Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California

Doctor of Science
in Electronics and Telecommunications with orientation in Telecommunications

---

Development of techniques for the analysis and processing of cardiac sound signals

Dissertation

submitted in partial fulfillment of the requirements for the degree of

Doctor of Science

Presented by:

Roilhi Frajo Ibarra Hernández

Ensenada, Baja California, México

2019

Thesis defended by

Roilhi Frajo Ibarra Hernández

and approved by the following committee

Dr. Miguel Ángel Alonso Arévalo
Thesis Director

Members of committee

Dra. Nancy Bertin

Dr. Salvador Villarreal Reyes

Dr. Roberto Conte Galván

Dr. Martín Osvaldo Méndez García

Dr. Daniel Sauceda Carvajal
Postgraduate Coordinator in
Electronics and Telecommunications

Dra. Rufina Hernández Martínez
Director of Graduate Studies

Resumen de la tesis que presenta Roilhi Frajo Ibarra Hernández como requisito parcial para la obtención del grado de Doctor en Ciencias en Electrónica y Telecomunicaciones con orientación en Telecomunicaciones.

Desarrollo de técnicas para el análisis y procesamiento de señales de audio cardiaco

Resumen aprobado por:

Dr. Miguel Ángel Alonso Arévalo
Director de tesis

La auscultación cardiaca ha prevalecido como una herramienta sencilla, económica y primaria para la inspección de enfermedades cardiovasculares. De esta técnica proviene el audio cardiaco o señal de fonocardiograma, elemento central de análisis de la tesis. Las limitaciones en el sistema auditivo humano en la banda de bajas frecuencias y la falta de expertos en auscultación ha restado su potencial como técnica de diagnóstico. Sin embargo, los recientes avances en el área del procesamiento digital de señales permiten desarrollar técnicas para fortalecer el análisis atomatizado de sonidos cardiacos. La presente tesis explota un modelo paramétrico de reconstrucción de sonidos cardiacos basado en Matching Pursuit con átomos tiempo-frecuencia y Codificación Predictiva Lineal. El modelo es rigurosamente evaluado por medio de pruebas objetivas y subjetivas con el apoyo de expertos en la salud. Los parámetros del modelo serán considerados para el diseño de un clasificador que detecte la presencia o ausencia de patologías en señales de audio cardiaco. Se comparó el desempeño de diversas técnicas de clasificación como métodos de aprendizaje máquina. Diversos métodos de extracción y selección de características así como de balanceo de clases fueron considerados en el diseño del clasificador. El objetivo del trabajo es explorar los métodos de aprendizaje máquina antes que implementar técnicas sofisticadas de clasificación. Los resultados generados muestran que los métodos de extracción de características influyen directamente en el desempeño del esquema de detección. También se reflejó en las pruebas que los métodos de clasificación de sencilla implementación pueden obtener resultados satisfactorios.

Palabras clave: Señales de audio cardiaco, procesamiento de señales, clasificación, modelado, evaluación, Matching Pursuit

Abstract of the thesis presented by Roilhi Frajo Ibarra Hernández as a partial requirement to obtain the Doctor of Science degree in Electronics and Telecommunications with orientation in Telecommunications.

Development of techniques for the analysis and processing of cardiac sound signals

Abstract approved by:

Dr. Miguel Ángel Alonso Arévalo
Thesis Director

Cardiac auscultation remains as one of the primary and low-cost screening tools for the detection of cardiovascular diseases. This technique is based on the analysis of the heart sound signal, formerly known as phonocardiogram, which is the central element under study of this thesis. The limitations in the human auditory system at low frequencies and often the lack of auscultation skills of general practitioners limits the potential and interest of the cardiac sound signal as diagnosis tool. However, recent advances in digital signal processing allow the development of techniques for the automated analysis of heart sounds. This thesis work proposes the use of a parametric reconstruction model for cardiac sounds based on the Matching Pursuit algorithm with time-frequency atoms and Linear Predictive Coding. The model was rigorously evaluated through objective and subjective tests with the assistance of health experts. Parameters of the model were considered for the design of a classifier for the detection of pathologies in phonocardiogram signals. We compared the performance of several classification schemes used in machine learning systems. Different methods of features extraction and features selection as well as the use of balancing techniques were considered for the classifier design. The obtained results show that the features extraction methods directly affect the screening method performance. The results also show that low-complexity classification schemes can achieve scores comparable to more complicated methods.

Keywords: Heart sound signals, signal processing, classification, modeling, evaluation, Matching Pursuit

Dedications

A Elsa, en cuyo camino me he cruzado con gran fortuna. A mamá y Jalcbe que juntos hemos cerrado y abierto ciclos y traspasado adversidades. A papá, a Debbie, sé que en algún punto volveremos a encontrarnos ...

Acknowledgements

Al Centro de Investigación Científica y de Educación Superior de Ensenada por ser mi casa durante más de 6 años y dejar una huella en mi formación académica, profesional y humana.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme el apoyo económico para realizar mis estudios de doctorado. No. de becario: 477876.

Al Dr. Miguel Ángel Alonso Arévalo, quien me abrió las puertas y me dio la maravillosa oportunidad de incursionar en el mundo de la investigación. Además de permitirme conocer varios lugares y entornos académicos. Sin duda ha dejado plasmada una huella en mi formación.

A Nancy Bertin, que me abrió las puertas de entrar al INRIA de Rennes, Francia y a todo su grupo de investigación PANAMA. Hicieron de mi estancia una época maravillosa que no olvidaré.

A mis compañeros cercanos de maestría con quien he creado vínculos inimaginables en especial a Rogelio Ojeda, quien estuvo muy cerca y pendiente de todo lo que he hecho. También agradecer a Ernesto Ortiz, Erick Calixto, Rafael García, Genaro Beltrán, Mario Rosas, Javier Ortiz, Fernanda, Julio Heredia, Eulogia, Sergio Toledo. Especial agradecimiento también a mis *paisanos* zacatecanos: Giovanny Cabral, Sergio Armas, Diana Díaz, Jairo Donlucas, Ramón Solís, Andrea Murillo, Guillermo González y al resto de la embajada zacatecana: Ismael Casarrubias, Javier Mendivil, Neto Cosío, con quienes viví grandes momentos.

A los compañeros de maestría que nos recibieron cuando llegamos al CICESE e hicimos una gran amistad Gabriel Echegaray, Anela Sánchez, Carlos Nieblas, Jonathan Monjardín y Gonzalo Cruz, Vero Rojas, Marcelo Maciel, Beatriz Stephens. A mis amigos

doctores egresados del CICESE: Rodrigo Méndez (mi gran amigo el hueón), Héctor Sánchez y sus familias (Héctor Sánchez *tío*, Teresa Esquivel, Carolina Román).

A los investigadores y personal del departamento de Electrónica y Telecomunicaciones del CICESE que me apoyaron y brindaron una sonrisa, sus conocimientos o una charla agradable: Roberto Conte, Ramón Muraoka Jaime Sánchez, Moisés Castro, Salvador Villarreal, Carmen Maya, Ricardo Núñez, Iván González, Erika y Aurorita.

Al club de atletismo del CICESE, con quienes hice gran hermandad: Brenda Anda, José Mojarro, Ana Laura Flores, Lili Espinoza, Erik Lemus, Bety Torres. Y al profesor Humberto Delgadillo por entrenarnos.

Al club de música del CICESE en todas sus facetas: Hugo Guillén (quien también contribuyó significativamente en mis conocimientos en Machine Learning), Luis Galaviz, Ceci Mozqueda, Eliana Rosas. A la nueva camada: Eduardo Verdugo, Paola Valdés, Miguel Luviano, Angel Regalado, Jesús, Juan Valle, José Chan. Y en especial a Laura Rebeca Pineda que nos apoyó mucho a que creciera este proyecto.

A mis amigos de la banda de óptica: Alma Karen, Jorge Munky López, Gibrahm Alejandro, Annita, Liliana Ávalos, Gaspar *Chezpi* Uriarte (†).

Al grupo de futbol y amigos de Los Compas F.C., quienes también hicieron mi estancia basante agradable: Pedro García, Andrea Acosta, Saúl Aguilar, Rogelio Carrasco, Rubén González, Francisco Franco, Carlos *Refu* Hernández, Gerardo Rangel, Víctor Aréchiga, Carlos Díaz, César Guevara y Marcos Aboytia.

Finalmente a mis compañeros de nueva camada de maestría: Rolando Díaz, Luis Mízquez, Gustavo Ramos, Jonathan Hirata, Iván Gutiérrez y Alan Calderón. También a Willem de Jonge, con quién pasamos grandes momentos.

Table of contents

Page

Table of contents

List of figures

Figure Page

List of figures

Figure Page

List of tables

Chapter 1. Introduction

The human heart is the most important organ since it provides blood to all parts of the body. It uses a pumping like method for the blood distribution. This pumping action involves mechanical and electrical activities produced by the heart in order to conduct the blood flow. The activity of a healthy heart is relevant for the daily activities of human body as blood carries important nutrients to the organs. According to the World Health Organization, cardiovascular diseases (known as CVDs) are the first of the top 10 main causes of morbidity (WHO, 2018). Heart diseases have been in the last years the world's biggest killers, accounting for a combined 15.2 million deaths in 2016. These diseases have remained the leading causes of death globally in the last 15 years. Mexico cannot be excluded from these statistics, since obesity and diabetes produced by bad nutrition and sedentary habits have contributed to make heart diseases one of the main causes of morbidity in our country [1].

A number of methods have been developed for the detection and diagnosis of CVDs. Some of these methods are highly sophisticated and provide reliable diagnosis of heart pathological states such as the magnetic resonance imaging (MRI), echocardiography, and cardiac catheterization. Although these methods are powerful tools for the evaluation of heart pathological states, they are limited in availability so their cost would be relatively high. Even with the availability and cost that these heart screening tools present, they require highly trained personnel for data acquisition and interpretation (Mahnke, 2009).

On the other hand, the heart sound signal (that we will call PCG, from the acronym of phonocardiogram) is still a powerful and popular tool to diagnose a pathological state of the heart. The signal is recorded when listening the activity in the heart valves during the auscultation, commonly performed with a device called stethoscope. The auscultation method remains as the first screening tool utilized by primary care providers. In fact, ECG and PCG exhibit similar features since both are highly correlated and *cyclostationary*, i.e. their statistics vary but are repetitive over a period. However, the PCG signal offers a distinct advantage to detect murmurs, which represent an abnormal heart condition, since it records acoustic properties of the heart valves.

[1] Information given by the National Institute of Geography and Statistics of Mexico (INEGI), more information is available online at: `http://www.beta.inegi.org.mx/contenidos/saladeprensa/boletines/2018/EstSociodemo/DEFUNCIONES2017.pdf`

Furthermore, the PCG signal offers a low cost method for diagnosis since we only need a stethoscope as the main device for the recording.

An accurately performed auscultation also requires trained personnel for a correct interpretation of the sounds. Therefore, physicians require auscultation skills to accurately perform the detection of a heart disease. In addition, just a fraction of the acoustic energy generated by the heart can be accurately detected by the human auditory system (Mahnke, 2009). Figure 1 shows the pressure levels and frequency region of speech, heart sounds and murmurs and the audibility threshold to illustrate the portion of energy commonly detected by the human auditory system.

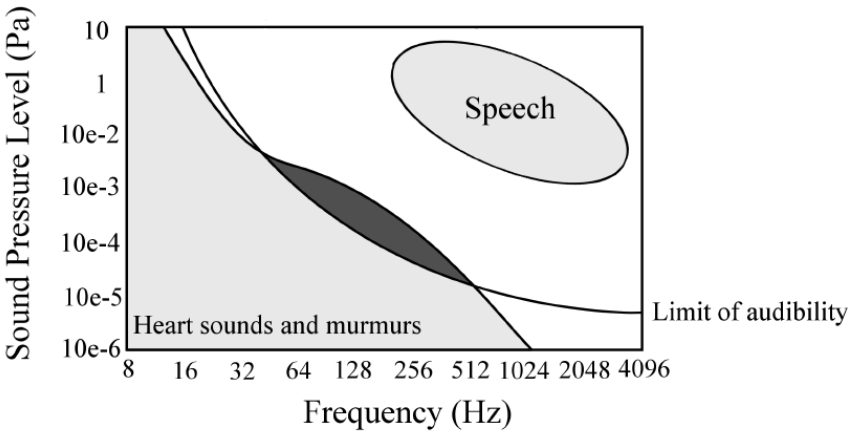

Figure 1. General frequency content of Heart sounds, murmurs and other physiological sounds (Springer *et al.*, 2016; Liu *et al.*, 2016).

Given these limitations for the implementation of the auscultation technique, the presently available computational tools for signal analysis and processing, and the need for an inexpensive tool for the cardiac diseases detection, a number of researchers started to develop methods for the automated analysis of heart sound signals, whose applications basically consist of three steps:

- **Pre-processing**: Comprises all methods for the signal quality assessment and relevant features extraction, such as filtering, noise removal and modeling for signal reconstruction.
- **Segmentation**: It is the extraction of the time periods to delineate the start and end points of the PCG signal main events.
- **Classification**: This step consists in the prediction of cardiac pathological states

from PCG recording features.

Modern stethoscopes are digital devices which contribute to the development of automated heart sounds analysis. Since they allow the visualization of the PCG waveform, the signal processing and analysis methods can be applied by using the recording samples.

