Biosignal analysis for cardiac arrhythmia detection using non-supervised techniques

José Luis Rodríguez Sotelo

Universidad Nacional de Colombia
Faculty of Engineering and Architecture
Department of Electrical, Electronics and Computer Engineering
Manizales

2010

Biosignal analysis for cardiac arrhythmia detection using non-supervised techniques

José Luis Rodríguez Sotelo
jlrodriguezso@unal.edu.co

Thesis document for partial fulfilment of the requirements for the
Doctorate degree in Engineering Sciences

Advisor
Prof. César Germán Castellanos Domínguez

Universidad Nacional de Colombia
Faculty of Engineering and Architecture
Department of Electrical, Electronics and Computing Engineering
Manizales
2010

Análisis de bioseñales en la identificación de arritmias cardíacas mediante técnicas no supervisadas

José Luis Rodríguez Sotelo

Trabajo de grado para optar al título de
Doctor en ingeniería
Línea de investigación en Automática

Director
Prof. César Germán Castellanos Domínguez

Universidad Nacional de Colombia
Facultad de Ingeniería y Arquitectura
Departamento de Ingeniería Eléctrica, Electrónica y Computación
Manizales
2010

To my parents and my sisters who have trusted in me...

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgments

First, I thank God for giving me the opportunity to develop this work successfully. You have enlightened and guided me in the right way. Thanks for your support when it was most required.

I would like to gratefully acknowledge the supervision of my advisor, Ph.D. Germán Castellanos for his continuous support in the Ph.D. program. He is responsible for involving me in an interesting research career these last 8 years. He showed me the main way to approach a research problem and the need for persistence when accomplishing any goal.

A special thanks goes to my co-advisor, Ph.D. David Cuesta Frau, who is with the *Universidad Politécnica de Valencia* (UPV), Spain. David's work represents the origin and the motivation of this thesis whose main objective is the development of non-supervised techniques applied to bio-signals. He gave me the opportunity to serve as Research Fellow at the GIICA group of UPV, spain, between 2008 and 2009.

Besides my advisors, I would like to thank the rest of my thesis committee: Ph.D. Edilson Delgado Trejos, who besides his valuable comments, gave important information from his Doctoral thesis, specifically in the chapters of physiology and assisted diagnosis of cardiac pathologies, and Ph.D. Mauricio Orozco who gave insightful comments and reviewed my work carefully.

I am also greatly indebted to Diego Peluffo, who, more than a great Ph.D. student that finished his Master of Science as a fundamental part of this work, is an unconditional great friend. Thank you, Diego!.

Let me also say "Thank You" to the following wonderful people at Research Group of Signal Processing and Recognition (Universidad Nacional de Colombia): Luis E. Avendaño, Julio García, Carlos D. Acosta, Franklin Sepúlveda, Genaro Daza, Luis Sánchez, Jorge A. Jaramillo, David Avendaño, Andrés Quiceno and all other people who contributed in every other way to this work.

I thank Ph.D. Pablo Laguna, Ph.D. Lior Wolf and Ph.D. Pierre Hansen, for their academic contributions.

During the course of this work, I was supported in part by the DIMA at the *Universidad Nacional de Colombia*, from *Apoyo a Tesis de Posgrado* project cod: 10242. I also was fully supported by COLCIENCIAS from *Doctorados Nacionales 2006* program between years 2006-2009.

I am grateful to the *Universidad Autónoma de Manizales*, specially to the Electronic and Automation Department and my friend, Ph.D. Carlos Murillo, who gave me the confidence to finish the doctoral program. Thank you to welcoming me in your institution.

Finally, but not least, I thank my family: my parents, Luz María y Carlos for unconditional support and encouragement to pursue my interests, even when the interests went beyond boundaries of engineering, research and of course, arts. My sisters, Luz Dary and Clara for believing in me.

José Luis Rodríguez Sotelo
July 2010

Notation

$s(t)$	signal in continuous time
T_s	sampling period
F_s	sampling frequency
$s[kT_s]$	signal in discrete time
$\boldsymbol{d}_j$	j-th ECG complex
n	number of observations (heartbeats)
p	number of variables (features)
$x_1, \ldots, x_p$	variables
$\boldsymbol{x} = (x_1, \ldots, x_p)^\mathsf{T}$	measurement vector
$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\mathsf{T}$	$n \times p$ matrix
$\boldsymbol{X} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix}$	
k	number of clusters
$\mathbb{R}^n$	n-dimensional Euclidean space
$\mathrm{tr}(\boldsymbol{X})$	trace of $\boldsymbol{X}$
$\mathrm{diag}(\boldsymbol{X})$	diagonal of $\boldsymbol{X}$
$\boldsymbol{X}^\mathsf{T}$	transpose of $\boldsymbol{X}$
DTW ($\mathrm{dtw}(\cdot, \cdot)$)	Dynamic Time Warpíng
$\langle \cdot, \cdot \rangle$	Inner product
$\langle \cdot, \cdot \rangle_{\boldsymbol{A}}$	M-inner product regarding matrix $\boldsymbol{A}$
$\mathbf{E}\{\cdot\}$	expectance operator
$\mathrm{times}(\boldsymbol{A}, \boldsymbol{B})$	array-wise multiplication

Acronyms

AAMI	Association for the Advancement of Medical Instrumentation
CBC	Center-Based Clustering
CWT	Continuous Wavelet Transform
DWT	Discrete Wavelet Transform
ECG	Electrocardiographic Signal
FC	Fuzzy Clustering
GA	Genetic Algorithms
GEMC	Gaussian Expectation Maxmimization-based Clustering
GIC	General Iterative Model
HMM	Hidden Markov Model
HRV	Heart Rate Variability
ICA	Independent Component Analysis
KLT	Karhunen Loéve Transform
LMS	Least Mean Square
MIT/BIH	Massachusetts Institute of Technology/Beth Israel Hospital
MLP	Multilayer Perceptron
MRA	Multiresolution Analysis
MSE	Mean Square Error
MSSC	Minimum Sum of Squares-based Clustering
NN	Neural Networks
NPCA	Nonlinear Principal Component Analysis
PCA	Principal Component Analysis
PNN	Probabilistic Neural Network
PRD	Percentage Root Difference
RLS	Recursive Least Square
SOM	Self-Organizing Maps
WPCA	Weighted Principal Componente Analysis
WT	Wavelet Transform

Abstract

A methodology for unsupervised Holter monitoring of cardiac arrhythmias is proposed based on variable–wise relevance analysis and partitional soft-clustering. Because of strong asymmetry among class observations the heartbeat–derived features are properly selected by their proper weighted linear projection based methods. To estimate the feature weights, two different distance measures are considered: Mean Square Error and M–inner product.

In the non-supervised classification stage, M–inner product approach can be also used as the initialization method of clustering stage offering the estimated number of groups of the partition. For clustering, it demonstrates that center-based clustering with an appropriate initialization can offer good performance from the point of view of cluster separability.

Additionally, in order to reduce computational cost, it is proposed to carry out a segment analysis by successive divisions along time, where each division is sequentially processed, and thus processing time is significantly reduced. Also, some appropriate supervised and non-supervised performance measures based on groups and spectral analysis are developed, which relate the clustering performance with the number of resultant groups and computational cost.

The experiments are done with a standard arrhythmia database of MIT/BIH and taking into account the AAMI standard (Association for the Advance of Medical Instrumentation).

Methodology shows comparable performance respect to others referenced studies, based on either supervised or unsupervised training.

Resumen

Se propone una metodología para análisis no supervisado de arritmias cardíacas de registros Holter basada en análisis de relevancia basado en variables y agrupamiento-suave particional. Debido a la fuerte asimetría entre clases, las características que representan a los latidos son seleccionadas apropiadamente empleando métodos de proyección lineal ponderada. Para estimar los pesos de las características se consideran dos medidas de distancia: El error cuadrático medio y el producto interno–M.

En la etapa de clasificación no supervisada, el enfoque del producto interno–M también puede ser usado como método de inicialización de la etapa de *clustering*, estimando el número de grupos de la partición. Para la etapa de agrupamiento, se demuestra que el método basado en centros, con una inicialización apropiada puede brindar buen desempeño desde el punto de vista de la separabilidad de grupos.

Adicionalmente, con del fin de reducir el costo computacional, se propone el análisis por segmentos, que se lleva a cabo dividiendo a los registros sucesivamente en el tiempo y procesando cada división de manera secuencial. También, se desarrollan medidas de desempeño supervisadas y no supervisadas basadas en análisis de grupos y análisis espectral, las cuales relacionan el desempeño de la partición con el número de grupos resultantes y el costo computacional.

Los experimentos se llevan a cabo usando la base de datos estándar de arritmias de la MIT/BIH y teniendo en cuenta el estándar de la AAMI (*Association for the Advance of Medical Instrumentation*).

La metodología muestra un desempeño comparable con respecto a otros trabajos de la literatura basados en análisis supervisado y no supervisado.

Part I

Preliminaries

Chapter 1

Introduction

The development of bio-signal analysis systems has become a major investigative field, due to technological progress in signal processing systems, and the large number of alternative solutions to a specific problem.

Electrocardiography is amongst the most studied type of bio-signals, since several decades of Electrocardiographic (ECG) signal research has made this basic discipline a tool for the diagnosis of cardiac disorders. Because of its simplicity, low cost and a non-invasive nature it, is still widely used despite the appearance of newer techniques.

This thesis covers the problem of long-term recording analysis corresponding to ECG signals of Holter records. The motivation for studying this issue focuses on the development of methods for cardiac arrhythmia analysis to identify particular events that occur at specific periods in time.

