arterial stiffness: methodological issues and clinical applications, European Heart Journal (2006) 27, 2588–2605;

[17] - Lee, Hae-Young, Oh, Byung-Hee, Aging and Arterial Stiffness, Circulation Journal : Vol. 74 (2010), No. 11;

[18] - Payne, Rupert A., *et al.*, Arterial Stiffness and Hypertension: Emerging Concepts, Hypertension 2010;55;9-14, American Heart Association, 2009;

[19] - Mackenzie, I. S., Wilkinson, I. B., Cockcroft, J. R., Assessment of arterial stiffness in clinical practice, Q J Med. 95, 2002, pp. 67-74.

[20] - Almeida, Vânia Maria Gomes de, Hemodinamic Parameters Assessment – An Improve of Methodologies, Universidade de Coimbra, Setembro 2009;

[21] - Nichols, Wilmer W., Clinical Measurement of Arterial Stiffness Obtained from Noninvasive Pressure Waveforms, AJH, 2005, Vol. 18, pp.3S-10S;

[22] - Murgo, J. P., *et al.*, Aortic input impedance in normal men: relationship to pressure wave forms, Circulation, 1980, Vol. 61, pp. 105-116;

[23] - Avolio, Alberto P., *et al.*, Arterial blood pressure measurement and pulse wave analysis—their role in enhancing cardiovascular assessment, Physiol. Meas., 2010, R1-R47;

[24] - Moens, AI, Die Pulscurve. (The Pulse.), EJ Brill, Leyden, The Netherlands, 1878;

[25] - Korteweg, DJ, Ueber die Fortpflanzungsgeschwindigkeit des Schalles in elastischen Röhren (On the velocity of propagation of sound in elastic tubes), Annalen der Physik und Chemie, New Series 5, 525-542, 1878;

[26] - Bramwell, JC, Hill, AV, The velocity of pulse wave in man, Proceedings of the Royal Society of London. Series B 93 (652): 298–306, 1922;

[27] - Meinders, Jan M., Hoeks, Arnold P. G., Simultaneous Assessment of Diameter and Pressure Waveforms in the Carotid Artery, Ultrasound in Med. & Biol., Vol. 30, No. 2, pp. 147–154, 2004;

[28] - Mitchell, Gary F., Arterial Stiffness and wave reflection: Biomarkers of cardiovascular risk, Artery Res., 2009, 3(2): 56-64;

[29] - Arnett, Donna K., Arterial Stiffness and Hypertension, University of Minnesota, USA;

[30] - http://www.pulsecor.com/augmentation-index.html (accessed in 14/07/2011);

[31] - Wnek, Gary E., Bowlin, Gary L., Biomaterials and biomedical engineering, Chemical Engineering Science*,* Informa Healthcare, USA, 2008;

[32] - Mehta, S. S., Detection of P and T waves in Electrocardiogram, WCECS, 2008;

[33] - Witten, Ian H., *et al.*, Data mining: practical machine learning tools and techniques -3rd edition, Elsevier, 2011;

[34] - Han, Jiawei, Kamber, Michelline, Data mining: concepts and techniques – 2nd edition, Elsevier, 2006;

[35] - Hadzic, Fedja, *et al.*, Mining of Data with Complex Structures, Springer, 2011;

[36] - Kohavi, Ron, Quinlan, Ross, Decision Tree Discovery, Stanford, 1999;

[37] - Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993;

[38] - Ho, Tin, Random Decision Forest, 3rd Int'l Conf. on Document Analysis and Recognition, pp. 278–282, 1995;

[39] - Breiman, Leo, Random Forests, Kluwer Academic Publishers, Machine Learning, 45, 5–32, 2001;

[40] - Breiman, L., Using adaptive bagging to debias regressions, Technical Report 547, Statistics Dept. UCB, 1999;

[41] - Ho, T. K., The random subspace method for constructing decision forests, *IEEE Trans. on Pattern Analysis and Machine Intelligence, 20*(8), 832–844, 1998;

[42] - Amit, Y., Geman, D., Shape quantization and recognition with randomized trees. *Neural Computation,9*, 1545–1588, 1998;

[43] - Pearl, J., "Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning" ,Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA. pp. 329–334, 1985.