The present work focus on the pre-processing and classification methods for the analysis of heart sound signals. It exploits the sparse representation of the PCG as a parametric model to obtain and explore a number of different signal features. Within the use of supervised machine learning tools, it is possible to map the generated input features for classifying the heart sound as normal or abnormal according to the absence or presence of a murmur or pathological state. The classification experiments conducted during the present work were performed using the PhysioNet/Computing in Cardiology Challenge 2016 database (PhysioNet/CinC Challenge 2016), which is the largest collection of heart sounds gathered by the scientific community in the world recorded with normal and abnormal conditions under a clinical environment (Liu *et al.*, 2016). A number of known machine learning classifiers and data mining techniques were rigorously evaluated.

1.1 General Objective

- To develop a heart sound signal model based on sparse representation and autoregressive processes. Use the model parameters to detect the presence or absence of pathologies in heart sounds.

1.1.1. Specific Objectives

In order to accomplish the general objective specified above, we set the following specific objectives.

- Enhance and complete the model described in our previous work (Ibarra, 2014). Perform objective evaluations by comparing different metrics.

- Conduct a formal subjective evaluation with the assistance of physicians and health experts using the MUSHRA method.
- Test the parameters of the model to be mapped as features for classification tasks (features extraction step). Develop a classification benchmark using the database provided by the Physionet/CinC 2016 Challenge.
- Compare the performance of different classifiers to detect pathological states from cardiac sound signals using the parameters of our reconstruction model.
- Compare the performance when using different ML feature selection and balancing methods. Compare also the efficiency of proposed method among different sets of features using features extraction methods found in the literature.

1.2 Methods of the thesis

According to the stated objectives, the present work has been organized in different phases. The following points make a recapitulation of the most important steps:

- **State of art review:** We revised the literature related to PCG signals automated analysis: reconstruction modeling, segmentation and classification of heart beats.
- **Collection and segmentation of PCGs:** We analyzed the settings given by the Physionet/CinC Challenge database, such as time duration, sampling frequency, segmentation annotations, etc.
- **PCG signal modeling:** We enhanced and completed a PCG signal modeling motivated by our previous work. Framework was basically summarized in the three following steps:
  - Sparse modeling: We compared the use of different time-frequency dictionaries for the sparse reconstruction of the PCG signal when using Matching Pursuit: Wavelet packets (WP), Modified Discrete Cosine Transform (MDCT), Gabor and Chirped Gabor atoms. We tested the Orthogonal Matching Pursuit reconstruction of the heart sounds.
  - Autorregresive modeling: We considered to reconstruct the residual part by using the Linear Predictive Coding method.

- Model evaluation: We tested the model using an objective evaluation by comparing the distortion and the plots in time, and time-frequency planes of the original and reconstructed signals. We also conducted a subjective evaluation by using the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) method. A number of physicians and health experts assessed the clinical quality of the reconstructed signals.

- **PCG signals classification:** We developed a benchmark for the classification of cardiac sounds. The methods employed can be stated as follows:

  - Features extraction: We mapped the MP and LPC parameters of the model to features. Then, we considered two approaches in order to have the same amount of features: averaging of the features and averaging of the heart cycles samples. A set of Mel Frequency Cepstral Frequency Coefficients (MFCCS) and wavelet parameters as features was also constructed to be compared.
  - Balancing of data: Since the database Physionet/Cinc 2016 Challenge contains a few number of PCG recordings labeled as abnormal compared the recordings labeled as normal we need to equalize the amount of samples. We conducted and compared the classification results using the Synthetic Minority Oversampling Technique (SMOTE) to balance the data set.
  - Cross validation (CV) test: This test was conducted in order to test different features sets given by the combination of cycles averaging or time averaging approaches with balanced or none balanced data cases. A number of known classification schemes were tested: Linear Regression, Linear Discriminant Analysis, Naïve Bayes, Classification and Regression Tree, K-nearest neighbors and Support Vector Machines.
  - Features selection: To reduce the dimensionality or number of features in the data sets, we tested the performance of three different approaches: feature importance amount (given by the random forest classifier), information gain and correlation features selection.
  - Implementation of the classifier: We conducted the evaluation of all the classification settings showed above using different metrics.
  - Our datasets need to be split into training (80% of the data) and testing (20% of the data) subsets.

- We compared the performance when combining different machine learning methods for features selection and the inclusion of the SMOTE oversampling technique.

1.3 Structure of the thesis

The organization of this manuscript is as follows, Chapter 2 details the methods and techniques employed for modeling of the PCG signal. It starts with the physiology of heart sounds and the mathematical aspects of the signal in time and frequency. Then, it enlists a brief review of the state of the art in modeling of cardiac sound signals. Chapter 2 also describes the methods employed in this research work for the parametric reconstruction of the PCG. Chapter 3 depicts the mechanism for the classification for heart sounds. It starts with a description of the techniques employed for the Physionet/CinC 2016 challenge and the database of sounds which was used to conduct the experiments of the thesis. We also present in this chapter a description of the classifiers taken in consideration for our task and their tuning to be set up. Chapter 4 reports how we conduct the features extraction procedure. It also includes the methods for balancing the data and a justification of why it is necessary to consider this problem. Chapter 5 reports the results of testing a number of classifiers as tools for the detection of pathological states in heart sound signals using different metrics. Finally, chapter 6 draws the conclusions reached after analyzing the results of the present work. It also presents some aspects to be discussed as future work. In order to summarize the research work reported in this thesis, Figure 2 depicts a *road-map* to detail the phases, research methods and productivity derived from this research.

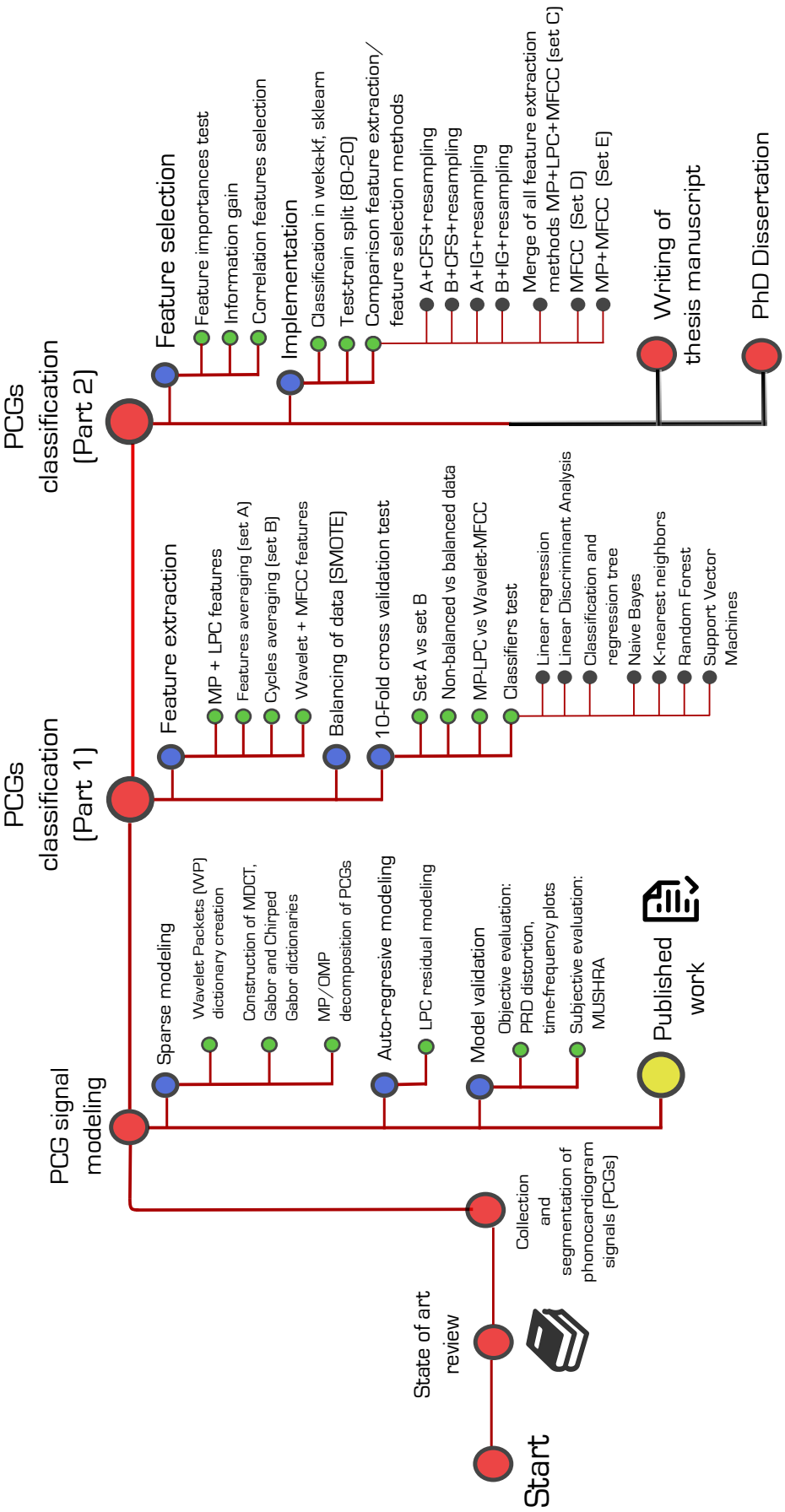

Figure 2. Road-map of this thesis work.

Chapter 2. PCG signal modeling

This chapter describes the characterization of the heart sounds. First, we sketch the physiological aspects concerned to the PCG signal and the procedures involved. We explain the PCG aspects in terms of time and frequency domains as electrical signal and how are characterized those that include murmurs or pathological states. Then, we present a review of the state of the art in the modeling of heart sounds, focusing on the time-frequency methods. Then, we explain the Matching Pursuit representation using Gabor atoms as a time-frequency modeling technique of the PCG signal and the LPC coding of the residual. This chapter also includes a review of the state of the art for modeling of cardiac sounds pointing out which techniques have been used to accurately represent heart sound signals.

2.1 Physiological characterization of heart sounds

The phonocardiogram or PCG signal is a vibroacoustic signal that represents the sounds generated by the heart and recorded from the chest with a device called stethoscope. The mechanical activity of the heart is then described by using this signal, which basically consists in the opening and closure activities of the heart valves. These components are actually gates which control the blood flow. The systematic opening of the valves regulates the blood for traveling through veins and arteries and separates the blood which contains oxygen from the amount of blood that does not (Abbas and Bassam, 2009).

There are four valves located into the heart, the aortic and pulmonary valves control the pass of blood out of the ventricles. The mitral and tricuspid valves lie between the atria and the ventricles and they control the flow between these places. Figure 3 shows the location of the heart valves. As mentioned before, the heart sound signal represents the mechanical activity of the heart, that means the function of the valves. A malfunctioning of any of this gates represents a turbulent blood flow, which is called murmur or rumble. This is in fact a pathological state.

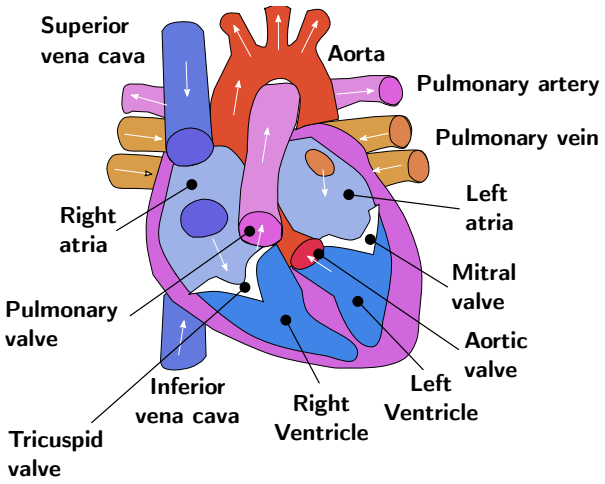

Figure 3. The heart anatomy. Location of heart valves and chambers.

2.2 Characterization of the PCG as an electrical signal

The morphological structure of the PCG signal consists of two main events or fundamental heart sounds (FHS): S1 and S2. Each FHS is directly related to the heart movements of systole and diastole. The first FHS or S1 is a sound which describes the opening of the aortic and pulmonary valves and the closure movements of the mitral and tricuspid valves. This sound is directly associated with the systolic movement, where the ventricles contracts to throw out the blood to the arteries. On the other hand, the second FHS or S2 sound describes the complement of the events that happen in S1: now the mitral and tricuspid valves are open to receive blood while the aortic and pulmonary valves close the conducts. S2 sound has two main components: A2 and P2 produced by the closure of the aortic and pulmonary valves respectively.

This procedure is called the systole and now the auricles contracts to throw out the blood to the ventricles. A junction in time of the two main FHS in sequential order (S1-S2) is called cardiac cycle. Another couple of sounds called S3 and S4 can be also present in a PCG, however, they represent a normal or healthy condition for pediatric PCG signals, while for adults they are actually related to a pathological state.

A frequently used method for the diagnosis and detection of heart diseases is the electrocardiogram (ECG), which can be interpreted as the polarization of all muscular nodes included in the heart. The ECG signal is actually a quasi-periodic electrical signal composed of three main waves: ’P’ wave, ’QRS’ complex and ’T’ wave. The FHS of the PCG signal are related with the ECG waves since the peak in the ’R’ wave and the

ending of 'T' wave are associated with the starting and end time positions of a PCG cycle. Figure 4 depicts this approximation.