These events are associated with cardiac disorders that may become potentially harmful to the patient. The developed methods are aimed at further development of specialized equipment that provide clinical monitoring for both the patient and the specialist, and that support the diagnosis in real time, improving mortality rates for heart problems specially for people in rural areas, to improve their access to this type of procedure that is currently not widely accesible.

The context analysis of such signals involves two major aspects that are studied in this work. The first one corresponds to the large amount of data stored in the records, reaching up to 100.000 heartbeats for its evaluation, which becomes a hard task for the specialist who evaluates the information and decides what heartbeats are important for a determined analysis. There are cases where only a few beats of the recording become essential in the diagnosis of some pathology or prevention of deadly diseases. Therefore, a detailed analysis of the entire record is needed.

The second aspect corresponds to the intrinsic characteristics of the signal, such as heart rate variability, morphological variety, among others, that could be the result of driving problems in the cardiac system or may be a patient's physical characteristics.

The electrical nature of ECG signals and its transmission toward acquisition devices, increases the noise sensitivity, which can completely alter the diagnostic information contained in the signal, changing the training processes in the identification of cardiac pathologies.

Consequently, both aspects have been strongly considered in the automatic ECG processing and analysis procedures to detect, classify, and cluster heartbeats. Thus, several methods have been reported in the scientific literature to carry out those classification–related tasks, using either supervised [4–7] or unsupervised [8–10] approaches. However, due to a large variability in ECG heartbeat morphology, the former methods tuned in a specific ECG dataset may have decreased performances in other datasets. In addition, these techniques require a considerable amount of known and labeled heartbeats which is not feasible when having long–term ECG monitoring. Regarding unsupervised methods, even their performance usually does not over-perform supervised training, yet former methods can be applied to a broader set of ECG recordings as they can dynamically adapt to new signal features. Moreover, additional factors must be taken into account, such as highly unbalanced classes, uncertainty of the number of classes, signal variability, artifacts, etc. Unsupervised analysis is preferred for Holter monitoring.

However, there are still some open issues when implementing unsupervised analyses, such as, computational cost, unbalanced clusters, unknown number of clusters and initial partition. In this thesis each of these problems is covered through unsupervised analysis tools in the selection/extraction features and partitional clustering areas, ending up in an unsupervised analysis methodology that can be implemented in oriented devices for analysis in real time. The proposed methodology, without requiring prior training or labeling of the samples by the specialist, can be applied to ECG signals that have great variability in time and morphology, identifying the main arrhythmias set by the AAMI standard.

1.1 General Objective

To design a non-supervised methodology for analysing ECG signals of Holter recordings including preprocessing, feature estimation, relevance analysis and clustering stages, in order to identify cardiac arrhythmias. Such methodology must be able to separate the main arrhythmia groups according to ANSI/AAMI EC57:1998 standards. Also, thinking about real time applications, a good trade-off between computational cost and performance must be guaranteed.

1.2 Specific Objectives

- To analyze and select methods for preprocessing and feature estimation taking into account QRS estimation, amplitude and time variability of ECG signals.

- To develop a non-supervised feature selection method that takes into account the feature variability and improves further classification tasks.

- To develop a clustering scheme for heartbeats in ECG signals guaranteing a good performance and low computational cost.

- To design a scheme to assess the methodology performance with respect to partition quality.

1.3 Outline of the Manuscript

The present thesis is divided into 4 parts:

- Part I corresponds to preliminaries, comprising Chapters 2 and 3. Chapter 2 presents introduction and objectives for the thesis. Chapter 3, presents a physiological background as well as a state of the art of ECG signal analysis focused on non-supervised classification.

- Part II is the theoretical background of the algorithms and methods developed in this study. First, in Chapter 5 the signal preprocessing and feature estimation are described. Chapter 6 establishes the algorithms and applications of ECG signals for relevance analysis methods. Chapter 7 describes the methods and tools used for heartbeat clustering, including all stages from estimation of number of groups to segment clustering.

- Part III comprises Chapters 8 and 9 that correspond to experimental setup and results and discussion, respectively. Chapter 8 shows the procedures to adjust parameters as well as variants of the methods described in Part II. In Chapter 9 all results that evaluate the performance system in each stage are shown. Also, a discussion is presented of the results in terms of performance measure and computational cost in comparison with another studies in the literature.

– Final conclusions of this study are presented in Part IV.

Chapter 2

Physiological Preliminaries

For the context of this work, the fundamental concepts regarding electrocardiographic (ECG) signals, its origin and the main types of ECG recordings, are presented [11]. Likewise, some types of cardiac arrhythmias, such as auricular and ventricular ectopic heartbeats, branch blocks, fusion beats, among others taking into account the standard of the Association for the Advancement of Medical Instrumentation (AAMI), are discussed in this chapter.

2.1 Electrocardiographic Signals

Electrocardiography is a discipline of medicine that is concerned with electrical disorder on heart and employs a electrocardiographic (ECG) signal that describes the behavior of electrical system on heart. In this subject, signals are captured by means of surface electrodes, being one of the most common medical tests for the exploration of cardiac activity to diagnose several kind of arrhythmias, conduction defects, heart attack, hypertrophy and other anomalies [12].

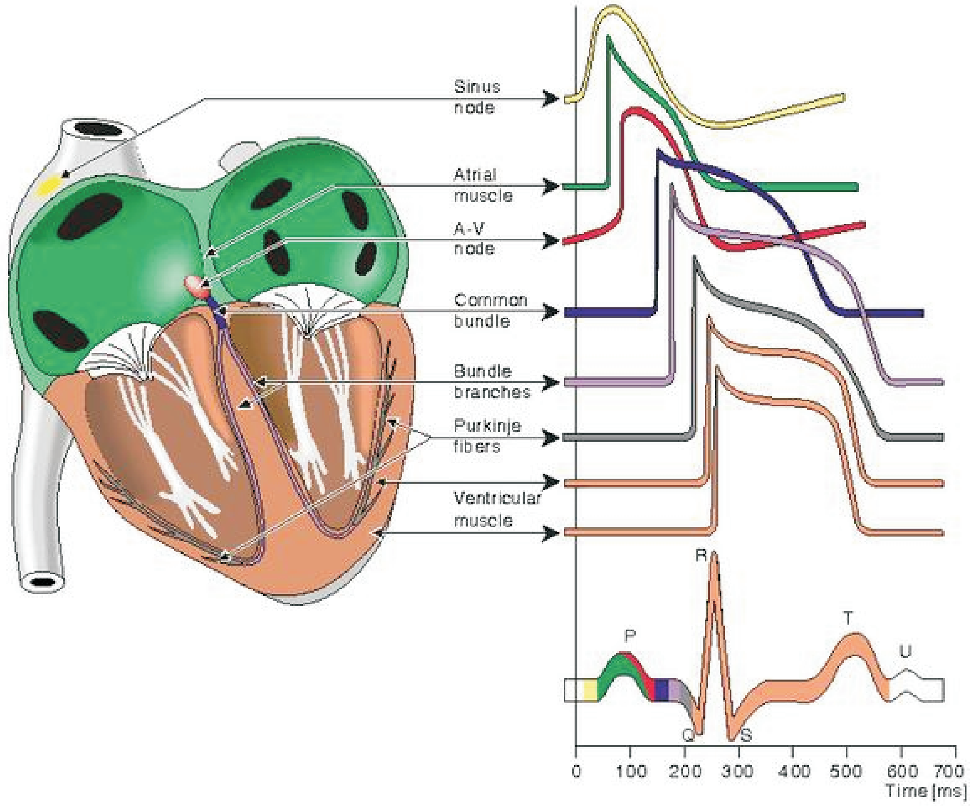
Myocardial electrical activity is stored into a electrical recording (ECG), as shown in Figure 2.1(a). A normal ECG signal follows a similar pattern to that shown in Figure 5.7(b). From the inside of the ventricular wall, different activation sites cause the formation of a front wave, which spreads through the ventricular mass to the outer wall of the heart muscle, so that the activation is performed cell by cell. After each ventricular region is depolarized, repolarization occurs, which is not a phenomenon of propagation. Because the duration of the action of the impulse on the epicardium (the outer part of the myocardium) is shorter than the action on the endocardium (the inner part of the myocardium), electrical activity shows to be spread from the epicardium towards the endocardium [13].

The electrical events involved in the dynamic cardiac are summarized in Table 2.1.