[44] - Ben-Gal, Irad, In Ruggeri, Fabrizio, Kennett, Ron S., Faltin, Frederick W., Bayesian Networks, *Encyclopedia of Statistics in Quality and Reliability*, John Wiley & Sons, 2007;

[45] - Thabtah, Fadi Abdeljaber, A review of associative classification mining, Knowledge Engineering Review, 22 (1), pp. 37-65, ISSN 0269-8889, 2007;

[46] - Cohen, W. W., Fast effective rule induction, Machine Learning: Proceedings of the Twelfth International Conference, Lake Tahoe, California, Morgan Kaufmann, 1995;

[47] - Fürnkranz, J., Widmer, G., Incremental reduced error pruning, Machine Learning: Proceedings of the Eleventh Annual Conference, New Brunswick, New Jersey, Morgan Kaufmann, (1994).

[48] - Cohen, W. W., Singer, Yoram, A simple, fast, and effective rule learner, Proceedings of the Sixteenth National Conference on Artificial Intelligence, 1999;

[49] - Almeida, V. G., *et al.*, Piezoelectric probe for pressure waveform estimation in flexible tubes and its application to the cardiovascular system, Sens. Actuators A: Phys. (2011), doi: 10.1016/j.sna.2011.04.048;

[50] - Murata Manufacturing Co. Ltd., Piezoelectric Sound Components. [Online]. Murata Manufacturing Co., Cat. No. P37E-23, Kyoto, Japan, 2010, Available at: http://www.murata.com/products/catalog/pdf/p37e.pdf#7BB_27 (accessed in 29/07/2011);

[51] - http://angioplasty-vir.com/index.php?page=procedures&option=2 (accessed in 27/07/2011);

[52] - Almeida, V. G., et al., A real time cardiac monitoring system – arterial pressure waveform capture and analysis, PEECS., Algarve, 2011;

[53] - http://rmgh.net/wiki/images/8/88/Limbleads.jpg (accessed in 29/07/2011);

[54] - http://www.biopac.com/disposable-electrode-100-education-specifications#LowerTab (accessed in 30/07/2011);

[55] - http://www.analog.com/library/analogDialogue/archives/29-3/low_power.html (accessed in 12/08/2011);

[56] - Li, Bing Nan, *et al.*, On an automatic delineator for arterial blood pressure waveforms, Elsevier, Biomedical Signal Processing and Control 5, 76-81, 2010;

[57] - Nürnberger, Jens, *et al.*, Left ventricular ejection time: a potential determinant of pulse wave velocity in young, healthy males, Journal of Hypertension, 2003, 21:2125-2132;

[58] - Newsham, Alexander C., et al., Development of an advanced database for clinical trials integrated with an electronic patient record system, Computers in biology and medicine, 41:575-586, 2011;

[59] - Almeida, *et al.*, Hemodynamic features extraction from a new arterial pressure waveform probe, Biosignals, Rome, 2011;

[60] - http://www.cs.waikato.ac.nz/~ml/weka/arff.html (accessed in 02/08/2011);

[61] - Kotsiantis, S. B., Supervised machine learning: A review of classification algorithms, Informatica, 2007, 31:249-268;

[62] - Breiman, L., Friedman, J., Olshen, R., Stone, C., Classification and Regression Trees, Wadsworth, Belmont, 1984;

[63] - Hartigan, J., Wang, M., A K-means clustering algorithm, Applied Statistics, 28, 100–108, 1979;

[64] - Lloyd, S., Least squares quantization in pcm, Bell Telephone Laboratories Paper, Marray Hill, 1957;

[65] - MacQueen, J., Some methods for classification and analysis of multivariate observations, Proc. 5th Berkeley Symposium, 281–297, 1967;

[66] - Ding, Chris, He, Xiaofeng, K-means clustering via principal component analysis, Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley;

[67] - Wilkinsom, Ian B., el al., The influence of heart rate on augmentation index and central arterial pressure in humans, Journal of physiology (2000), 525.1:263-270;