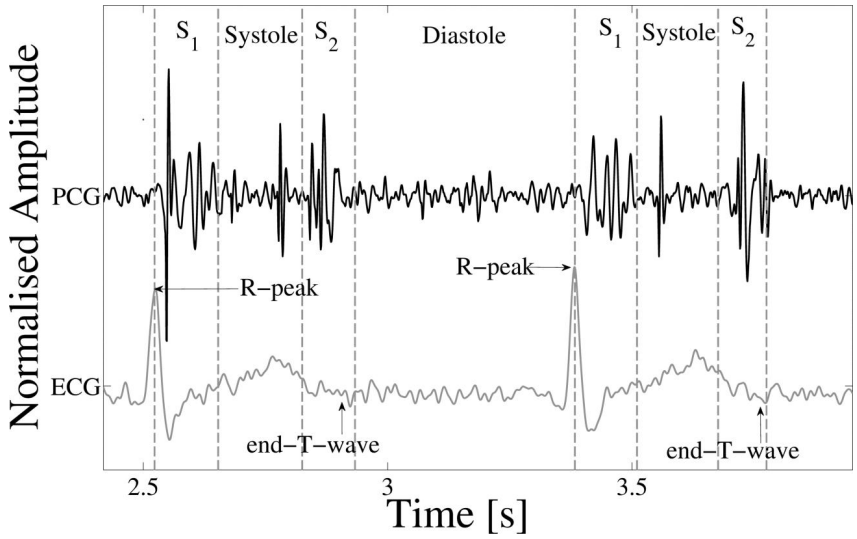

Figure 4. Relation between the ECG and PCG waveforms (Liu *et al.*, 2016).

If a PCG contains only a S1-S2 junction is associated with a normal heart state or a healthy patient. The interval where there is no FHS is called *silence*. Some authors handle the *systolic* and *diastolic* silence intervals according to their location (after S1 or after S2). Figure 5 shows the waveform or time shape of a cardiac cycle of a PCG signal in a normal state. When a malfunction in the heart valves is presented they

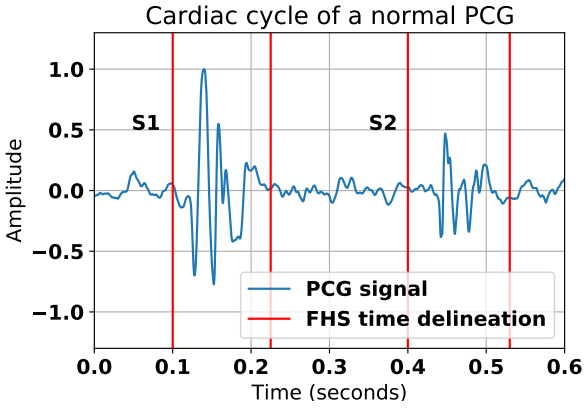

Figure 5. The waveform of a normal cycle extracted from a PCG.

obstruct the blood flow. This effect produces a turbulence which can be seen from the PCG as a murmur or rumble. An abrupt change in the frequency content of the signal can be seen when this pathological state is present. Figure 6 illustrates this effect by showing a cycle of an abnormal PCG. The S1 and S2 waves have typical features in

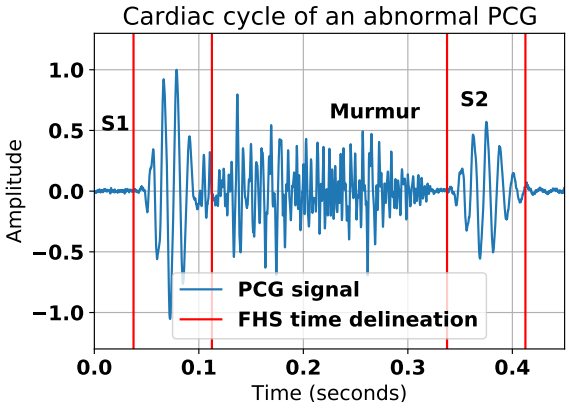

Figure 6. The waveform of an abnormal cycle extracted from a PCG.

time and frequency domains. Although there is no specific time length for each FHS a number of authors define the average time of these intervals. Each FHS has also a specific bandwidth or frequency content, it is generally below the 150 Hz for normal S1 and S2 sounds (Abbas and Bassam, 2009). This information is presented in Table 1. As a complement, Figure 7 displays the time waveforms for both normal and abnormal heart cycles and its associated spectrograms. Notice that as Table 1 indicates, the main frequency content lies below 150 Hz for normal sounds while for abnormal sounds it can reach approximately 500 Hz (where the murmur is situated).

Table 1. The characteristic time length and frequency content of the fundamental heart sounds.

Event	Time length (s)	Frequency content (Hz)
S1	0.1-0.12	20-150
S2	0.08-0.12	50-60
S3	0.04-0.05	20-50
S4	0.04-0.05	<25

2.3 PCG signal modeling state of the art

This section makes a review of previous research done in the area of cardiac audio modeling, focusing on the signal representation in the TF plane. These works can be classified according to the methods used as:

1. *Pole-zero models*: In order to derive frequency domain features of porcine prost-

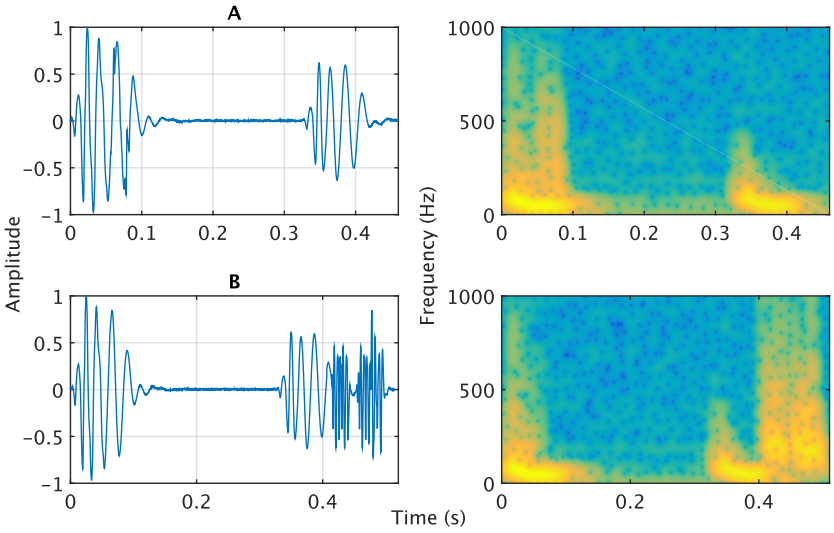

Figure 7. A: The plot in time of a normal cycle and its associated spectrogram, **B:** the waveform of a heart cycle containing a murmur (abnormal state) and its correspondent spectrogram.

hetic heart valves, Joo (Joo *et al.*, 1984) determine to use the heart sound by the Steglitz-McBride pole-zero modeling method. The characteristics of heart sounds from patients were used to determine the state of the valves as normal or pathological. The use of all pole models as LPC have been used also to represent the spectral features of cardiac sounds. Agostihno and Souza (Agostihno and Souza, 1997) took the LPC method to synthesize heart sound signals with a multi-pulse method of filter excitation. However, the choice of the time window represented a problem in the frequency resolution, since they fixed it as rectangular with 30ms of time resolution. The impulse excitation and 5-th order filter were determined as parameters to model the heart sounds. In the research work presented by Redlarsky (Redlarski *et al.*, 2014), a representation of the PCG sounds also by the LPC algorithm was done. The LPC coefficients were coupled in pairs as parameters to a Support Vector Machine classifier in order to detect different heart murmurs. A spectral *fitness* comparison was done to find the appropriate filter order for the phonocardiogram spectral representation. The research found that with a 24-th filter order a 99.7% of spectral fitness can be reached, while with the 5-th order that Agostihno introduces reachs only a 85.5% of spectral matching factor. The all-pole methods presented above aim to find the most appropriate spectral representation of cardiac sounds having a lack of information about time resolution. Although they are easy to compute, it is hard to find coherence in time to fix a

window because of the phonocardiogram is a quasi-stationary signal.

2. *Damped-sinusoidal models*: The reconstruction of heart sounds produced by prosthetic heart valves with damped sinusoids was used by Köymen (Köymen *et al.*, 1987). This research assumes that the rapidly increasing of the valves can be seen as an impulsive excitation, then, a PCG can be reconstructed as a damped-sinusoid sum. The account of the waves parameters such as the amplitude and frequency information of the sinusoids was the first step. Then, dominant peaks of the PCG spectra in the frequency band of 200-500Hz were observed. Another interesting model of the PCG by decaying sinusoids, focusing on the second heart sound was presented by Leung (Leung *et al.*, 1998). In this work, the approach of the Gaussian modulation to decompose the second heart for the diagnosis of pediatric heart diseases was done. The peaks of the short time Fourier transform (STFT) determined the optimal parameters in the sinusoid to best represent the sound. An automatic technique to measure and compare similar patterns was performed by calculating the mean and standard deviation values of the fixed split. The model was also used by Tang (Tang *et al.*, 2010) to find the quasi-stationarity of the heart sound signals and separate the components from noise using fuzzy detection. The same author uses this model in (Tang *et al.*, 2016) to achieve signal compression and transmission over low data rates. A distortion percent of less than 5% (in terms of percent-root-mean-square difference) was obtained by applying the method to a set of signals. Besides, in the calculation results the compresion ratios from 20 to 149 depending on the presence or absence of murmurs (at heavier murmurs lower compression rates). The residue between the chosen sinusoids and the original signal is modeled by vector quantization. Although this model aims to reach a low distortion and a high level of compression rate, the vector quantization method for the residual representation is complex in terms of calculations and requires a previous training stage.

3. *Chirp signal models*: Using transient nonlinear chirp signals, Xu modeled the components of the second heart sound s_2 A2 and P2 (Xu *et al.*, 2001a). Each member of s_2 can be modeled as a mono-component nonlinear chirp signals of short duration, in fact, the sound was considered as modulated in frequency. With this hypothesis, the signal was modeled as a function of polynomial phase. The Wigner-Ville distribution was used to depict the time frequency distribution of the synthesized

signal. The same set of nonlinear chirp signals were used also by the same author (Xu *et al.*, 2001b) to extract the A2 and P2 components of the second heart sound signal. The phonocardiograms were recorded from pigs with pulmonary hypertension. This work was resumed by Djebbari (Djebbari and Bereksi-Reguig, 2011), who created a *Valvular Component Chirp Model (VCCM)* from the chirp wavelet used by Xu as mother for the time scale analysis. Comparing with the Discrete Meyer and the Morlet Wavelets. The cross correlation reached was 98.59% with a polynomial order of 15. In the spectrogram comparison the VCCM model represented more than the 80% of the spectral components for both s_1 and s_2 sounds.

4. *Sparse signal models*: By using time-frequency Gabor waves, Zhang et al. (Zhang *et al.*, 1998) represented the normal heart sounds and murmurs. A set of gaussian modulated sinusoidal waves were selected to make a compact model which aims time-frequency analysis-synthesis of the heart sounds. The application of the MP method to two sets of murmur signals give as a result a normalized-mean square root error (NRMSE) of less than 2.2% while the algorithm reached between 39 to 67 atoms per signal if normal and 125 to 536 in the presence of a murmur or noise. These results reflect that the Matching Pursuit method is suitable to represent the transients and complex structures in the PCG, in this way Nieblas et al. (Nieblas *et al.*, 2013) also used the Matching Pursuit algorithm for a cardiac sound segmentation in different pathological PCG signals. A time frequency distribution was depicted to support the calculation of the segmentation points of normal sounds and murmurs. Other approaches of the application of Matching Pursuit in cardiac sounds that can be founded in literature are the automatic detection of the cardiac cycles from the PCG signals (Sava *et al.*, 1998) and the analysis of the first heart sound for the detection of the activity in the mitral and tricuspid valves (Wang *et al.*, 2004). The framework presented by Jabbari (Jabbari and Ghassemian, 2011) the use of the Multivariate Matching Pursuit (MMP) was employed in order to model heart systolic murmurs. In this variation, the best matched elementary waveforms conformed a new set of functions used to perform again the algorithm. Thus, the combination of consecutive selected elements enabled the model for systolic intervals. This research work concludes that recurrent wavelets can be adapted to reconstruct pathological heart sounds with the aid of a cycle-averaging method.

5. *Wavelet models*: The research done by Martinez Alajarin (Martínez-Alajarín and Ruiz-Merino, 2004) presents the PCG signal compression in blocks. The number of coefficients were selected by the Wavelet Packet transform procedure. An amount of 99.9% of the retained energy can be represented by the WP decomposition when using a Daubechies-8 wavelet. Manikandan (Manikandan and Dandapat, 2007) also presents a compression method based on wavelets using an energy based treshold to retain the significant coefficients. Reed et al. (Reed *et al.*, 2004) applied the wavelet decomposition of heart sounds to construct a prototype of computer aided diagnosis. A neural-network classifier takes the meaningful features of the selected wavelets which could help in heart sound analysis. The Wavelet packet of heart sounds was useful for Liang (Huiying *et al.*, 1997) to perform a segmentation of the sound. Some meaningful components were extracted to separate the first and second heart sound and the systolic-diastolic periods. Normal and abnormal sounds were used to test the algorithm. The algorithm obtained 93% of accuracy after being applied over 77 recordings that contained 1165 cycles.

Although many signal processing techniques have been applied, there is a lack of knowledge in testing the modeling of heart sound signals. Matching Pursuit was presented in this section as a prominent method for the phonocardiogram processing due to the adaptive decomposition of the parts of cardiac sounds. However, authors could not ensure the use of cardiac sounds as prominent diagnostic tools. For this reason, this work evaluates the MP-LPC modeling of PCG signals through objective and subjective tests, considering the opinion given by clinicians and health experts as listeners in the MUSHRA subjective analysis.