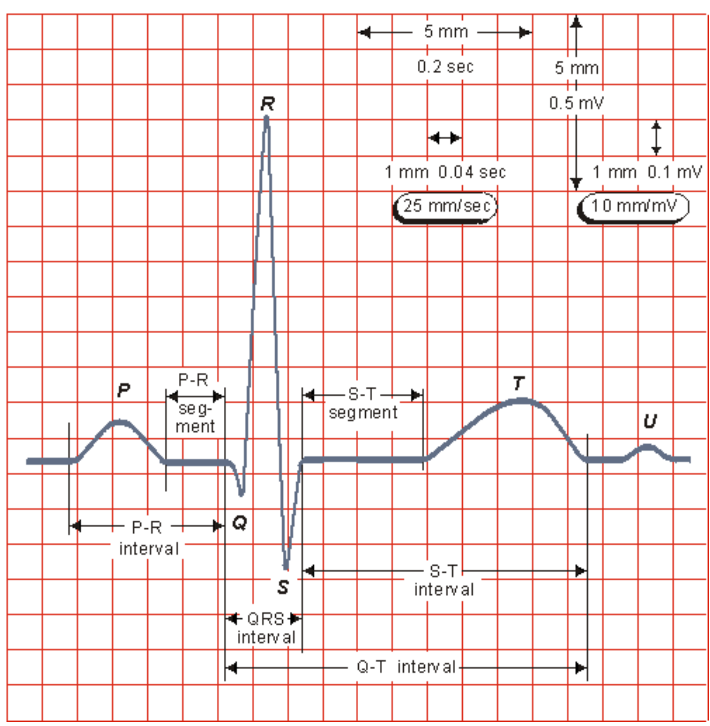
Table 2.1: Electrical events in heart tissue. Terminology in interpretation of ECG[a]. Conduction rate[b]

Location	*Event*	*Time* [ms]	*ECG*[a]	*Rate* [m/s][b]	*Frequency* [$1/min$]
Sinoatrial Node	Orig. of impulse	0		0.05	$70-80$
Right Auricula	Depolarization	5	P	$0.8-1.0$	
Left Auricula	Depolarization	85	P	$0.8-1.0$	
AV node	Impulse arrival	50	PQ	$0.02-0.05$	
AV node	Impulse output	125	PQ		
His bundle	Activated	130		$1.0-1.5$	$20-40$
Branch of bundle	Activated	145		$1.0-1.5$	$20-40$
Purkinje fibers	Activated	150		$3.0-3.5$	$20-40$
Endocardium					
Septum	Depolarization	175	QRS	0.3 (axial)	$20-40$
Left ventricle	Depolarization	190	QRS	–	$20-40$
Epicardium					
Left ventricle	Depolarization	225	QRS	0.8	$20-40$
Right ventricle	Depolarization	250	QRS	(transv.)	$20-40$
Epicardium					
Left ventricle	Repolarization	400			
Right ventricle	Repolarization				
Endocardium					
Left ventricle	Repolarization	600	T	0.5	

A normal ECG pattern consists of P wave, QRS complex, and T wave. The QRS complex, in turn, includes three separate waves: Q, R and S, all these are generated when the cardiac impulse goes through the ventricles. In the ECG signal, waves Q and S are, generally, significantly less prominent than the wavelength of P and sometimes may be missing. The P wave depends on electrical currents generated when the atria depolarize before contraction, and the QRS complex is produced by currents arising when the ventricles depolarize prior to contract. Therefore, P wave as well as the components of the QRS complex correspond to depolarization waves. The T wave, which is caused by currents arising when the ventricles recover from the depolarization state, is known as the repolarization wave. In essence, the ECG is composed of waves



(a) The origin of the electrical activity [11]



(b) Pattern of the normal ECG signal [11].

Figure 2.1: Electrophysiology of the heart

of depolarization and repolarization (see Figure 2.1(a)). *QRS* complexes are present in most of heartbeats that are associated with ventricular electrical activity and contain important clinical information, so its signal to noise ratio is the highest among all waves present in the ECG signal [13].

2.2 Ambulatory Electrocardiography

During the last two decades, acquisition systems for physiological signals have been developed and improved and they are then lighter, smaller and capable of recording multiple signals up to 48 hours. These systems, also called ambulatory record systems, are used in ECG analysis to detect infrequent arrhythmias or transient abnormalities in heart function often related to the stresses of everyday life, such as transient ischemic events or silent myocardial ischemia. Some types of diseases can not be detected in short-time ECG or 12 leads ECG recordings. These devices are called Holter recorders, named after an ECG recorder developed by N. J. Holter in 1961 [14]. The first Holter recorder used a tape to store large signal files, since the capacity of solid state memory were not available at that time. Now, signals are recorded in flash-type semiconductor memories, which can be transferred to a workstation for further analysis [14].

The increase of the health costs makes an urgent need to develop ambulatory systems that reduce the permanence of patients in hospitals. Therefore the design of a portable, low cost, high performance and simple system that allows automated analysis and diagnosis is a necessity. Such equipment must integrate various data analysis techniques such as: signal processing, pattern recognition, decision making and human-machine interaction. The existing portable devices have been reduced, for technological reasons, to record the signal over a period of time, which is constrained by the storage capacity of the devices. A typical signal of 24 hours consists of approximately 100.000 heartbeats which can be grouped morphologically into a much smaller number of classes. In most of classes where the heartbeat has a typical pattern, it is enough to know the number of heartbeats and a representative template of the morphology for grouping. In the lapse of time where cardiac activity presents anomalies or symptoms of illness, it is necessary the complete record of the signal. This is possible if the portable device for analysis in addition to record the signal, is also capable of testing it.

2.3 Cardiac Arrhythmias

The pathologies observed using the ECG are divided into three categories:

1. Heart rhythm disturbances, or arrhythmias.
2. Dysfunctions of blood perfusion in the myocardium or cardiac ischemia.
3. Chronic disorders of the mechanical structure of the heart, such as left ventricular hypertrophy.

This work is focused on the study of the first type of pathologies described above. In particular, the experiments are developed over the entire QRS complexes that are associated with ventricular electrical activity. They contain clinic important information, for example its morphology presents significant changes in abnormal ventricular heartbeats. QRS complex is also present in most of the heartbeats and its signal to noise ratio is the highest among all waves present in the signal.

The main problem of ECG analysis is the wide variability into signal morphology, not only among patients but also due to patient movements, changes in the electrical conduction, characteristics of the body, among others. Because of this, it is not possible to form a training set that takes into account all cases of interest. In addition, the ECG signal is contaminated by several noise sources, both external sources (interference of the power line, movement of the electrodes) and biological sources (muscle movement that causes high-frequency interference and the breathing that causes displacement of the baseline). Then, this kind of analysis requires special care to choose appropriate techniques for signal conditioning (pre-processing), since the quality of input signal for the further classifier has a direct impact on the performance of it.

2.3.1 Not imminently life-threatening cardiac arrhythmias

Broadly speaking, arrhythmias can be divided into two groups. The first group includes ventricular fibrillation and tachycardia which are life-threatening and require immediate therapy with a defibrillator. Detection of these arrhythmias is well researched and successful detectors have been developed with high sensitivity and specificity. This study analyzes the second group, which includes arrhythmias that are not imminently life-threatening but may require therapy to prevent further problems.

In accordance to the AAMI standard (ANSI/AAMI EC57:1998/(R)2003) [4], the following arrhythmia groups shown in Table 2.2 are of interest to be examined: normal–labeled heartbeat recordings (termed N), Supraventricular ectopic beat (Sv), Ventricular ectopic beat (V), Fusion beat (F), as well as unknown beat class (Q) is taken into consideration. All classes listed are assumed to be present during Holter analysis.

The MIT/BIH arrhythmia database [15] is among the most representative database to evaluate the design of algorithms related to the analysis of cardiac arrythmias. The database contains several types of beats within each group of arrhythmias recommended by the AAMI, for example, in the Normal group can be found LBBB arrhythmia type (Left bundle Branch Block), RBBB (Right Bundle Branch Block), AE (Atrial Escape) and NE (junctional Nodal Escape). The used arrhythmias classification can be seen in Table 2.2.

Table 2.2: Set of analyzed arrhythmias according to the AAMI standard.

AAMI heartbeat	Description	MIT/BIH heartbeat types
N	Any beat not in the Sv,V,F or Q classes	Normal (N), Left Bundle Branch Block ($LBBB$), Right Bundle Branch Block ($RBBB$), Atrial Escape (AE), Nodal (junctional) escape beat(NE)
Sv	Supraventricular ectopic beat	Atrial Premature (AP), Aberrated Atrial Premature (aAP), Nodal (junctional) Premature (NP), Supraventricular Premature (SP),
V	Ventricular ectopic beat	Premature Ventricular Contraction (PVC), Ventricular escape (VE)
F	Fusion beat	Fusion of ventricular and normal (fVN), Fusion of paced and normal beat (fPN)
Q	Unknown beat	Paced (P), Unclassified (Q)

2.3.2 Group of arrhythmias *N*

It corresponds to any beat that does not belonge to Sv, V, F or Q classes (Table 2.2), as is shown in Figure 2.2.

Within this kind of heartbeats are normal beats (N), bundle branch block (LBBB and RBBB), and escape and atrial nodal beats (AE and NE).

Bundle Branch Block (BBB) is a disorder in the conduction of electrical impulse in the ventricles [16]. The electrical impulse conduction to the ventricles is via the His bundle and its divisions: right and left bundle branch. When one of these branches is altered, the electrical impulse spreads through the ventricular muscle itself rather than by the Purkinje system. This reduce the conduction velocity, so if there is blockage in one of the branches occurs a prolongation of the QRS complex because it lasts as long as depolarization spreads through the ventricles [17]. Branch blocks also generate morphological changes (R-prime) in the QRS complex.

In the left bundle branch block (LBBB), cardiac depolarization is propagated by the right ventricle much faster than the left ventricle. Therefore, the left ventricle remains polarized longer than the right one. This is reflected in an extension and a morphological change (RR') of the QRS in left precordial leads (V5 and V6). Moreover, in the right bundle branch block (RBBB), the impulse conduction through the right ventricle is delayed regarding the left one, in this way, the QRS is prolonged and generates a morphology named rsR in the right precordial leads (V1 and V2).