2.4 Description of the signal databases used in the research

In the past few years many methods for the automated analysis of heart sounds demonstrated the potential to accurately perform a segmentation and detect a pathological state in clinical applications. However, most researchers have used their own PCG recordings which are not publicly available. As a consequence, there is a lack of high-quality, standardized and rigorously validated databases. On the other hand, the use of open databases to evaluate the automatic analysis of heart sounds allows a

fair comparison of the algorithms in the literature. In the previous years we can find a number of different open databases for the analysis of heart sounds, these are shown in Table 2.

Table 2. The databases considered for this research.

Name	Number of recordings	Normal/ abnormal recordings	Sampling frequency	Available from
Littmann	14	13 abnormal 1 normal	11025 Hz	http://www.litman.com/3M/en_US/littmann-sthetoscopes/
eGeneral medical	38	37 abnormal 1 normal	4000 and 8000 Hz	http://www.egeneralmedical.com/listoheartmumur.html
HSCT11	206	206 normal	11025 Hz	http://www.diit.unicit.it/hsct11/
PhysioNet/CinC Challenge 2016	3153	665 abnormal 2488 normal	2000 Hz	https:www.physionet.org/challenge/2016/

Another two open databases can be also found online, one of them provided by the University of Michigan (MHSDB) (UMHS, n.d.), which contains a total of 23 heart sound recordings. The number, quality and length of the recordings was a reason to avoid this collection of sounds to conduct the experiments in this thesis. Additionally, we have the PASCAL database (Bentley *et al.*, 2011) provides another public database comprising 176 recordings for heart sound segmentation. Although there is a relatively large number of recordings they have a frequency content below 195 Hz due to the application of a low-pass filter, which removes useful components of the heart sound to be used for diagnosis.

In a first stage, we evaluate the diversity of pathological states, such as rumbles or murmurs contained in the collections of signals described in Table 2. We noticed that the Littmann database contains a limited number of recordings, although they have a variety of abnormal sounds presented with a good signal quality. Since its purpose is the biometrical identification of patients, the HSCT 11 sounds database contains only heart sounds of healthy subjects. In this sense, we finally considered for our experiments the eGeneral medical heart sounds database (eGeneral Medical, n.d.) to test the reconstruction of heart sounds by using the proposed MP-LPC method, since it contains a variety of murmurs and abnormal states.

On the other hand, the Physionet/CinC Challenge 2016 (Liu *et al.*, 2016) heart sounds collection will be used for the PCG classification task. We will analyze this dataset and the best-ranked methods used for the challenge in detail in Chapter 3. The experiments conducted in Chapter 4 and Chapter 5 will be tested using this database recordings.

2.5 The reference model for PCGs reconstruction

We consider that the phonocardiographic signal can be represented as the summation of two components: a deterministic and a stochastic part. Let $x_h(t)$ be the deterministic part and $x_n(t)$ the stochastic part so our PCG signal $x(t)$ is represented as follows:

$$x(t) = x_h(t) + x_n(t). \tag{1}$$

In the expression (1), the $x_h(t)$ component is related to the harmonic-like behavior of the PCG and contains the majority of its acoustic energy. For this part, we consider the Matching Pursuit (MP) algorithm to perform an analysis-synthesis procedure. We represent the harmonic components of the signal as a linear combination of elementary waveforms called atoms which belong to a time-frequency dictionary. The second part, $x_n(t)$ also called residual, is obtained as a result of the subtraction between the deterministic signal from the original heart sound recording. We consider that $x_n(t)$ be accurately represented as an autoregressive process using the Linear Predictive Coding (LPC) technique. A block-diagram of the model is illustrated in Figure 8. In the first step, $x(t)$ is decomposed as the linear combination of waveforms taken from a dictionary. Then, the LPC modeling was considered to represent the residual signal. As output, we get parameters for storage, transmission or classification purposes.

The proposed heart sounds model was evaluated by performing subjective and objective tests. We reconstructed different PCG signals that correspond to different pathological cardiac sounds. As an objective evaluation, we considered the metric of percentage of Root-Mean-Square Difference to test the distortion produced by the original heart sound and the reconstructed signal. For the subjective test we conducted a formal methodology for perceptual evaluation of audio quality with the assistance of medical experts following the procedures described in the MUSHRA test (ITU-R, 2001).

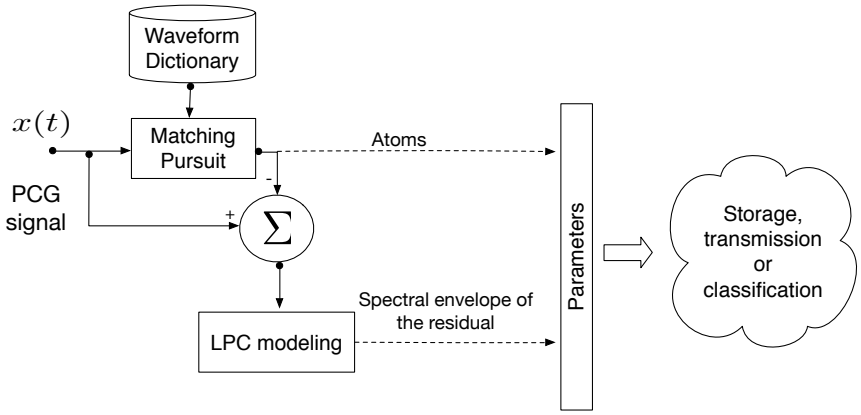

Figure 8. The proposed model for PGC signals reconstruction.

2.5.1. Reconstruction of the deterministic part using The Matching Pursuit method.

In the area of audio signal processing, it is assumed that this natural phenomena commonly have signal components which vary considerably in time and frequency, thus they are nonstationary. In the recent years, sparse representations have become a powerful tool for the analysis and processing of non stationary signals.

A sparse method is a way to accurately approximate a signal $\mathbf{x} \in \mathbb{R}^N$ as a linear combination of the K columns $\mathbf{d}_k$, called atoms, of a matrix $\mathbf{D} \in \mathbb{R}^{N \times K}$, where $N \ll K$, called dictionary. The representation of $\mathbf{x}$ may either be exact $\mathbf{D}\boldsymbol{\alpha} = \mathbf{x}$ or approximate $\mathbf{D}\boldsymbol{\alpha} \approx \mathbf{x}$. The vector $\boldsymbol{\alpha} \in \mathbb{R}^K$ stores the representation coefficients of the signal $\mathbf{x}$ (Chen *et al.*, 2001; Elad, 2010).

Given that $N \ll K$, the dictionary $\mathbf{D}$ is a full-rank, overcomplete and redundant matrix and the representation problem has an infinite number of solutions, therefore, constrains on the solution must be set. The choice of the *sparsest* representation, which is actually the core of the sparse methods, resides on seeking the $\boldsymbol{\alpha}$ vector with the fewest non-zero components. This problem is indeed NP-hard, and its solution can often be found using approximation algorithms.

A sparse representation can also provide a flexible decomposition to reconstruct the signal components whose localizations in time and frequency vary widely. In the fields of signal processing and harmonic analysis we can find a number of applications that use the decomposition of a signal over a family of functions. In known applications

we found this linear representations using a single basis, for example Fourier or wavelet basis. However, Fourier representations are not completely accurate when having functions well localized in time and wavelet representations cannot be well adapted to represent functions whose Fourier transforms have a narrow frequency support. Nevertheless, for non stationary signals it is often necessary to use for sparse representation a dictionary of elementary waveforms called time-frequency atoms. Matching Pursuit (MP) is a greedy-sparse method proposed in (Mallat and Zhang, 1993) for the adaptive decomposition of nonstationary signals using time-frequency dictionaries of atoms. MP in fact, is an approximation method which finds the number of non-zero solutions one at a time.

The MP algorithm optimizes the adaptive approximation of a signal by choosing the atom that *best-match*. In each step or iteration we select an atom by calculating the maximal inner product between the atoms of the dictionary and a signal $\mathbf{x}$. The signal is represented as the linear combination of M selected atoms as:

$$\mathbf{x} = \sum_{m=1}^{M} \alpha_m \cdot \mathbf{g}_{\gamma_m} + \mathbf{R}_M, \tag{2}$$

where $\mathbf{x}$ is the signal to reconstruct, $\mathbf{g}_{\gamma_m}$ is the atom which belongs to the dictionary $\mathscr{D}$ where $\gamma \in \Gamma$, that is selected at the iteration m, having $m = 0, 1, 2, \cdots M$, α_m is the value of the maximal inner product at the m-th and is used for weighting, and $\mathbf{R}_M$ is called the residual signal produced when subtracting the linear combination of the M selected atoms with the original signal $\mathbf{x}$. The whole method is detailed in Algorithm 1.

Algorithm 1 The Matching Pursuit method

Input: $\mathbf{x}$, $\mathscr{D} = \{\mathbf{g}_\gamma, \gamma \in \Gamma\}$
Output: $\alpha_m, \mathbf{g}_{\gamma_m}$
  Initialize: $\mathbf{R} = \mathbf{x}$, $\alpha_m = 0$
  repeat
    $\mathbf{g}_{\gamma_m} = \text{argmax}_{\gamma \in \Gamma} |\langle \mathbf{R}, \mathbf{g}_\gamma \rangle|$
    $\alpha_m = \langle \mathbf{R}, \mathbf{g}_\gamma \rangle$
    $\mathbf{R} = \mathbf{R} - \alpha_m \cdot \mathbf{g}_{\gamma_m}$
  until The number of M iterations or the level of energy desired in the decomposition has been reached.

Looking at Algorithm 1, it is noticed that we may need an infinite number M of atoms to perform an exact decomposition. In this way, the MP algorithm needs a stop criteria. One choice for this issue can be a desired number of M iterations depending on

the memory storage capacity. In fact, the number of atoms cannot be a parameter for measuring the quality of the reconstruction by itself. However, it is possible to calculate a ratio between the amount of energy of the original $d_0 = \|\mathbf{x}\|^2$ and the residual signal $d_M = \|\mathbf{R}_M\|^2$, which can also be known as energy decay [1]. We can easily calculate a percentage of energy decay P as follows:

$$P = \frac{d_M}{d_0} \times 100. \tag{3}$$

We have now determined a stop criteria to choose an adequate number of atoms by evaluating the P percentage.

2.5.2. Selecting a dictionary for the MP decomposition

Selecting an appropriate dictionary for the decomposition of the signal under analysis is a crucial step for the performance of the MP algorithm. The atoms of the dictionary should best represent the PCG signal components. A set of time-frequency waveforms can be generated by the modulation, scaling and translation of a window function $w(t) \in \mathscr{L}^2(\mathbb{R})$. This window has unit energy $\|w(t)\| = 1$, is real and centered at 0.

The family of functions $\mathscr{D} = \{g_\gamma, \gamma \in \Gamma\}$ is very redundant (Mallat, 1999). It will be used to compute a sparse representation of the PCG $\mathbf{x}$. According to the choice of the window we can have different families of functions.

In this work we evaluate the performance of reconstructing heart sounds when using as dictionaries four different mathematical functions that have been proposed in the literature and are considered to successfully represent the characteristics of PCG or other sound signals. Specifically, we selected the functions: Modified Discrete Cosine Transform (MDCT) packets (Ravelli *et al.*, 2008; Fuchs *et al.*, 2015),wavelet packets using Daubechies-6 as mother wavelet (Martínez-Alajarín and Ruiz-Merino, 2004; Huiying *et al.*, 1997), Gabor functions (Ibarra, 2014; Nieblas *et al.*, 2013; Ibarra, 2014; Ibarra *et al.*, 2015; Sava *et al.*, 1998; Wang *et al.*, 2004; Jabbari and Ghassemian, 2011) and Gaussian chirplet functions (Djebbari and Bereksi-Reguig, 2011; Xu *et al.*, 2001a,b). These sets of waveforms are defined by the choice of the window function as follows:

[1] From Algorithm 1, since $R_0 = \mathbf{x}$, let d_0 be the initial energy of the signal and d_M the energy of the residual at the M-th iteration.

- **Gabor atoms** also know as Morlet wavelet atoms, are based on a Gaussian window function:

$$g_\gamma(t) = \frac{1}{\sqrt{s}} w\left(\frac{t-u}{s}\right) e^{i2\pi\xi(t-u)}, \tag{4}$$

where $\gamma = (s,u,\xi)$ is an index of a set $\Gamma = \mathbb{R}^+ \times \mathbb{R}^2$ described by the modulation frequency ξ, scaling s and traslation u parameters. We say that $g_\gamma(t)$ is *well-concentrated* because its energy is mostly concentrated at the neighborhood of u whose size is proportional to s. Gabor functions are based on a Gaussian window function:

$$w(t) = \sqrt[4]{2} e^{-\pi t^2}. \tag{5}$$

- **Chirplet atoms**: they are waves constructed when adding the parameter c as the chirp rate to the modulated Gaussian window defined in (5).

$$g_{(s,u,\xi,c)} = \frac{1}{\sqrt{s}} w\left(\frac{t-u}{s}\right) \exp\left[j\left(\xi(t-u) + \frac{c}{2}(t-u)^2\right)\right], \tag{6}$$

- **MDCT** atoms:

$$g_{p,k}(t) = w_p(t)\sqrt{\frac{1}{M}} \times \cos\left[(t-a_p) + \frac{M+1}{2}\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right], \tag{7}$$

where the analysis is done in frames of M samples and in $k = 0,1,\cdots,M-1$ iterations, whilst $w_p(t)$ is the window function defined as:

$$w_p(t) = -\sin\left[\left((t-a_p)+\frac{1}{2}\right)\frac{\pi}{2M}\right], \tag{8}$$

where a_p is the start of the p-th frame. This window function is complementary in energy, it is $w_p^2(t) + w_p^2(t+N) = 1$.

- **Wavelet** atoms are a family of orthonormal bases of vectors well localized in time and frequency. For a signal of N samples each vector g_γ has the index $\gamma = (j,p,k)$. A discrete window function is dilated by 2^j centered at $2^j(p+1/2)$ and modulated by a sinusoid of frequency $2\pi 2^j(k+1/2)$. The dictionary $\mathcal{D}_{\gamma\in\Gamma}$ contains $(N+1)\log_2(N)$ atoms. In this work we employed a dictionary based on the Daubechies-6 (DB6) as mother wavelet.

2.5.3. Discrete time version of MP

Since the model requires an implementation in discrete time, let us define the methods for this domain. For instance, the expression given in equation 2 will turn into:

$$\mathbf{x} = \sum_{m=1}^{M} \alpha_m \cdot \widehat{\mathbf{g}}_m + \mathbf{r}, \tag{9}$$

where the signal $\mathbf{x}$ to be analyzed and the residual $\mathbf{r}$ have length N and a sampling period T_s. The discrete time-frequency dictionaries employed are composed by the union of J blocks: $\mathscr{D} = \cup_{j=1}^{J} \mathscr{D}_j$ for $j = 1, 2, \cdots, J$. It allows the use of J different time scales, shifting periods and frequency modulations. Since each block has its own predefined set of parameters, the waveform of a discrete-time Gabor atom is the following:

$$\mathbf{g}_{j,n,k}(l) = w_j(lT_s - nT_j) \exp\left(\frac{2i\pi klT_s}{K_j}\right) \text{ for } 1 \le l \le L, \tag{10}$$

where the time location or window shift is defined as nT_j for $n = 0, 1, \cdots N-1$, the window length or scale L_j and is modulated at a frequency k/K_j, K_j is a predefined number of possible frequencies (FFT size), T_s is the sampling period and L the number of time samples. The Matching Pursuit Toolkit (MPTK) provides a fast and flexible implementation in MATLAB for the sparse representation of audio signals using the discrete MP algorithm (Krstulovic and Gribonval, 2006).

Fig. 9 depicts the waveform of a Gabor atom, which is actually a cosine modulated Gaussian window. On the right side of Fig. 9 other atoms are shown in order to see the effect of changing the modulation frequency.

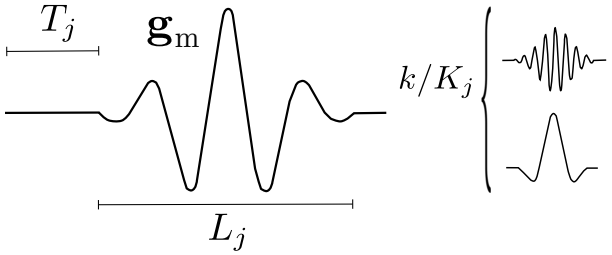

Figure 9. Left: time waveform of a Gabor atom and its predefined parameters. Right: Effect of varying the modulation frequency.

2.5.3.1. Dictionary evaluation for the MP decomposition

We considered MP for reconstruction of heart sounds because it is a method which allows an adaptive time-frequency decomposition of a non-stationary signal with variable components. In order to assess the performance of the dictionaries mentioned before, we conducted a first test, we use the set of the signals included in the eGeneral medical database. We computed the decay in the energy profile in decibels by using the conversion:

$$P_{dB} = 10 \log_{10} \frac{d_M}{d_0},$$

where d_M and d_0 have been defined in the equation (3). The P_{dB} was calculated when iterating the MP decomposition from 1 to 200 atoms. We first divide in cardiac cycles each PCG recording. For each iteration, the resulting decay energy in decibels was averaged in terms of the number of cycles. Then, the resulting P_{dB} was averaged for the whole number of recordings. As a result, we got four different curves shown in Figure 10.

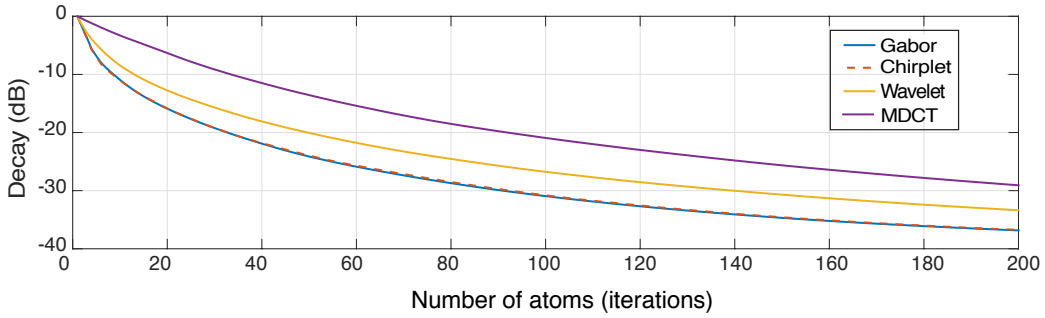

Figure 10. Decay energy ratio in decibels curves using different MP dictionaries.

According to the results depicted by the curves shown in Figure 10, we can see that the MDCT and Wavelet families of atoms exhibit a slower extraction profile, it means that we need a higher number of atoms to extract the most possible energy from the decomposition. In contrast, Gabor and Chirplet dictionaries show the fastest decay profile. Nonetheless, according to the equation 6, chirp atoms require a chirp rate parameter in frequency c, which implies the use of more computational resources. As a conclusion of this evaluation, we decided to use Gabor dictionaries of waveforms, since this atoms displayed a slightly better performance among the other families. A similar evaluation is performed in the research conducted by Gribonval, who compared the performance between the ordinary and chirplet Gabor dictionaries to perform the

MP decomposition of voice and orchestra sounds (Gribonval, 2001).

Hence we have selected the Gabor dictionaries for the MP reconstruction, the next step is to find an adequate of number of M atoms to represent a PCG signal. For this purpose, we developed a second test. The parameter to evaluate was the percentage of energy P described in the expression (3). Following the procedure conducted in the previous test, each recording was segmented in cardiac cycles. Then, we decomposed each of the cycles for all the recordings varying the number of atoms from 1 to 200. At each iteration we calculate the percentage P and then it was averaged for the number of cycles of the PCG signal. Then, the average percentage was calculated for all the recordings included in the eGeneral set. We obtained as a result a curve that exhibits an asymptotic behavior, which indicates that the signal is not exactly sparse but only compressible. Figure 11 displays the results of this test.

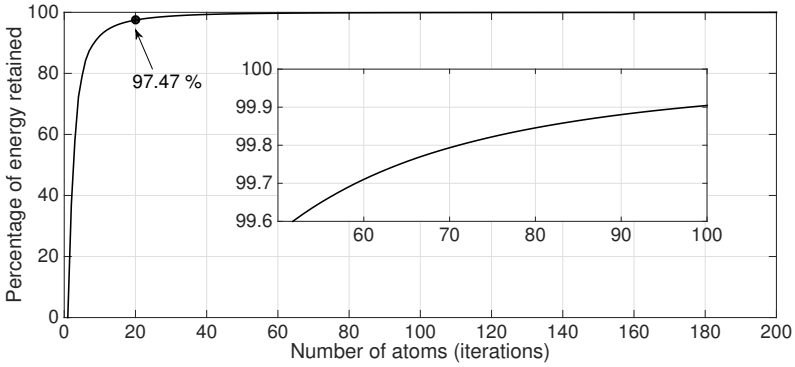

Figure 11. Percentage of energy retained by the MP decomposition of the eGeneral recordings.

The curve displayed in Figure 11 presents an asymptotic behavior. It is due to the greedy nature of the MP algorithm. Fig. 11 includes a zoom box to show that the 100% of the energy will never be attained, regardless of the number of atoms used in the decomposition. However, we considered that when selecting $M = 20$ atoms in average ≈97% of the energy can be represented. This number of atoms will be considered for the next experiments.

2.5.4. Reconstruction of the stochastic part using the LPC coding

In the last section we observed that the signal energy in the time-frequency plane cannot be totally represented by using a finite number of iterations when using the MP

algorithm. In fact, the remaining part or residual signal $R(t)$ turns into a low correlation sequence with respect to the waveforms of the dictionary as the number of iterations increases. We consider that $R(t)$ contain useful components of the signal which are still audible. This component of the signal is $x_n(t)$, it has a *noise-like* behavior due to its low correlation with the harmonic components.

The reconstruction of $x_n(t)$ will enhance the quality of our model, in this sense, instead to model the time waveform we can actually represent it by the envelope of its spectrum. This procedure can be performed when using the Linear Predictive Coding algorithm (LPC), which has been used successfully in the processing of speech signals (Quatieri, 2006). It is assumed that the vocal tract can be modeled as an all-pole filter where the poles can accurately represent the spectral envelope of the signal. The basic principle of LPC is that a sample $\hat{r}_n$ of the residual sequence $\mathbf{r}$ can be predicted as the linear combination of the past p samples as follows:

$$\hat{r}_n = -\sum_{k=1}^{p} h_k r_{n-k} + e_n, \tag{11}$$

where $n = 0, 1, \cdots, N-1$ and the h_k belong to a set of coefficients of a p-th order infinite impulse response filter computed in terms of the minimum mean square error (MMSE) criterion, e_n is an error sequence given by the diference between the original and predicted sequences. In order to find the p values we need to solve the system of equations (called *Yule-Walker equations*) derived when calculating the squared error of (11). In the literature, the Levinson recursion is a commonly used procedure to find the h_k coefficients of the filter (Makhoul, 1975; Vaidyanathan, 2007; Rabiner and Schafer, 1978). Usually the method is conducted by dividing the sequence in frames and then calculate a set of p coefficients for each slice. The reconstruction or synthesis of each frame is performed by exciting the generated filter with a train of pulses separated by a pitch period T_p (voiced sound case) or a white noise sequence (unvoiced sound case).

2.5.4.1. Filter order selection

It is noticed that a finite order of p coefficients for the analysis of the residual frames is needed. It plays an important role for the LPC reconstruction in terms of memory and because of the *whitening* effect of the all-pole filter generated by the LPC coefficients.

This effect is described as a flattening in the output spectrum when p is increased. A measure of the spectral flatness (SFM) in the power spectral density (PSD) $S_{x_n x_n}(e^{j\omega})$ of $x_n(t)$ can be computed (Jayant and Noll, 1984) as:

$$\gamma^2 = \frac{\exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log_e S_{x_n x_n}(e^{j\omega}) d\omega\right]}{\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{x_n x_n}(e^{j\omega}) d\omega} = \frac{\eta_{x_n}^2}{\sigma_{x_n}^2}, \tag{12}$$

where $S_{x_n x_n}(e^{j\omega})$ is considered as a zero-mean process with variance $\sigma_{x_n}^2$. Notice that $\gamma^2 = 1$ if $S_{x_n x_n}(e^{j\omega}) = \sigma_{x_n}^2$, denoting the PSD of a white noise process. Consider also that $0 \leq \gamma^2 \leq 1$, which is an important metric to describe the shape of a PSD using a single value.

We conducted an evaluation in order to find an adequate value for the LPC filter order by using the parameter γ^2. It was calculated and then averaged for the residual signals computed after the MP decomposition of the dataset. Figure 12 shows the value of γ^2 as a function of the filter order (from 1 to 30). We analyzed three different stages: when the residual signal contains 1%, 5% and 10% of the original signal energy after being decomposed by the MP algorithm. In addition, during our analysis we considered both sample rates 4 kHz and 8 kHz. As expected, higher flatness γ^2 values are obtained

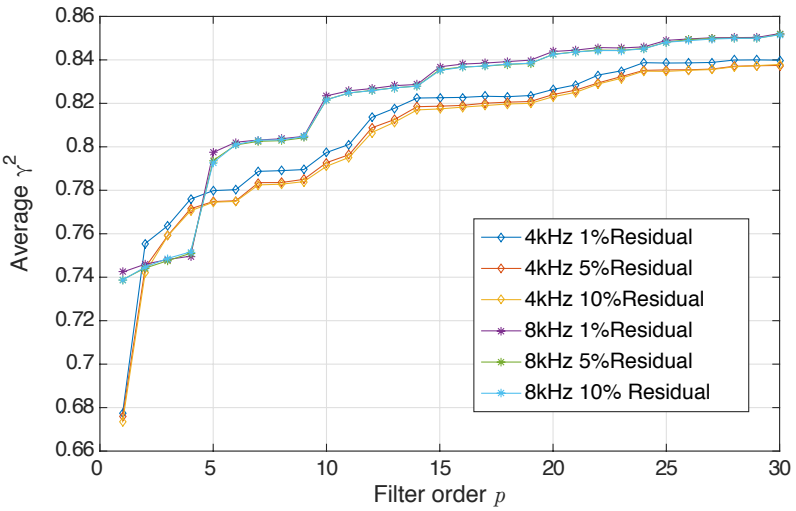

Figure 12. Average spectral flatness measure of the reconstructed residual PCG signals.

as the LPC filter order increases. However, considering that a growth in the number of coefficients affects the dimensionality of the representation and increases the cost in the memory storage, we considered that a filter order of $p = 15$ is enough. In fact, we can observe in the Figure 12 that the flatness shows a rapid increase up to $p = 15$

and then the growth is much less significant. In contrast, the signals sampled at 4 kHz exhibit a slightly smaller flatness levels for all cases.