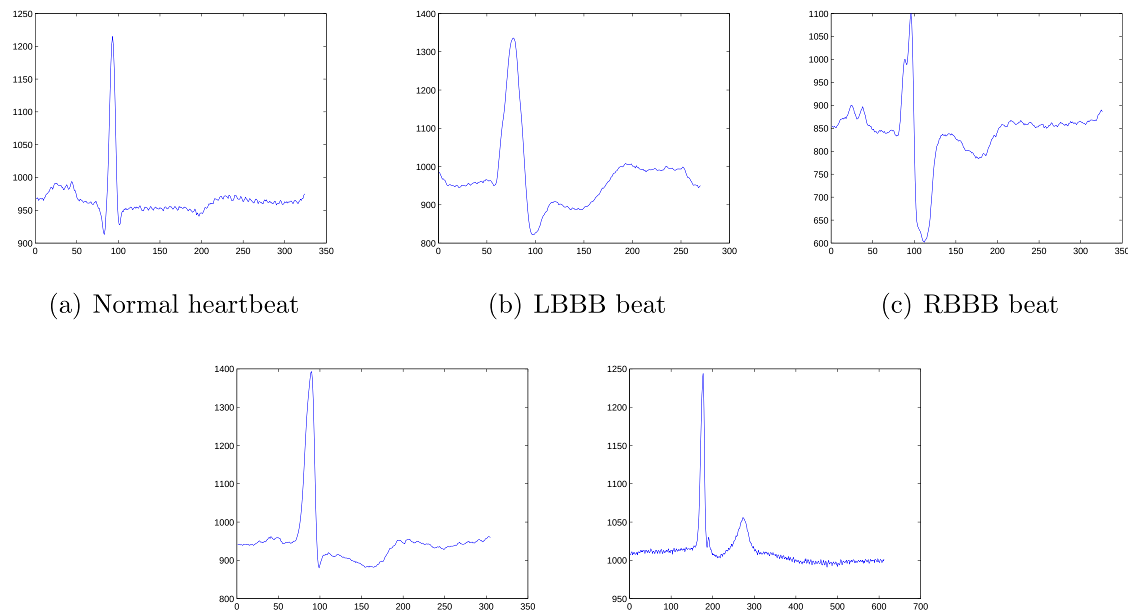
The presence of BBB, does not necessarily mean heart disease, it can occur in healthy patients, which has a good prognosis and do not progress to higher degree block [18]. However, in some studies [19–21] was found that the presence of right bundle branch block is correlated with arterial hypertension, heart failure, coronary disease, pulmonary embolism, and increased mortality and the emergence of left bundle branch block increases risk of coronary heart disease, mortality and ventricular myocardial infarction [22], [23]. Thus, can be seen that is necessary to detect such arrhythmias because of the prognostic value they have.

The escape heartbeats are characterized by appearing eventually interrupting the pace of rate base. The most common are those who are ahead on that cadence or extrasistoles and those who move away or escape heartbeats. Depending on the morphology of the waves its origin can be known (atrial, nodal or ventricular) and the type of existing AV conduction.

2.3.3 Group of arrhythmias type *Sv*

It corresponds both to atrial and supraventricular premature beats and its variants. An example is illustrated in Figure 2.3.

An Atrial Premature Beat (APB), also called Atrial Ectopic Beat (AEB) or Premature Atrial Contraction (PAC) is an extra heartbeat caused by electrical activation of the atrium from an abnormal site before a normal heartbeat can occur. Generally, APB's occur in healthy people that rarely have symptoms. It is common among people who have lung problems and more common among adult people than young people. Recent



(a) Normal heartbeat (b) LBBB beat (c) RBBB beat

(d) Atrial Escape beat (e) Nodal escape beat

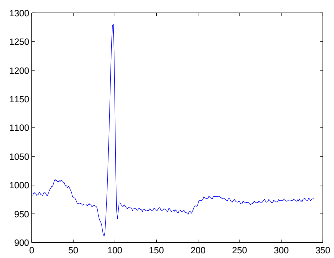
Figure 2.2: Heartbeats of N group, extracted from MIT/BIH database.

studies on risk factors for stroke have shown that frequent APB heartbeats are an independent risk factor for having a stroke [10].

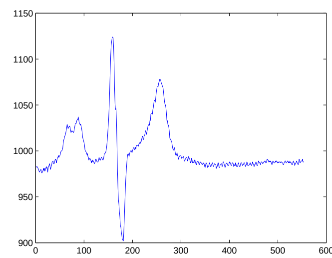
Although often the APB's, are considered a benign phenomenon, it has been shown in clinical practice that frequent APB's could be an early symptom of heart failure and may precede atrial fibrillation.

Frequent APB's can be an indicator for other risk factors, such as severe hypertension, asymptomatic atherosclerosis, structural abnormalities that cause stroke, calcified mitral valve or enlargement of the left atrium, which could cause an increase in the formation of thromboembolism [24].

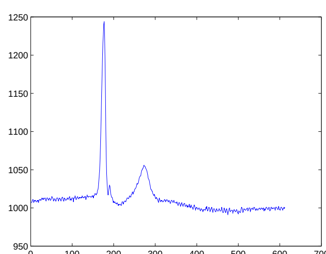
Usually, experts analyze the Holter recordings for the detection of APB beats due to their frequency and found that their detection is complicated in the sense of the nature of the APB beats, because they exhibit similar morphological characteristics to a normal heartbeat specifically in ventricular depolarization and repolarization, that is to say similar morphology between the QRS complex and T wave, with respect to a normal heartbeat which hold the majority in the record, using atrial depolarization for its detection, that is to say the PR interval as the P wave presents morphological variations with respect to the common P-wave record, however there may be beating that does not contain P wave, due to overlap with the previous T wave beat resulting in a slight increase in amplitude of the latter. Another more effective technique of APB heartbeat detection is by analysis of heart rate variability (HRV) as from the physiological point of view, before completion of ventricular repolarization, there is a premature excitement in the atrial area other than the sinus node causing the appearance of a premature beat. If the premature excitement is sudden there will be a delay in the activation of the sinus node for the next cardiac cycle, represented in a pattern of increase and consecutive decrease heart rate. The drawback with this technique is that if occur continuous premature excitations or continuous premature beats, this pattern disappears and causes increased heart rate in some cases interpreted as normal pace due to the shooting conditions of the recording, reducing possibility of success in the detection of APB beats through the HRV.



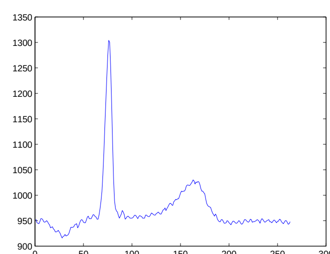
(a) Atrial Premature Beat (APB)



(b) Aberrated APB



(c) Nodal Premature Beat



(d) Supraventricular Premature Beat

Figure 2.3: Heartbeats of Sv group, extracted from MIT/BIH database.

2.3.4 Group of arrhythmias type *V*

A ventricular premature beat (ventricular ectopic beat, premature ventricular contraction) is an extra heartbeat resulting from abnormal electrical activation originated in the ventricles before a normal heartbeat would occur, as can be illustrated in Figure

2.4.

The main symptom is a perception of a skipped heartbeat. ECG is used to make the diagnosis. Avoiding things that trigger these beats, such as stress, caffeine, and alcohol, is usually sufficient treatment.

Ventricular premature beats are common, particularly among older people. This arrhythmia may be caused by physical or emotional stress, intake of caffeine (in beverages and foods) or alcohol, or use of cold or hay fever remedies containing drugs that stimulate the heart, such as pseudoephedrine. Other causes include coronary artery disease (especially during or shortly after a heart attack) and disorders that cause ventricles to enlarge, such as heart failure and heart valve disorders.

The VE beats are hardly found in ECG of 12-leads, using Holter recordings for its detection [25]. The VE's can be identified with the following criteria of structural abnormality in the ECG [26, 27]:

- The QRS duration is higher than the average dominant QRS due to abnormal activation of the ventricle which is carried out through intra-myocardial functional pathways.

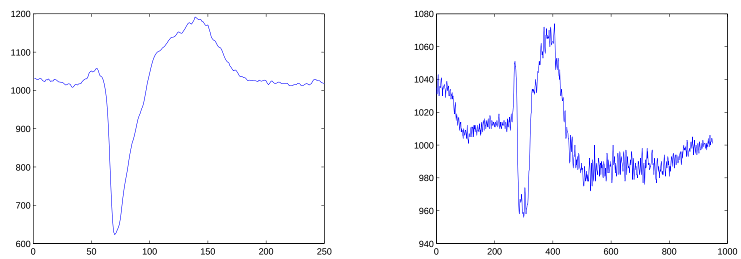
- Different morphologies in the QRS complexes are present. Preceding P waves that do not occur prematurely. The T wave is often found in the opposite direction of the R wave. If the heartbeats originate from a single focus, all the VPC have the same morphology, although different from the normal morphology.

- Usually, the ventricular extrasystoles are premature, the RR intervals are shorter than the average RR and usually it can be found a complete compensatory pause in the heartbeat.

- The VE's originated from the left ventricle normally produce heartbeat patterns of right bundle branch block (RBBB) and the VE's originated from the right ventricle normally produce heartbeat patterns associated with left bundle branch block (LBBB).

A ventricular escape beat is a self-generated electrical discharge initiated by the ventricles and causing their contraction; normally the heart rhythm begins in the atria of the heart and is subsequently transmitted to the ventricles. The ventricular escape beat follows a long pause in ventricular rhythm and acts to prevent cardiac arrest. It indicates a failure of the electrical conduction system of the heart to stimulate the ventricles (which would lead to the absence of heartbeats, unless ventricular escape beats occur).

Ventricular escape beats occur when the rate of electrical discharge reaching the ventricles (normally initiated by the heart's sinoatrial node (SA node), transmitted to the atrioventricular node (AV node), and then further transmitted to the ventricles) falls below the base rate determined by the ventricular pacemaker cells. An escape beat usually occurs around 2–3 s after an electrical impulse has failed to reach the ventricles.



(a) Premature Ventricular Contraction (PVC)

(b) Ventricular escape beat

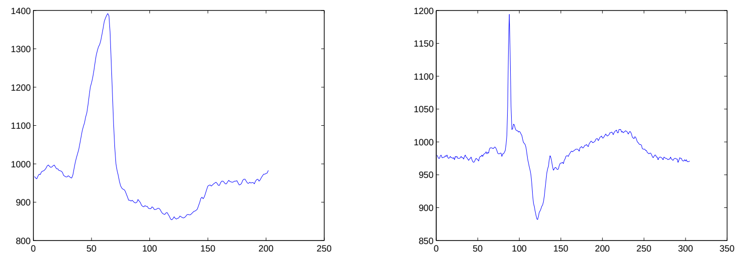
Figure 2.4: Heartbeats of V group, extracted from MIT/BIH database.

2.3.5 Group of arrhythmias type F

Fusion heartbeat occur when either the atria or the ventricles are activated by two simultaneously invading impulses and can be assessed in the P wave or the QRS complex of the ECG. An atrial fusion beat results when the sinus beat coincides with an atrial ectopic beat, when two atrial ectopic beats coincide, or when an atrial or sinus beat coincide with retrograde conduction from a junctional focus. A ventricular fusion beat results when a ventricular beat coincides with a sinus beat, a ventricular ectopic beat, or a junctional beat. A couple of examples are shown in Figure 2.5.

2.3.6 Group of arrhythmias type Q

Unclassified heartbeats (heartbeats Q) correspond to heartbeats that do not contain relevant medical information, mainly due to some external conditions as artifacts, electrode disconnection, saturation of acquisition system, or heartbeats originated by pacemakers. It is necessary to isolate this kind of heartbeats from the training space in

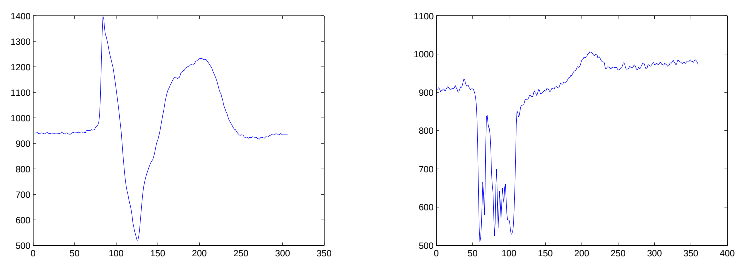


(a) Fusion of ventricular and normal beat

(b) Fusion of paced and normal beat

Figure 2.5: Heartbeats of F group, extracted from MIT/BIH database.

order to obtain an adequate diagnosis. Normally, due to a small quantity of this type of heartbeats and their low importance in the diagnosis, are considered as outliers. In Figure 2.6 two types of Q heartbeats are shown, corresponding to Paced beat and Unclassified beat.



(a) Paced beat

(b) Unclassified beats

Figure 2.6: Heartbeats of Q group, extracted from MIT/BIH database.

Chapter 3

Assisted Diagnosis of Cardiac Pathologies

In this chapter, the development of assisted diagnosis that is related to the following key questions: *How to interpret the current records in time according to the possible diagnostic decisions?*, *What are the essential characteristics?, and finally, how to extract the necessary information that is embedded in the obtained signals?* is discussed. These questions are typically related to automation systems, particularly in the training of intelligent systems oriented to pattern recognition. Also, both the requirements and the state of the art of computer-assisted diagnosis systems are presented.

3.1 Computer-Assisted Diagnosis and its Applications

Doctors ought to identify the symptoms and signs that are related with physiological activity that serve to determine the normality or abnormality states, associated with potential diseases. The identification is made through sensory perception, sometimes with the mediation of instruments or devices that magnify basic functional signals or facilitate judgments about the degree of normality or abnormality of different functional states of the body. This model has several limitations: high subjectivity when is interpreted by a medical doctor, lack of storage and replication capability for future analysis, if a more precise advice on diagnosis or treatment decisions is required, finally, the bias and other human mistakes, despite the skills generated by the medical training and human resource training. According to this, it is highly important to develop supporting equipment and technology to facilitate the exploration and auscultation to objectively improve the quality of a clinical decision, or in other cases, to provide sufficient information to the specialist in order to support an adequate prognosis.

The advent of new electronic technologies, including medical instrumentation, digital signal processing and imaging, has shown that it can improve the quality of the collection, analysis, precision and sensitivity in the signs issued by the human body system. In this regard, particular interest has arisen around the non-invasive medical diagnosis, because, electronic records of biological signals, obtained on the surface of the body, reflect the internal behavior of the body or some of its parts, and may be adequate to provide essential clinical information without resorting to invasive. However, the model of medical diagnosis requires objective methods related to process models, measured in terms of sensitivity and specificity, ensuring that technological aids allow to facilitate the capture, analysis and the proper storage of the different signals recorded, so allowing the subsequent interpretation of the information by other observers.

Today, advances related to signal analysis and automation systems, make the diagnostic support more interesting in the use of automated processing and identification of functional states of the body, providing diagnostic support to the specialized medical personal, that interprets and makes final decisions about the diagnosis. As a result, assisted diagnosis, in which a digital processing system provides more evidence and information to a specialist, helps the quality of the clinical verdict.

Since the 80s it was intended to build expert systems in the medical field under exploratory conditions with little or none clinical expertise as one of the most important applications of artificial intelligence [29]. The advantages related to the high values of speed, accuracy and memory for data storage systems that offer modern digital processors, allow the development of information processing systems on a vast amount of medical data for diagnosis. This will increase productivity and efficiency of diagnosis, both therapeutic and preventive actions economically justified by the use of computers to solve medical problems. However, it should be emphasized that the assisted diagnosis is always taken as consultative means. A computer can not replace the medical verdict, however, may suggest a series of decisions with a certain level of reliability, so that they are finally accepted or rejected by the doctor. The specialist uses the response of the computer as a second opinion, but it is the doctor who makes the final decision.

The use of digital devices has become an effective alternative in the process of finding an accurate diagnosis, which has gradually ceased to be an individual decision.

For many years the diagnosis of complex diseases has been taking place not by a doctor but by a group of doctors, each one specialized in a field. In recent decades, more and more records of tests have involved complex diagnoses and are decrypted by a group of specially trained doctors (clinically specialized). In this sense, in addition to its high capacity to process information, computer systems provide the ability to share information globally. Therefore the use of digital processing systems in medicine can resolve two closely related tasks: a) greater accuracy in diagnosis and b) expanding coverage in the applied field of an accurate diagnosis. In general, we can say that the class of diseases for which assisted diagnosis proves to be important, are determined by two factors: first, in diseases where accurate diagnosis is necessary for the success of therapy, is to say, diagnoses that require consideration of many signs or symptoms. As a rule, these are diseases that do not require emergency surgery, but require complex operational techniques. Second, in an acute illness, where the diagnosis also requires precision, more importantly, urgent. In these cases the diagnosis is often made before obtaining the results of additional tests (if these are possible) and concludes with minimal information to run fast-action procedures (surgery, for example).

The number of submitted papers related to research and development of diagnostic applications assisted in IEEE EMB Society, Bio-signals and publications in IEEE Transactions on Biomedical Engineering from 2000 to 2007 are listed in Table 3.1. Most of these presentations have focused their studies in the area of pre and extraction process / selection of features geared to the recognition of functional states.

Table 3.1: Publication of works in science context

Application field	*Number of papers*
Electrocardiography	210
Phonocardiography	73

3.2 Assisted Diagnostic Considerations

Clinical diagnosis, among others, has the following specifications [11]:

Subjective nature. Although the specialist requires additional information and evidence available, he has the final verdict, and thus the quality of the verdict may change from specialist to specialist. Matching the respective studies are considered acceptable to the medical opinion of several specialists, may show values of around 60-70%, which generate enough uncertainty in the training of automatic systems and impose heavy restrictions on their effectiveness.

Qualitative nature. Much of the evidence and information obtained from the records of the activity of the human body, clinical examinations, and others, has been parameterized, and according to their values it has been defined qualitative judging bands, however, there is an ample scope for subjective rating scale on final clinical verdict. The generation of features should be accompanied by a stage where values are defined for each representative functional state, to comply with the conditions of consistency and effectiveness imposed on the estimation procedures, to the effect that the results can be reproducible and transportable. With regard to identification procedures, they must have a close relationship with the qualitative assessment of the functional states and not just restricted to detection.

Local nature. The trial on the normality or abnormality of functional states of the human body, besides being subjective, is not always widespread and universal. For example, the biochemical composition, morphological, anthropometric, biological and clinical varies according to factors such as geographic location, ethnic, social origin, human activity, among others, taking into account the specific values that different communities have developed in the process of local adaptation. In addition, the location may be referred to time, to the extent that the behavior of the organism is non-stationary nature. The body, as a complex system, has dynamic compensation mechanisms that make the measurement of physiological variables change over time.

Heuristic transcendence. The practice of medicine throughout its history has empirically developed a set of complex and refined diagnostic variables, many of them based on direct sensory perception of the doctor, who is not always easy to model and parameterize (for example, the ring of the heart sounds, the morphological changes of ECG recording, etc). However, these clinical diagnostic variables are useful both for discovering states of normality or abnormality, such as for therapeutic treatment. This aspect generates a major constraint in the generation of new variables in the automatic diagnostic models, as it is always preferable that the feature analysis has a physical sense and be easily understood by medical personnel.

Besides the above, in the design of automatic diagnostic systems the following aspects must be taken into account:

Irregular patterns of representation. In medical practice, it is common that several states of normality or abnormality of the human body, be responsible for identical models of representation, is to say, the patient is given a diagnosis made by a single state of functionality, when in fact can have several abnormal states, or otherwise, when the patient receives a diagnosis indicating the presence of various dysfunctions and ,in fact, only has one of them. This issue becomes more complex due to the non-stationary nature of the clinical diagnosis.