2.5.5. Objective evaluation of the model

2.5.5.1. Measure of distortion

For the objective assessment of the model we compared the distortion produced when reconstructing the different pathological recordings using our MP-LPC approach. A common used measure for the PCG signals modeling evaluation (Martínez-Alajarín and Ruiz-Merino, 2004; Tang *et al.*, 2016) to find the produced distortion is the percentage of root-square mean difference or PRD, which can be computed as:

$$\mathrm{PRD} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{N}(x_i - \mu_x)^2}} \times 100, \tag{13}$$

where x_i and $\hat{x}_i$ are the original and the reconstructed signals of length N samples. Visual inspection of the original PCG and the reconstructed signal waveforms might not present noticeable differences, resulting in a low PRD. In a visual scale at a cardiac cycle level, usually PRD values below 5% are not easily distinguishable.

Table 3 presents the PRD values calculation when applying the proposed MP+LPC model to the set of signals of the eGeneral database. The average number of atoms per cardiac cycle required to reach 99%, 95% and 90% of the energy of the PCG during the MP decomposition is also shown. It can be noticed that the majority of the recordings required less than 200 atoms per cycle for the reconstruction when using the 99% energy criterion. The first 11 recordings correspond to cardiac sounds with the presence of a murmur during the diastolic phase. In this subset, we can see that recording number 7 corresponding to mitral stenosis has the highest level of distortion with a $\mathrm{PRD} = 7.31\%$. Nonetheless, the PCG recording which requires the highest average number of atoms per cycle was the aortic stenosis for the three criteria. For the normal heart sound case (with no pathological events), the reconstruction presents a distortion of 5.49% and requires an average of 19 atoms per cardiac cycle for the 99% energy criterion. The average PRD for the entire set of reconstructed PCG signals

is 5.00% and the average number of atoms per cardiac cycle is 42.76 for the 99% energy criterion.

2.5.5.2. Graphical quality assessment of the model

We analyze the performance of the model by visual inspection. To see graphically the effects of reconstructing a heart cycle by using the proposed model, we generated the plots of the waveform, the power spectral density and the spectrograms for the original PCG $x(t)$, the residual signal $R_M(t)$ and the reconstructed signal $\hat{x}(t)$. The tunning parameters of the model were a number of $M = 15$ atoms to represent the 99% of the signal represented in two cycles and an order of $p = 15$ filter coefficients for the LPC representation of the residual signal.

All plots were joined in Figure 13, were we observe that the adaptive time-frequency representation given by the proposed model copes with the transitory components of the signal, making it a robust method for modeling abrupt changes in frequency of the murmurs. In detail, Figure 13 **A** shows a similar shape in the time waveforms of the original and reconstructed signals when using the model. In addition, Figure 13 **B** shows a similar morphology between the frequency spectra. Spectrograms shown in 13 **C** and 13 **D** from the original and reconstructed signal are also similar in shape.

2.5.6. Subjective evaluation of the model

The audio distortion measurements like the PRD do not necessarily correlate well with the perceived audio quality of the reconstructed signal. For this reason, we decided to use a methodology for the subjective evaluation of cardiac audio quality. In the present work the subjective sound quality of the reconstructed PCG signal was evaluated by following the ITU-R BS.1534-3 recommendation, better known as MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) (ITU-R, 2001). This methodology has been popular considered and suitable to assess audio quality since it gives accurate and reliable results (Mason, 2002).

The MUSHRA evaluation consists of the following steps: first, the subject is presented with all different processed versions (stimuli) of a single PCG audio (item) at the same time, that is, the stimuli are all simultaneously available (Mason, 2002). Then,

Table 3. PRD results using the proposed MP+LPC reconstruction model on the eGeneral PCG signals dataset.

No.	Name of PCG recording	99%-1% energy		95%-5% energy		90%-10% energy	
		atoms/ cycle	PRD	atoms/ cycle	PRD	atoms/ cycle	PRD
1	Atrial Septal Defect	105	6.39	44	12.66	29	17.06
2	Aortic Regurgitation	177	6.98	77	10.98	48	14.58
3	Aortic Stenosis 2	74	3.61	36	7.56	23	11.06
4	Atrial Gallop	22	8.08	11	16.00	7	22.05
5	Fixed S2 Split	13	3.33	7	10.70	7	14.68
6	Fixed S2 Split 2	20	5.75	10	11.96	7	17.79
7	Mitral Stenosis	153	7.31	54	13.24	36	18.00
8	Physiologic S2 Split	30	3.68	11	8.49	8	13.30
9	Physiologic S2 Split 2	12	3.30	7	10.72	7	14.69
10	S3 Gallop	31	4.39	14	10.20	9	14.41
11	S4 Gallop	35	2.88	12	8.68	8	14.08
12	Summation Gallop	16	4.62	8	10.04	7	14.27
13	Summation Gallop 2	28	6.36	11	14.19	7	21.22
14	Ventricular Gallop S3	24	5.85	13	12.11	9	17.39
15	DiastolicWideS2Split	31	3.70	11	7.82	7	12.83
16	EarlyAorticStenosis	26	4.91	10	10.44	7	14.99
17	Ejection Murmur	75	4.08	30	9.54	19	13.22
18	Ejection Murmur 2	35	3.33	11	7.37	7	10.68
19	Systolic Aortic Stenosis	30	5.00	12	10.70	7	15.74
20	Mild Aortic Stenosis	47	4.54	19	10.05	12	13.35
21	Normal Heart Sound	19	5.49	9	11.87	7	15.79
22	Pericardial Friction Rub 2	31	6.41	13	11.91	8	16.09
23	Systolic Aortic Stenosis 3	90	5.89	38	11.82	25	16.24
24	Systolic Mitral Prolapse	37	5.30	17	11.67	11	15.94
25	Systolic Mitral Prolapse 3	72	5.33	33	11.91	21	16.54
26	Systolic Mitral Regurgitation	57	4.79	19	9.78	8	13.49
27	Pulmonary Stenosis 2	46	3.72	14	8.31	7	13.40
28	Systolic Split S1-S3	14	4.17	8	9.59	7	14.83
29	Diastolic Rumble	41	4.97	20	10.47	13	14.97
30	Ejection Clic	20	5.16	10	10.53	7	15.79
31	Early Systolic Murmur	26	4.56	11	9.74	7	14.97
32	Late Systolic Murmur	31	4.89	12	10.48	8	15.10
33	Opening Snap	23	5.43	12	12.15	8	17.09
34	S3	25	5.41	13	11.67	9	16.73
35	S4	23	5.52	11	11.75	8	16.90
36	Pansystolic Murmur	46	4.09	14	8.59	8	14.02
37	Normal Split S1	19	5.39	9	11.71	7	15.67
38	Normal Split S2	21	5.64	11	12.15	8	16.97
	Average	**42.76**	**5.00**	**17.94**	**10.77**	**11.78**	**15.41**

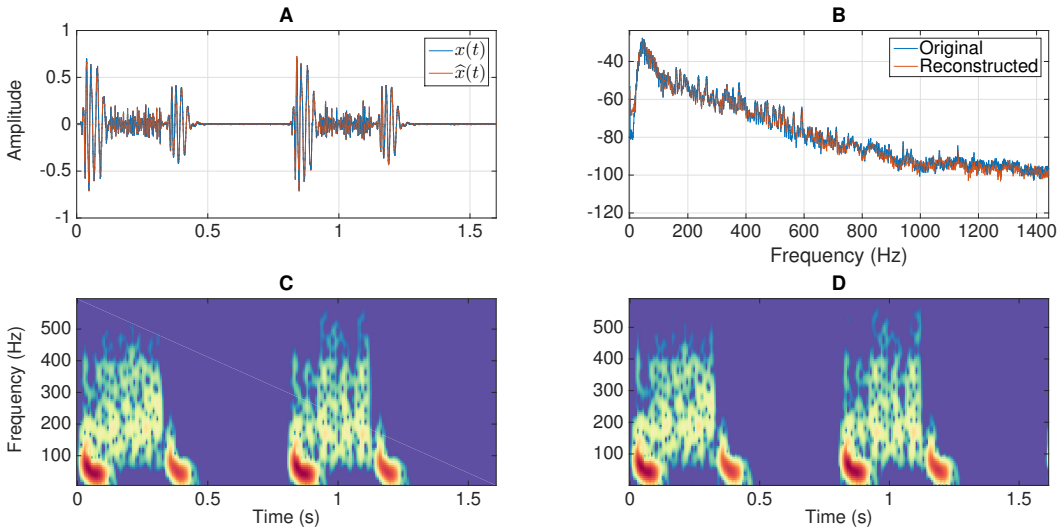

Figure 13. Plots of the waveforms (A), the frequency spectrums (B), and the spectrograms (C,D) of the original and reconstructed signals respectively.

the subject listens to each audio and is required to assign a grade to each stimuli in order to indicate its perceptual quality. The evaluation is performed considering a numerical scale from 0 to 100 with descriptive marks associated with the intervals as indicated in Figure 14. The purpose of using intervals is to have no restrictions about the numerical values assigned, any integer value can be used. A version of the original audio signal is presented as a reference and other codified version as an anchor to ensure that the full range of the scale is used, no matter the quality of the PCG model variant under the test. For broadband audio signals the anchor is a stimuli which corresponds to a low-pass filtered version of the original signal. However, since PCG signal frequency content is predominant in the lower bands (<1 kHz) we decided to use as anchor a noisy version of the original PCG signal. Specifically, white Gaussian noise with a signal to noise ratio of 30 dB was added to the original PCG signal.

Expert listeners are preferred for subjective tests, in our case we require participants with medical background and ample experience in cardiac auscultation. For this research work, a group of faculty members of the School of Medicine at the Xochicalco University in Ensenada, Mexico kindly volunteered to participate in the subjective evaluation of the proposed PCG model, they did not receive any economical compensation for their contribution. The MUSHRA evaluation procedure was explained to all participants and then they signed an informed consent agreement according to the ethical guidelines and principles of the International Declaration of Helsinki. A group

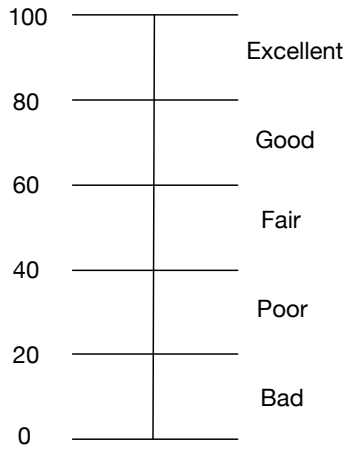

Figure 14. The numerical scale with descriptive terms associated to score the quality of the signals in the MUSHRA test.

of fifteen physicians were selected to participate in the test, 7 female and 8 male. The participants age lies in the interval from 27 to 59 years with an average of 44.53 years with standard deviation of $\pm$13.6 years. All the participants have at least five years of professional practicing experience and eight of them are medical specialists in various fields.

During the evaluation, the MUSHRA recommendation advises to present to the participants a number of audio items of approximately 1.5 times the number of stimuli under test (ITU-R, 2001). In our case, we decided to evaluate four additional stimuli besides the reference and the anchor. The four stimuli evaluated per audio item are the following variants of our MP-LPC based PCG model: three signals sampled at 8 kHz where $x_h(t)$ contains 99%, 95% and 90% of the energy in the original signal respectively. Finally, the fourth stimuli corresponds to a signal sampled at 4 kHz where $x_h(t)$ contains 99% of the energy of the original signal, in all cases the remaining energy is modeled using LPC.

Following the aforementioned advice, during the MUSHRA evaluation we presented ten different audio items to the participants. The PCG audio items selected are the following: 1) aortic regurgitation, 2) fixed S2 Split, 3) ventricular gallop S3, 4) pansystolic murmur, 5) late systolic murmur, 6) early systolic murmur, 7) S3, 8) S4, 9) systolic pulmonary stenosis and 10) normal heart sound. We developed a graphical user interface (GUI) to conduct the MUSHRA evaluation as shown in Figure 15.

The hardware used to reproduce the sounds were professional monitor headphones

Audio-Technica ATH-M50x connected to a MacBook Pro retina display notebook. The evaluation took place in an office room in a quiet environment. In order to allow the participants to familiarize with the GUI and the test, a training stage of two additional audio items was provided. The whole test took in average 17.94 minutes with standard deviation of ± 7.97 minutes.

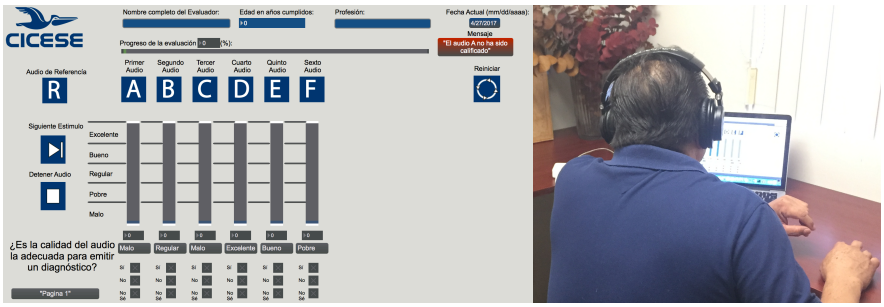

Figure 15. The designed GUI for the MUSHRA test(left). A listener performing the evaluation (right).