Asymmetry in the population distribution of the classes. In general, the population do not have the same number of people in each class to identify, and although this factor is taken into account in the design of experiments related to the training of automated systems, there is great difficulty in achieving sufficient number of samples per class for machine learning. As a result there is a need to maximize the information provided by each clinical observation, or record of generating strategies that involve the pursuit and extraction of more information about each case. With assisted diagnosis, however, it is required that the performance level of the computer response be high. For example, if the sensitivity in detecting computer injury is less than the average sensitivity of the doctors, it would be difficult to justify the use of assisted diagnosis.

Adjustability of diagnosis. The system of health services is divided into different categories or levels that meet the different needs of the service. For example, there are primary care centers located in remote areas or access problems, for implementing the basic perception of normality and abnormality of patients, where the doctor is often confronted with many technical limitations to the registration of bio-signals. For its guidance, in these centers, it is preferable to develop automatic identification systems oriented to screening and / or centralized systems that feed on disease prevention. In these systems development emphasis is on highly accurate detection of abnormal operating conditions, which present a very different nature compared to normal states. Is for this reason, that there are strong restrictions on the acquisition stage of records. The third and fourth level centers located in large cities, who have qualified and have fewer restrictions on the provision of equipment, automated systems require more emphasis on highly accurate differentiation between various states of abnormality nature which are very similar. The solution to both requirements in a single automated system becomes very complex in its implementation and management, preferably parallel development and integration platform that includes additional aid, such as the use of Information and Communications Technologies.

Complexity and high cost. The registration of bio-signals and the formation of the respective databases are a key issue in the development of automated systems that support the clinical diagnosis. First, the high costs are related to the location, transport and preparation of the patients for taking a correct record of the bio-signal. Next, the appropriate labeling of records and the tuning of the various processing algorithms, add a considerable burden to the budget of the research, because it requires highly qualified professional staff with expert advice of the databases, as well as assistance in training the system, not counting the costs of the use of medical equipment that are borne by hospitals and health centers.

The regulations in the area. The area of health, by its impact and importance, is well regulated, both in their professional practice, as in other aspects related to the provision of health services. For example, in Colombia, the conditions of accreditation for institutions providing health services in the form of telemedicine, are defined by Resolution No 1995 of 1999 that establishes rules for handling medical records, including electronic filing form. The transmission and storage of bio-signal and medical imaging has as standard DICOM, and so on.

In this work are presented different models for the development of automated systems, and depending on the ECG signals, allow to take the best objective and collaborative clinical decisions. The determination of risk factors for cardiovascular disease in our Colombian media is critical. In the state of Caldas, the top 10 leading causes of death in 1999 (the latest statistical records found in the particular Territorial Directorate of Health) indicates that mortality from cardiovascular and pulmonary diseases are at the top, surpassing the deaths caused due to violence.

The use of control and prevention methods of cardiovascular disease, involves, among others, the development of more effective tools in the diagnosis of the cardiac function, in particular, there are two ways in the centers of primary and secondary levels of attention, that are considered:

Analysis of the heart's electrical activity through use of electrocardiography, which is done for direct quantification of different parameters (rhythm, frequency, interval estimation, amplitudes), aimed at the identification of electrical events related to ischemic disorders.

Mechanical analysis of the heart through the acoustic signals obtained by auscultation, seeking recognition of abnormalities related to heart murmurs.

3.3 Automatic Identification into the Assisted Diagnosis

In [30], discusses about necessities that are often required for automatic identification related to the training quality that is needed to define the implementation of procedures under crucial and adequate conditions in the resolution of specific tasks of pattern recognition (see Figure 3.1). In this sense, there is a higher value that restricts the training quality given by the sample size and a lower limit given by the clinical knowledge, where the designed procedures attempt to capture the physiological dynamic in order to add stability between the results of automatic system and clinical inspection in a range of quality [11]. The problem of considering a training sample, big enough

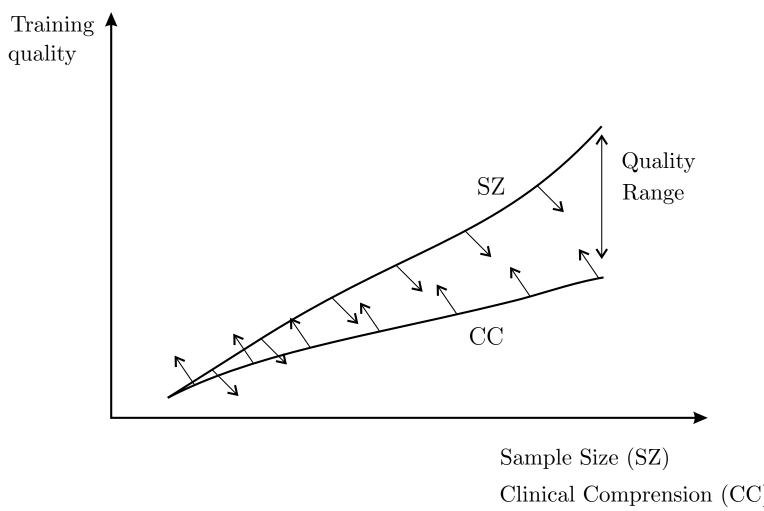


Figure 3.1: Requirements for training quality

in size has its origin in the complex and expensive task of acquisition and appropriate labeling of pathological records, therefore, the databases available for training usually have a small number of records. On the other hand, clinical knowledge is not always easily accessible due to the differences between the structure of medical knowledge and understanding of the software designer [28]. Having the intention of improving the training quality it could be feasible to improve the representation quality by increasing the number of considered features in the analysis, however, in [31] the following relationship are presented:

$$\text{Representation Quality} = \#\text{ of features} \times \frac{|R_o \cap R_e|}{|R_\eta|}$$

Where the representation quality is directly proportional to the number of features, however, is subjected to two rigorous restrictions: the intersection between the representation given by the estimated characteristics (R_e) and natural representation of analysis objects (R_o) must be higher and also the considered characteristics in the procedures must minimize the noise representation (R_η) due to physical or physiological condition of the subjects (athletic conditions, feeding habits, etc). In this way, not always a high number means high quality in the representation of a physiological phenomenon [11].

In the terminology of medical diagnosis, each of the patterns can be identified in terms of formalized symptoms through an electronic record of signals or a set of images of the patient. Meanwhile, the gathered classes represent the range of possible diagnoses or medical claims. In this sense, solutions based on deterministic logic have been raised, but it is more common to use different metric approaches, in particular probabilistic. An ordered set of relevant features form a vector in a multi-dimensional space, and this is a point in that space of representation. The interpretation can be based on the features, geometric proximity in the coordinated space and, additionally, it can be associated with significant symptoms of the disease under study. According to this, the diagnostic method based on clinical preceding (complete or partial) takes as true that the cases related in the same condition are usually identified by characteristics of similar symptoms, this is, into the feature space of multiple dimensions. The match between cases is determined by the distance between their representation points [32]. Nevertheless, there are unsolved problems in cases where the characteristics have different Gaussian probability density or exist non-linear relationships between them. In most cases, hypothesis about approximate probability distributions is used, based on some additional information. Thus, using these hypotheses for each of the pathologies, we can establish relationships and distances in different points in space in order to distinguish among classes. It is also important to consider that it has a relatively small number of features that can find a complete interpretation in their meanings. In a more general case, the distance in the feature space can be used as diagnostic criteria [33].

On the other hand, the information that identifies each of the symptoms is known, and therefore, it is possible to find the characteristics that best describe the diagnosis. Because of this, the feature space has complex anisotropic properties. A partial report can be done in the representation space by giving to the coordinates different weights, which are determined by calculations based on statistical properties of features [33]. This operation means to change the scale in different directions in the representation space. Some algorithms use evaluation functions based on the discriminating power and carry out more complex transformations of coordinates.

The algorithms decision-making are based on the proximity in the feature space, depending on the location of the point according to the areas that belong to different functional states. This implies that the decision surfaces must pass through areas, where the small number of representation points determines the limit between regions with much higher density. The determination of decision about restrictions can be done using conventional methods of objective classification [34] [35]. It is important to notice that increasing the number of diseases to diagnose, requires more complex representation methods, which are usually involved in different phases of the combination process among deterministic logic, metric approaches and information sources.

In general, pattern recognition can be divided into a series of stages as it is shown in Figure 3.2 [28].

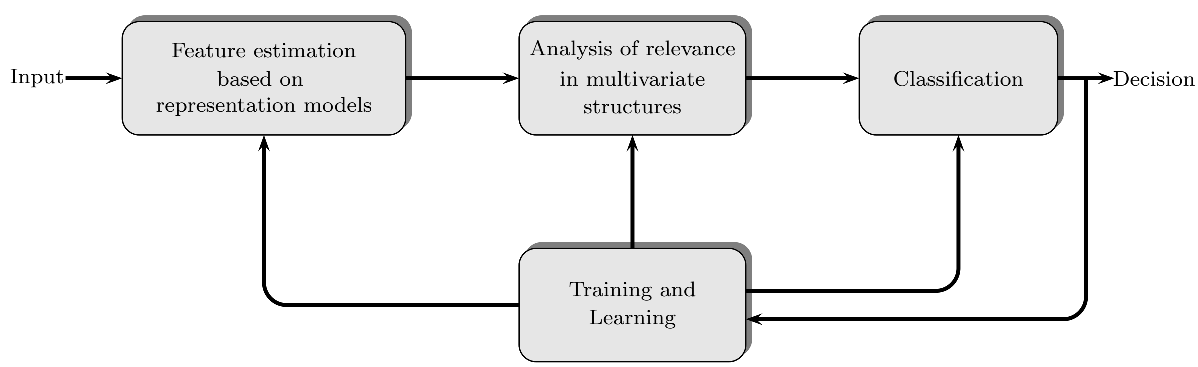


Figure 3.2: Block diagram corresponding to the pattern recognition

The first stage corresponds to the feature estimation based on representation models, which extract the information related to the description of the functional states to identify, and consists on transforming each of the patterns of representation in measurable quantities. In the next stage, called relevance analysis of multi-varied structures, it chooses a reduced set of features representing the patterns, in order to preserve the discriminant information of the classes, that are required to identify in the attended diagnosis.