During the subjective test the audio items order is fixed, however, the position of the six different stimuli was random. Each listener was asked to score in a scale from 1 to 100 the quality of the reproductions (comparing each one with the reference). The subjects were also asked if the sound quality was adequate for diagnostic purposes, the only available answers to this question were “yes”, “no” and “not sure”. The participant was allowed to reproduce the reference signal as many times as desired until she/he felt confident with the score. A post-screening stage examining all the results was conducted after the test, since some participants can bias the results when they fail to adequately grade the hidden reference with the top score and do not grade the noisy version with the lowest note. In that case, the results of that participant are not considered in the statistical analysis. Under the above mentioned conditions, the results from three participants were not taken into account in the statistical analysis.

The obtained averaged scores of the subjective evaluation are shown in Figure 16, where we follow the statistical analysis methodology proposed in the MUSHRA recommendation. The first step in this procedure is to calculate the mean score for each of the presentations (stimuli) $\bar{u}_{jk}$ as follows:

$$\bar{u}_{jk} = \frac{1}{N} \sum_{i=1}^{N} u_{ijk}, \tag{14}$$

where u_i represents the score of each i participant for a given test condition j and audio sequence k. N is the number of participants. For the presentation of the results

all the mean scores should have an associated confidence interval, derived from the standard deviation and size of each sample $[\bar{u}_{jk}+\delta_{jk}, \bar{u}_{jk}-\delta_{jk}]$, where the value $\delta_{jk} = t_{0.05}(S_{jk}/\sqrt{N})$ is obtained from the t value of 0.05, $t_{0.05}$ and the standard deviation S_{jk} given by:

$$S_{jk} = \sqrt{\sum_{i=1}^{N} \frac{(\bar{u}_{jk} - u_{ijk})}{(N-1)}}. \tag{15}$$

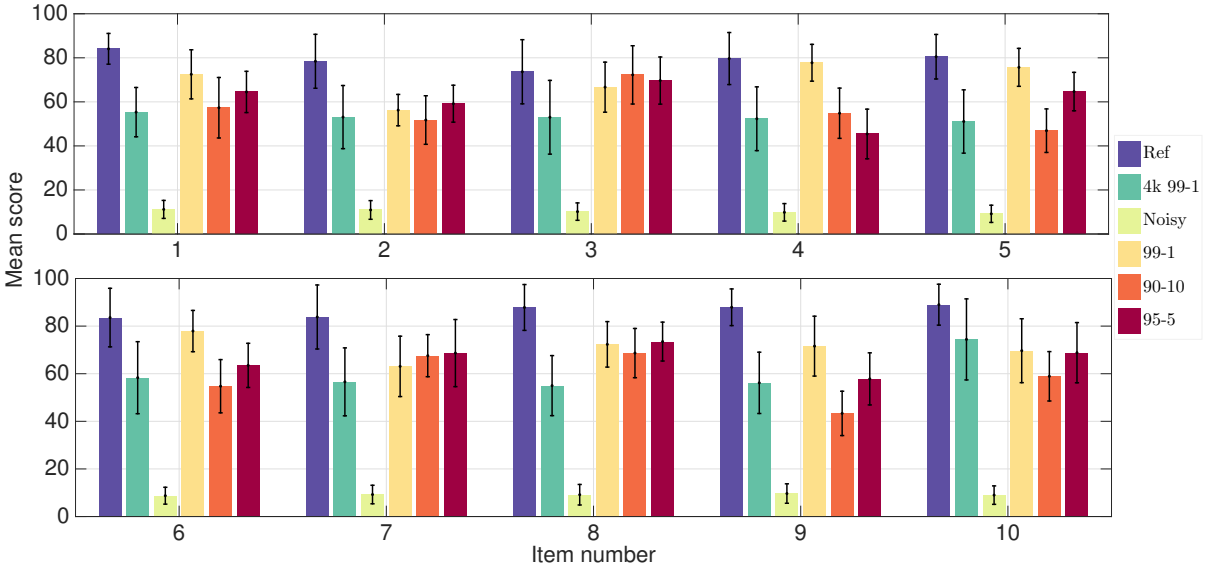

Figure 16. Average MUSHRA scores and error bars associated with the 95% of confidence interval.

For the 10 audio items evaluated, the 95% confidence interval error bars of the reference signal overlap with the error bars of at least one of the variants of the proposed model. In some cases, the score assigned by the participants to the reconstructed PCG signals is practically identical to the score of the reference heart sounds. Among the model variants, the results suggest a better performance for the stimulus reconstructed by 99% of the energy in MP and a sampling rate of 8 kHz. This result reinforces the outcomes of the subjective test since, the mentioned variant also has the lowest PRD. The preference of this variant was in a way expected, since it uses more data to approximate heart sound signals. Nevertheless, it was not expected that for the normal heart sound the listeners would prefer the 4 kHz sampled version with the 99% criterion. After the evaluation we asked every participant their opinion about the reconstructed sounds and in general they found them very realistic. The participants mentioned that when listening carefully to the reconstructed versions sometimes they

perceived artifacts which do not correspond to real sounds produced by the human body. Nevertheless, they also mentioned that the quality of reconstructed heart sounds was adequate for diagnosing purposes.

We have presented in this Chapter a reconstruction model for PCG signals representation based on MP sparse and LPC autoregresive modeling. The model was tested in a PCG recording database which contains a variety of abnormal states. It was rigorously evaluated by conducting subjective and objective tests to examine the quality of signals reconstruction. We are not aware of any other work in the field of PCG signal processing conducted such as the one presented in this thesis.

Chapter 3. Classification of heart sound signals

This chapter gives a general view of the classification of heart sounds to detect abnormalities which may lead to the presence of a pathological state. We start by defining the principal aspects of machine learning as a tool for data analysis to then enclose the concept of classifier. A number of common methods for the classification task are enlisted. Since the methods tested in the next chapters use as training data the Physionet/CinC 2016 sounds database we describe in detail the purposes of this challenge and some interesting results obtained by the participants. We contrast the methods of features extraction and classifiers used to then build our scheme for heart sounds screening.

3.1 Machine Learning and classification

Machine Learning (ML) arises as an intersection of the statistics, artificial intelligence and computer science disciplines. ML is simply about extracting knowledge from data to find patterns to easily take automatic decisions (Müller *et al.*, 2016). In recent years, ML methods have become ubiquitous in everyday life: the automatic recommendations of which movies to watch, to what food to order or which products to buy, personalize our music playlists and the face recognition of our friends or another subjects in different online pictures and the automatic detection of spam e-mails are examples of applications which apply ML methods.

Apart from the commercial applications mentioned before, ML has a huge influence on the way data is managed by researchers today. A number of ML tools have been applied to diverse scientific issues such as understanding stars (Ball *et al.*, 2006), find distant planets, analyze DNA sequences (DePristo *et al.*, 2011) and provide personal treatments for cancer (Kourou *et al.*, 2015; Cruz and Wishart, 2006), apnea (Xie and Minn, 2012) and other medical applications (Polat *et al.*, 2008; Shipp *et al.*, 2002; Dreiseitl *et al.*, 2001).

In the field of ML, we have basically two main types of tasks: *supervised* and *unsupervised*. The main difference between both types reside that in supervised learning we have prior knowledge of what the output values should be according to the input samples. On the other hand, the unsupervised learning has no labeled outputs, so the

point is to infer the patterns present within a set of points. Supervised learning is one of the most commonly used and successful type of machine learning, its main goal is to make predictions to new, never-before seen data after a training stage.

One of the main types of supervised learning is *classification*. It has as goal to predict a class label from a predefined set of possibilities. Thus, we can see the classification task as a discrete supervised learning method. In our research, we look for the screening of abnormal states from heart sounds. We considered a binary classification task, whose labels will be *normal* or *abnormal* according to the absence or presence of murmurs or anomalies in the PCG signal.

Input data for classification is commonly presented as a matrix $\mathbf{X}$ whose target values or class labels are stored in a vector $\mathbf{y}$. The columns of $\mathbf{X}$ are called *instances* and the rows *features* or *attributes*. Dimensionality is given by the number of features N_f presented. In fact, we desire to classify an instance with the aid of some parameters that characterize it. The most common used example as the *hello world* of classification is the iris dataset (Fisher, 1936). The instances are different kinds of flowers whose attributes are the petal length, petal width, sepal length and sepal width (numerical values). As class labels we have the specie which each flower belongs to: iris-setosa, iris-virginica or iris-versicolor. The classification procedure mainly consists of the following stages:

- Split the dataset in *train* and *test* subsets. The train subset will feed a classification algorithm while the test set will be used to make predictions. Consider that train subset contains class labels while test subset do not. Some common applications divide 70%-30% or 80%-20% of the data for the train and test sets respectively.
- Train a classifier method. In this stage, a *prediction model* will be constructed when finding relationships between the parameters of the input and the class label assigned.
- Test the created model when feeding it with the test subset parameters. This model will predict a class label for any input of this *new* data.

The model assessment is performed when comparing the output predicted values $\widehat{y}_i$

with the original class labels y_i of the test set. In binary classification, we have four possible outcomes:

- True positive TP: It happens when the model predicts correctly the positive class. In our case, it will be the right prediction of an abnormal state.
- True negative TN: This outcome happens when the model predicts correctly the negative class. For our research, it will happen when the normal condition is correctly predicted. The accuracy and capacity to predict a murmur will be directly affected when this amount is large.
- False positive FP: It happens when the positive class is incorrectly predicted by the model. In our benchmark, it will happen as an incorrect prediction of an abnormal state, as a consequence, this amount has a negative repercussion in the accuracy and performance of our model.
- False negative FN: This outcome happens when the negative class is incorrectly predicted. This can be traduced as an incorrect prediction of a normal state and will not affect the performance in our application, since the person is healthy.

A common way to report the last mentioned outcome results is the *confusion matrix*, which allows the visulalization of the performance of the algorithm. The rows of the matrix represent the instances in a predicted class while each column represents the instances in an actual class (or viceversa). The confusion matrix for classification in a binary case is displayed in Table 4.

Table 4. Confusion matrix for a binary classification.

	Positive condition	Negative condition
Predicted condition positive	TP	FP
Predicted condition negative	FN	TN

The values of the confusion matrix help to obtain the metrics which describe the model assessment. Most of this evaluation quantities can be described as percentages

or as a values between 0 and 1. The performance metrics that we will use during this research work will be the following:

- **Accuracy:** ACC is the *precision* of the model to correctly detect the positive or negative conditions. The model will be accurate if the actual and predicted values are close to each other. The mathematical expression for accuracy is the following:
$$ACC = \frac{TP + TN}{TP + FN + TN + FP}. \tag{16}$$

- **Sensitivity:** SE which is also called the *true positive rate*, is the measure of the proportion of actual positives which are correctly identified. Mathematically, SE is defined as:
$$SE = \frac{TP}{TP + FN} \times 100. \tag{17}$$

- **Specificity:** SP or true negative rate, measures the proportion of actual samples in negative condition which are correctly identified. It is defined by the following expression:
$$SP = \frac{TN}{TN + FP} \times 100. \tag{18}$$

- **Matthews Correlation Coefficient:** MCC acts like the Pearson Correlation Coefficient to measure the quality in binary classification schemes. It can be interpreted as perfect prediction (if $MCC = 1$), no better than positive prediction (if $MCC = 0$) or a total disagreement between the two classes (when $MCC = -1$). It is computed as follows:
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TN + FP)(TP + FN)(TN + FP)}}. \tag{19}$$

3.2 Classification Methods

Classification is a supervised learning problem, it predicts the category or *class* which a new observation belongs. A classifier needs to know how to characterize and evaluate unlabeled data to perform the predictions, it has a specific set of dynamic rules, which includes an interpretation procedure to handle vague or unknown values, all tailored to the type of inputs being examined during the training phase.

Hence a wide variety of methods have been proposed in the literature for the classification task in supervised learning, the appropriate selection of the classifier technique is an important step in terms of getting the best performance. For this issue, the present work considers to take in account the Occam's razor principle. This criteria helped us to select the first group of algorithms to execute the classification of heart sound signals.

The Occam's razor states that given more than one suitable algorithm with comparable trade-offs, the one that is least complex to deploy and easiest to interpret should be used. According to this principle, we will only enlist those methods considered as less complex in terms of the training stage. Classification algorithms enlisted in the next sections try to find separable patterns in the data, especially linear discriminators and classification trees.

In the classification methods of Machine Learning a common presented issue is the problem of *overfitting*. It happens when the classifier has been over-trained and the generated model starts to make incorrect predictions.

3.2.1. Naive Bayes (NB)

The Naive Bayes classifier assumes independence among predictors. It means that the presence of a particular feature in a class is unrelated to the presence of any other feature (Chan *et al.*, 1982). This model is easy to build, especially in very large datasets. The Bayes theorem provides a way to calculate the posterior probability $P(c|x)$ of a class label or target c given a predictor x, given by:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)},$$

where $P(c)$ is the prior probability of the class, $P(x|c)$ is the likelihood or the probability of the predictor x given the class c and $P(x)$ is the prior probability of the predictor.

3.2.2. K-nearest neighbors (KNN)

KNN is one of the simplest algorithms for classification. It is based on similarity measures, given by a distance function. The measure function is commonly the Euclidean

distance

$$E_i = \|x_i - x_0\|.$$

The training step consists of storing the feature vectors and class labels of the training samples. The classification phase consists in analyzing the closest k number of instances after being saved and then return the most common class as the prediction (Weinberger and Saul, 2009; Cover and Hart, 1967).