This procedure is also called dimensional reduction and is subjected to a relevance condition, which seeks precision in the identification, computational low-cost and reduction of the operation complexity. Finally, the classification stage is responsible for assigning a class label to a specific pattern, using a selected feature set for this. In every stage runs at the same time the training and learning process, being this, the most important procedure of pattern recognition, now that, is where the operating parameters are determined and adjusted.

Estimation of features. For the estimation of features is important to do previously the reduction of disturbances that are commonly found in bio-signals and the segmentation of the events that characterize the functional states, using strategies ranging from linear filters to the whole time-frequency process. The magnitude of the problem that involves the signal preprocessing, and its importance is reflected in the amount of work in this area reported in the literature (see Table 3.2), also the fact of being the first stage of the bio-signals processing. The percentages of every process stage are shown in Table 3.2, and are obtained from the quantities that were exposed in Table 3.1.

Table 3.2: Quantity of works per processing stage

Area of application	Signal preprocessing	Feature estimation	Dimensional reduction	Classification strategies
ECG	28%	33%	29%	10%
PCG	37%	31%	6%	26%

Although bio-signals are processes with a wide non-stationary character, the mathematical equipment employed in the analysis of bio-signals, was Fourier, particularly, the short time interval processes.

The problem is that the trigonometric functions can not locate the analysis at fixed intervals of time (for example, during the appearance of abrupt changes in different functional states of the body). Because of this, the Fourier transform does not take into account that the periodic or quasi-periodic parameters (amplitude, frequency, phase) can evolve over time, creating difficulties related to insufficient length of the signal over the period of analysis, the presence of non-stationary origin of fluctuations, loss of data in few moments of time, etc. The Inadequate conventional methods of spectral analysis, in the description and analysis of the non-stationary bio-signals, has required new forms of mathematics and statistics representation. A particular interest is for wavelets, based on a class of special functions with localization properties both in physical space (time), as in the Fourier space. It is important to notice, that although the wavelet functions are widely used in bios-signal analysis with non-stationary structure, the Fourier's harmonic analysis is not excluded [36]. Otherwise, the system of representation in the form of point values is a basic strategy for characterization of random variables and their appropriate use of influences on the effectiveness of the system. However, the non-stationary nature of the bio-signals makes the representation difficult through these point values, so it is necessary the generation of random features in arrays (contours or surfaces), seeking the dynamic changes of these new features, revealing important information to the concept of classification over a range of analysis, e.g., time, frequency, wavelet representation scales, etc. In the particular case of the automatic recognition of heart disease, it is reported a progress in the specific cases of electrocardiography [37] [38] and phonocardiography [39] [40].

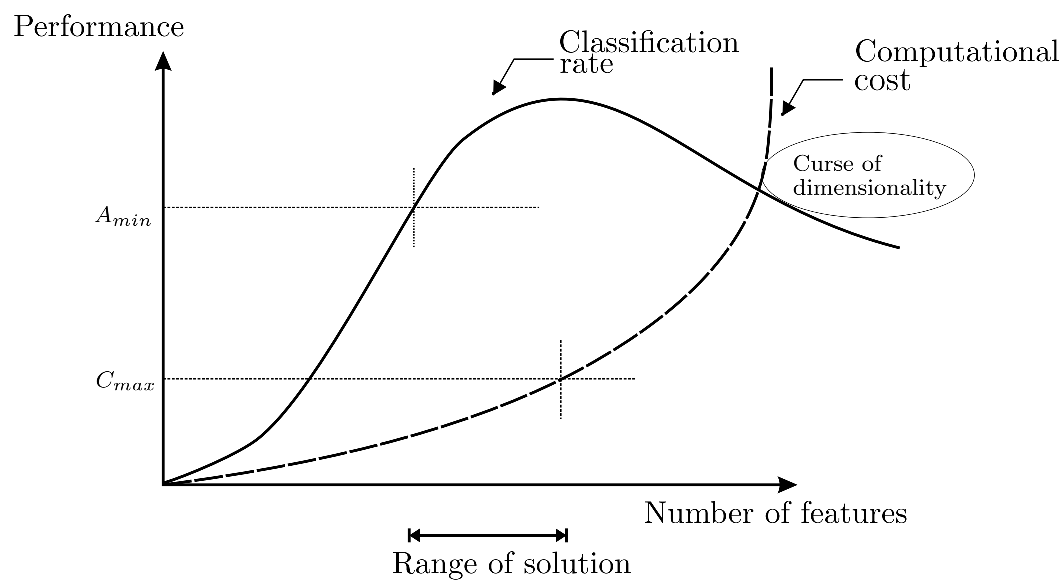


Figure 3.3: Solution range

Analysis of relevance. There is a strong restriction in basic methods of learning, consisting on the number of characteristics or descriptive variables, and must be much less than the number of records available for training; while the number of features increases, requires an exponential increase of observations to keep the density of the given sample (*curse of dimensionality* [41]). It can be shown in Figure 3.3, for a training record-set given, there are a number of characteristics over which the classifier performance will decrease instead of improving from a [42], which is possibly related to the fact that in high dimensionality data, not all measured variables are relevant in terms of representation [43]. Thus, Figure 3.3 shows the acceptable compromise values between the computational complexity (maximum value permitted C_{max}) and the success rate of classification (minimum acceptable value A_{min}), which determine an appropriate operating range for the number of features where the reduced space does not compromise the classification accuracy of the detection system. According to this, the reduction of dimensions made on representation spaces can be taken as a task of optimization associated with a condition of relevance, taking into account that the relevance is which defines the context of reduced representation [28]. In practice, the problem consists on defining exactly the goal of relevance in order to the reduced-space be supported with enough information, allowing a high ability of generalization when evaluating new examples are required, in this way, the skill of the detection system does not depend only on the efficiency of the classifier [11]. In [44] shows different important aspects of analysis and introduces an approach to the problem according to the perspective of the following definition:

Definition 3.3.1 (Purpose of relevance)**.** *A characteristic ξ is relevant to the objective c if exists a couple of examples A and B in the space of representation, such that A and B differ only in their assignation to ξ and $c(A) \neq c(B)$.*

In general, a limited number of features simplify the representation space, becoming the analysis less dense and exhausting, in order to obtain a classified stage faster and with minimal memory requirements. However, an excessive reduction in size could lead to a loss of representation, reducing the quality and accuracy of the recognition system.

Classification. In the automated diagnosis, this stage defines the class or functional status that belongs to a represented pattern analysis in a feature space. In the bio-signals analysis, right after they have been estimated on the features about the set of training records, each signal is represented by a vector $x = [x_1, \ldots, x_p]$, which is known as a measure vector or pattern vector. For an effective classification, it is desirable that the patterns for every class form groups or clusters.

The characteristics which are common along the patterns of a particular class are known as intra-class features. The discriminant features that represent the differences in various pattern classes are called inter-class features. The classification problem is to generate limits or borders of decision, which optimally separate the data according to different kinds of patterns in the training set (see Figure 3.4), and different supervised and non-supervised classification techniques have been proposed in order to solve the problem [45]. The disadvantage comes out when the classes which are discriminated have non-Gaussian probability distribution and separation borders are highly non-linear, since the conventional techniques of classification for these cases become ineffective. On the other hand, for the classifier evaluation, the set of samples available can be divided into two groups: training set and validation set. The basic method of validation is known as simple validation and the steps to evaluate a classifier using this method can be listed as follows:

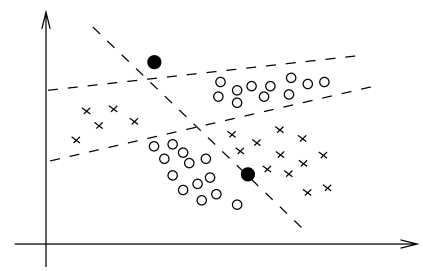


Figure 3.4: Generalization capacity

1. Separate randomly from the total sample size N_s, two sets of size N_t and N_v, where $N_t + N_v = N_s$.

2. Build the classifier using the training set.

3. Using the classifier, just assign a class label to each pattern of the validation set. The correct class for each pattern evaluated must be known previously, in order to obtain the number of cases correctly classified N_{CC}.

4. Calculate the classification accuracy by the expression: $P = 100 * N_{CC}/N_v$.

The advantage of this method is that it does not use much processing time; however, the result may have an unpleasant variability, since the test depends largely on the choice of data which form the training and validation sets. Because of this, there is, besides other techniques, cross-validation (cv), which is often used when the sample size is not large enough, but leads to increased computational complexity [46].

Chapter 4

ECG Analysis: A review

Over the last two decades, the acquisition systems of portable physiological signals are more lightweight and small, and can also record multiple signals up to 48 hours. These long-time recording systems, also called ambulatory record systems, are used in studies of electrocardiogram (ECG) to detect infrequent arrhythmias or transitory cardiac abnormalities in heart function often related to the everyday stress, such as transient ischemic episodes or silent myocardial ischemia. This kind of systems takes importance when pathologies can not be detected by using only ECG clinical records of 12-leads. Devices for ECG recording are called Holter records, named after an ECG recorder developed by NJ Holter in 1961. The first Holter-recorder used a tape to store large files of signals, because the ability of solid state memory was not available at that moment. Nowadays, the signals are recorded in flash-type semiconductor memories, which can be transferred to a workstation for their posterior analysis.

Usually, the analysis of recordings is carried out off-line, and is known as automated support, which takes into account several factors like the length of recording; having stored in a single record more than 100.000 heartbeats to be examined, and it is very important not to discard any heartbeat because the diagnosis only depends on some of them, which implies a reduction of process time for automatic task. Another factors to be also taken into consideration are those related to disturbances on the signal such as motion artifacts, baseline wander, EMG noise, among others.

In order to get a full analysis of the records, it is required a series of stages which include digital processing techniques for signals. Commonly, these stages include pre-processing, segmentation, feature extraction/selection and classification. A correct selection of these techniques is mandatory, taking into account the factors that may affect the signal. This work reviews the main techniques used in the analysis of ECG signals. In preprocessing is highlighted the effectiveness of non-stationary analysis to separate into bands the frequency-features of the signal and thus be able to remove disturbances that affect the analysis of the signal. Segmentation has similar methodologies, but uses different combination of techniques for signals processing. It also reviews the extraction and feature selection techniques, which reduce significantly the computational cost for algorithms, but providing same results.

There exist methods for extracting small sets of representative beats from the total ECG record. Two different approaches are possible in the design of procedures for extracting representative beats: supervised and unsupervised approaches. In the first case, it is necessary to have a complete set of beats manually labeled, and in that way classifying every new heartbeat of the signal under an analysis of one known class, giving as result an automatic diagnosis; while unsupervised techniques do not require the label of heartbeats, and are more flexible. This type of techniques is the most frequently used and is based on examining all the beats and using an appropriate measure of dissimilarity for the unsupervised classification (or clustering) applied on the set.

Next, some works of the literature, which have used suitable tools to carry out each of the procedures described above, will be exposed.

4.1 Noise Reduction of ECG Signals

Noise reduction of ECG signals has became a vast research subject and there are studies that use from classical techniques to new noise estimation methods, including wavelet denoising, dimensionality reduction and classifiers. In general, there are two types of noises: artificial type and biological type. Powerline interference, impulsive noise, electronic devices noise and electrodes contact artifacts are found among the former type. Muscle contraction noise (EMG), ECG baseline wander and ECG motion artifacts are found among the latter type.

Noise reduction methods focus mainly on pathology diagnosis, i.e., the signal, after having been filtered, should not lose its intrinsic characteristics such as morphology and duration, which is a complex task since there are some noises mixed with the signal spectrum. There are several techniques to address the problem such as [47] in which two filters in cascade are applied to eliminate high frequency noise, the first is a filter of order 1 and the second is a filter of order 2. The structure of difference equation for the filter is $y[n] = y[n-1] + x[n] - x[n-m]$. To reduce baseline drift a high pass filter type *IIR* with cutoff frequency at 0.05 *Hz* is implemented. The problem related to this type of filters is the presence of secondary lobes and low selectivity. There is another technique, called mobile average filters, whose main disadvantage is the low selectivity and the masking of high frequency details. In [48] it is presented a possible form for the implementation of this technique.

Another known technique is the use of time-varying filtering, in which the cut-off frequency varies according to the low-frequency properties of the ECG signal with respect to their average. In [49], this technique is implemented, adapting the cut-off frequency of the filter to the current "level" of baseline wander in the signal. If no baseline wander appears, the filter with the lowest cut-off frequency is used, and so on. The adaptation is related to the error between the output of the currently selected filter and the filter output for the highest cut-off frequency.

One of the most implemented techniques for the biosignals filtering is the signal average, but the main disadvantage is that signal and noise must satisfy certain conditions. For instance, the noise to be detected must be invariant in time, stationary and additive, and not correlated with the signal cycles. These restrictions make the method be limited in its benefits. However, in [50] this technique is implemented.

There are other techniques that are based on morphology recognition and use signal approximations in order to reduce noise in general. In [51] and [52], the signal is represented by means of orthogonal expansions with a small number of coefficients using inner product and adaptive estimation according a LMS criterion.

Adaptive filtering is another well established technique for interference rejection. It exhibits remarkable results in quasi-stationary noises such as power line interference and baseline wander. In [53] an incremental estimation filter is proposed, which estimates the noise output as a reference of the last three samples of the contaminated signal and which has a tolerance of $\pm 100 mHz$ depending on some adaptation parameters. In [54] a method with notable results called Adaptive Sinusoidal Interference Multiple Canceler (ASIMC), in which, from a given input signal, $r(kTs) = s(kTs) + \sum_{i=1}^{nw} \alpha_i cos(\omega_0 kT_s + \phi_i)$, the aim is to extract the desired signal $s(kTs)$ from the contaminated signal $r(kTs)$ without distortion. The estimated parameters of phase $(\hat{\phi}_i)$ and amplitude $(\hat{\alpha}_i)$ are iteratively adapted by minimizing the mean square value of an error function $e(kT_s)$ which depends on $r(kT_s)$ and on the estimated parameters. The algorithm works correctly when amplitude and phase have small variations, but its main fault lies in frequency variations that may reduce its performance.

In [55], a general method to remove white Gaussian noise with zero mean and unitary variance, using different thresholding methods applied to Wavelet decomposition coefficients, is proposed. This approach has been successful in reducing EMG noise, power line interference and baseline wander in ECG signals. In [56], a methodology to reject power line interference using 33 families, 7 decomposition types and 8 threshold levels, is proposed. The search for the best combination was done heuristically, calculating distortion measures for each iteration. A test with 105 recordings of the QT database showed the best results with the mother wavelet coiflet 2, decomposition level 2 and heuristic sure-hard-thresholding. The disadvantage is that the combination fits properly with this particular database, exhibiting less favorable outcomes in different signal types. To improve this drawback, mother functions fitted to the ECG signal ensure that a wide range of signals can adequately be represented with the found mother wavelet.

In [57] a new method for power line interference elimination, using an algorithm that extracts a specific signal component as a central block and follows its variations in time is proposed. A set of nonlinear differential equations governs the algorithm dynamics. The descent gradient method to minimize the least square error between the input signal and the desired sinusoidal signal is used. Poincaré´s method is used to test the mathematical properties of the algorithm.

In [58], based on the work of [59], an orthonormal wavelet which can be fitted to any signal of interest using Multiresolution analysis (MRA) is developed. The MRA decomposes a signal $f(x)$ into a serie of detail functions, W_j and an approximation function V_j. That is, $f(x)$ is projected in W_j and V_j, where $V_{j-1} = V_j + W_j$ is a vector sum and where W_j and V_j are orthonormal. The projection of $f(x)$ in W_j and V_j produces the detail functions $g_j(x)$ and $f_J(x)$, such that $f(x) = f_J(x) + \sum_{j=1}^{J} g_j(x)$. The orthonormal bases of W_j and V_j are given by the wavelet $\Psi_{j,k}(x)$ and the scaling function $\phi_{j,k}(x)$. The adjustment process is carried out in two steps. First the spectral amplitudes between the ECG and the wavelet are adjusted and, in the second one, the delay group is adjusted. But whatever the orthonormality conditions of amplitude and phase, the application is suboptimal. This method allows generating wavelets that have optimal adjustment to the desired signal.

In general the techniques mentioned above are used to reduce interferences, whose spectrum does not overlap with the spectrum of the signal, having trouble for noise of motion artifacts or electrodes disconnection types.

In [60], an analysis of ambulatory ECG recordings, focusing the objective on the motion artifacts analysis taking as reference the Body Movement Activity (BMA) is developed. An arbitrary *BMA* during the ECG recording causes electrodes movement, which induces noise in the signal also known as motion artifacts, which is a challenging task due to the artifact and the ECG signal overlap in their spectra. As it is not possible to analyze the signal in the spectral domain, the Principal Component Analysis (PCA) decomposition is performed. The analysis is based on specific types of motion artifacts through the characterization and classification of the main *BMA*, such as sitting and relaxing, hands movement, walking, climbing or going down stairs, among others, taken from the ECG signal. As a result an assessment regarding the performance of the classifier and false-detection rate, where the highest values obtained were 0.985 and 0.014 respectively is carried out. Performance decreases when *BMA* types increase.

In [61], two methods to remove artifacts from ECG signals acquired by portable systems are discussed and tested. The first method uses an adaptive filtering scheme, which requires a reference input, composed by an electrode placed strategically in the right shoulder of the patient, which functions as an accelerometer. The accelerometer located in the z-axis, vertically parallel to the surface of the body, had the highest correlation with the artifacts in motion present in the recording. Two widely used algorithms in adaptive filters were implemented, the LMS (Least Mean Square) and RLS (Recursive Least Square).

The RLS algorithm had better results than the LMS algorithm, for the *DII* lead, taking into account mean and standard deviation of an error measurement that indicates the noise residual percentage after applying the proposed filtering techniques. It is noteworthy that the convergence rate of the RLS algorithm is an order of magnitude larger than the LMS. Although this means a better performance for RLS, its computational complexity increases. The second method was based on ICA (Independent Component Analysis), using the multichannel nature of the ECG recording.

Two leads *DI* and *DII* were processed by the basic and convolutive ICA models. Taking into account the residual noise percentage present in *DII* lead, evaluated after the artifacts removal stage, the basic ICA model showed better results. The end result demonstrates experimentally that the basic ICA model overcomes the adaptive filtering, but ICA remains a powerful tool for these applications. However, the selected performance index does not provide complete information on the morphology of the ECG signal recovered. Nonetheless, it is noteworthy that the adaptive filtering technique can be effective only if the reference input is strongly correlated with motion artifacts added to the ECG signal.