3.2.3. Linear discriminant analysis (LDA)

LDA is an algorithm which attempts to express one dependent variable as a linear combination of other features or measurements (Friedman *et al.*, 2001). Considering a set of attributes **X**, and for each sample a related known class **y**. Assume that the k classes have a common covariance matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma} \forall k$. Having these assumptions, the procedure of comparing two classes k and l leads to a linear decision surface, which can be seen by comparing the log-probability rations $\log[P(y = k|\mathbf{X}) / P(y = l|\mathbf{X})]$:

$$\log\left(\frac{P(y=k|\mathbf{X})}{P(y=l|\mathbf{X})}\right) = \log\left(\frac{P(\mathbf{X}|y=k)P(y=k)}{P(\mathbf{X}|y=l)P(y=l)}\right) = 0.$$

Then, the linear discriminant function is defined as:

$$(\mu_k - \mu_l)^T \Sigma^{-1} \mathbf{X} = \frac{1}{2}(\mu_k^T \Sigma^{-1} \mu_k - \mu_l^T \Sigma^{-1} \mu_l) - \log\frac{P(y=k)}{P(y=l)},$$

where μ_k and μ_l correspond to the mean values of the k and l classes.

3.2.4. Logistic regression (LR)

Despite its name, LR is a linear model for classification rather than regression. In this method, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function (Schmidt *et al.*, 2017). This model arises an $\mathscr{L}^2$ optimization problem[1] for binary classes, when the cost function:

$$\min_{\mathbf{w},c} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \log(\exp(-y_i(\mathbf{X}_i^T\mathbf{w} + c)) + 1),$$

[1] Also known as *least-absolute* deviations, it basically consists in minimizing the absolute differences between the true y_i and predicted $\hat{y}_i$ values.

is minimized. In this function, $\mathbf{w} \in \mathbb{R}^p$ is a vector of weights as used in linear regression tasks.

3.2.5. Classification and regression tree (CART)

A decision tree or classification tree is a tree-like diagram in which each internal node is labeled with an input feature. A tree can be learned by splitting the source set into subsets on an attribute value test. This procedure is repeated on each derived subset in a recursive manner called recursive partitioning (Breiman, 2017). The process is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. In CART method, the Gini impurity measures how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled accordig to the distribution of labels in the subset. The Gini impurity for a set of items with J classes is calculated as follows:

$$\mathbf{I}_G(p) = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = 1 - \sum_{i=1}^{J} p_i^2,$$

where $i \in \{1,2,,\cdots J\}$, and let p_i be the fraction of items labeled with the class i in the set and $\sum_{k \neq i} p_k = 1 - p_i$ the probability of a mistake when categorizing that item.

3.2.6. Random Forests (RF)

A Random Forest can be easily defined as a combination of tree predictors. Each tree depends on the values of a random vector Θ sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). The RF classifier consists of a collection of tree structured classifiers $\{h(\mathbf{x}, \Theta_k, k = 1, \cdots)\}$. Each tree casts a unit vote for the most popular class at input **x**. In RF, it is possible to determine the importance of each feature after splitting an instance. This metric can be used for feature selection.

3.2.7. Support Vector Machines (SVM)

In the SVM method a given set of data (binary labelled) is separated with an *hyperplane*, which is maximally distant from them (Cortes and Vapnik, 1995). An input space of N data points $\{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbf{X}$ can be separated by a hyperplane $\mathbf{H} : \mathbf{w}^* \mathbf{X} - b = 0$.

This $\mathbf{H}$ hyperplane is placed by determining two parallell hyperplanes $\mathbf{H}_1, \mathbf{H}_2$ that have maximum margin $2/\|\mathbf{w}\|$ and establishing the condition that there are no data points between them. SVM can use a **kernel trick** to enable the operation in a high dimensional feature space when the inner products between of all the given pairs of data are known. The following functions are commonly used in SVM:

- Gaussian radial basis function: $f(x_i, x_j) = e^{-\|x_i, x_j\|^2/2\sigma^2}$
- Polynomial kernel: $f(x_i, x_j) = (x_i \cdot x_j + m)^p$.

3.3 The Physionet/CinC 2016 challenge

The Physionet research group in collaboration with the Computing in Cardiology 2016 conference (CinC 2016) encouraged researchers for the development of algorithms to classify heart sound recordings. They collected the largest open database of heart sound signals from a variety of clinical and nonclinical environments (Liu *et al.*, 2016). This data was used for the challenge and consists of recordings from eight independent databases sourced from seven different contributing research groups. This subsets are divided by the following groups:

- The Massachusetts Institute of Technology heart sounds database (United States), which comprises 409 recordings from 121 subjects. The recording lengths vary from 9 to 37 seconds.
- The Aalborg University heart sounds database (Denmark), composed by 695 recordings. A total of 151 subjects participated as volunteers. The average length of the recordings is 8 seconds.
- The Aristotle University of Thessaloniki heart sounds database (Greece), which was recorded from 45 subjects. It includes 45 recordings with recording lengths varying from 10 to 122 seconds.
- The K. N. Toosi University of Technology (Iran) heart sounds database, including 174 PCG recordings from 28 healthy subjects and 16 with a diagnosed heart disease. Each recording has an exact length of 15 seconds.

- The University of Haute Alsace heart sounds database (France), composed by 79 recordings: 39 in normal state and 40 in abnormal. The recording length varies from 7 to 17 seconds.
- The Dalian University of Technology heart sounds database (Australia), which includes 338 recordings from 174 healthy volunteers and 335 diagnosed with a CVD. The record length varies from 3 to 98 seconds.
- The Shiraz University heart sounds database (South Korea). It comprises 114 recordings from 79 healthy subjects and 33 healthy/unhealthy patients. The recording length vary from 30 to 60 seconds approximately.
- The Skejby Sygehus Hospital heart sounds database (Denmark), which consists of 35 recordings taken from 12 normal subjects and 23 pathological patients with heart valve defect. The recording length varies approximately from 15 to 69 seconds.
- The Shiraz University fetal heart sounds database (South Korea), which was constructed using 119 recordings made from 109 pregnant women. The data includes individual recordings from 99 subjects, two recordings from three subjects and seven cases of twins that were recorded individually. The average duration of each record was about 90 seconds. Since the challenge was focused on adult heart sound signals, this database was not included. However, it is available in the online Physionet dataset.

As described in Table 2, the distribution of recordings is 665 labeled as abnormal and 2488 as normal[2], since it was taken from unhealthy and healthy patients. Thus, dataset is unbalanced. A *blind* dataset was not released to the public to help the sponsors testing the submitted methods. Another thing to consider is that some recordings are under noisy conditions, which provides authenticity to the challenge. Participants were encouraged to submit their algorithms to the Physionet dataset and then present their research works in the Computing in Cardiology 2016 (CinC) conference. The sponsors of the challenge included a benckmark of heart sounds classification based on features extracted from the segmentation method proposed by Springer (Springer *et al.*, 2016).

[2]Originally, 4430 recordings were registered in the database (as described in (Clifford *et al.*, 2016)). However, during the challenge, sponsors suggested to remove some recordings under high noisy conditions. After taking out this audios we conserved the 3,153 employed by this study.

The output annotations from the time delineation of heart sounds using the Springer method were uploaded. After the challenge took place, sponsors uploaded a hand-corrected annotations of the signals segmentation. Participants of the Physionet/CinC 2016 Challenge were encouraged in 2017 to improve their classification methods in order to be published in the Physiological Measurement journal special issue. The scoring procedure was based on the number of recordings classified as normal, abnormal or unsure and the quality of the signals as noisy or clean. Table 5 shows the possible outcomes.

Table 5. Physionet/CinC 2016 challenge metrics, where A: Abnormal, U: unsure, N: normal.

Reference label	Weights	Entry's output		
		A(1)	U(0)	N(-1)
A, clean	wa_1	Aa_1	Aq_1	An_1
A, noisy	wa_2	Aa_2	Aq_2	An_2
N, clean	wn_1	Na_1	Nq_1	Nn_1
N, noisy	wn_2	Na_2	Nq_2	Nn_2

According to these values, it was defined a modified sensitivity Se' and specificity Sp', defined as:

$$Se' = \frac{wa_1 \cdot Aa_1}{Aa_1 + Aq_1 + An_1} + \frac{wa_2 \cdot (Aa_2 + Aq_2)}{Aa_2 + Aq_2 + An_2}$$

$$Sp' = \frac{wn_1 \cdot Nn_1}{Na_1 + Nq_1 + Nn_1} + \frac{wn_2 \cdot (Nn_2 + Nq_2)}{Na_2 + Nq_2 + Nn_2},$$

where wa_1 and wa_2, also described in Table 5, represent the percentages of good signal quality and poor signal quality recordings in all abnormal recordings, respectively. As a complement, wn_1 and wn_2 are the percentages of good and poor signal quality in all normal recordings. The global metric for ranking the classification benchmarks is the overall score or M_Acc, which is defined as:

$$MAcc = \frac{Se' + Sp'}{2}.$$

A total of 348 entries were submitted in the challenge by 48 teams. Results from the conference and the scientific journal publication will be discussed in the next section.

3.4 Literature review of cardiac sounds classification

The top scored ranked by M_Acc entries of the Physionet/CinC 2016 challenge will be discussed in this section. We will put forward also the highest scored methods of the journal Physiological Measurement according to the *Recent Advances in Heart Sounds Analysis* special issue.

The entry proposed by Potes et al (Potes *et al.*, 2016), obtained the highest overall score $MAcc = 86.02$ having also the highest sensitivity $Se' = 96.33$. Nonetheless, the obtained specificity was relatively low $Sp' = 77.81$ The conducted methodology was based on the use of a voting scheme between an AdaBoost classifier and a convolutional neural network (CNN). A total of 124 features in time and frequency domains were extracted, including Mel-Frequency Cepstral Coefficients (MFCC), diastolic and systolic time lengths from segmentation, and ratios of the amplitudes. The median power of nine frequency bands was also used as input feature.

Zabihi et al (Zabihi *et al.*, 2016), obtained the second place in the challenge ranking with a $MAcc = 85.90$, $Se' = 86.91$ and $Sp' = 84.90$. A set of 40 features in the time, frequency and time-frequency domain were extracted, including LPC coefficients, MFCC coefficients, entropy and wavelet (Daubechies 4) based features. Then, a subset of 18 features was selected by using a wrapper-based feature selection scheme. For the classification task, an ensemble of 20 feedforward Artificial Neural Networks (ANNs) with two hidden layers in each, and 25 hidden neurons at each layer were used.

The third ranked method was developed in (Kay and Agarwal, 2016), and obtained scores of $MAcc = 85.2$, $Se' = 87.43$ $Sp' = 82.97$. The features extracted were the MFCCs, Morlet wavelet based features, and signal complexity (spectral entropy of time lengths of beats and cycles). Dimensionality was then reduced to 70 features by using a student-t statistical test followed by a Principal Component Analysis (PCA) reduction. The classification was performed via a fully connected two-hidden-layer neural network trained by error back-propagation. The authors included an updated version of their algorithm in the special issue (Kay and Agarwal, 2017). They excluded the “training-e” set for training since the recording sensor type for training-e set is different from others. Notwithstanding, they obtained a significantly worse score of $MAcc = 58.0$. A reason of this decrease is because 69% of recordings in the test set are from dataset-e. It indicates

that the algorithm is sensitive to the recording type and struggles to generalize from one dataset to another.

On the other hand, in (Abdollahpur *et al.*, 2016) it is proposed a *cycle-quality assesment* method to examine the quality of the input PCG signals. The features were extracted only in the cycles with higher signal quality and reliable segmentation[3]. A feature set of 90 elements was obtained. Among the features extracted we find the mean values of the energy, the mean and standard deviation values of the cycles and beat lengths, wavelet entropy and wavelet decomposition based features (using a Daubechies 2 mother wavelet) and the mean values of the MFCC coefficients. The method selected for classification is based in a voting scheme of tree neural network classifiers. The method obtained scores of $Se' = 81$, $Sp' = 87$ and $MAcc = 84$. The authors submitted an updated version of the algorithm (Abdollahpur *et al.*, 2017) obtaining a scores of $Se' = 76.96$, $Sp' = 88.31$ and $MAcc = 82.60$.

The Physionet/CinC challenge provided a global open assessment of heart sound algorithms for classification. It gives to the research community the potential benefits of well-characterized heart sound and comparing their methods for PCG signals automated analysis. However, when analyzing the results between the open and blind datasets the scores were not as high as in the open datasets.

Among the observations made by the sponsor researchers (Clifford *et al.*, 2017), the classifiers performance is proportional to the recording sources and requires improvement. It was shown that there is not a best *high-quality* classifier, since among the top six entrants a 2% of variation is presented in the scores. The features extraction stage is crucial for the classification improvement in the scores. Among the most relevant features extracted in the submitted algorithms we find the MFCC coefficients, wavelet-based and time- frequency parameters. They also indicated in these work that it is more important to focus on the labelling and features extraction preprocessing than on the classifier.

In the research conducted by (Ismail *et al.*, 2018), a wide revision of methods for heart beat localization and classification is described. The paper concludes that the most challenging scenario is the analysis of noisy signals, where time-frequency, and

[3]The none corrected by hand Springer annotations

wavelet methods have been popular choices for researchers. Another challenge problem will be the real time implementation of practical applications of heart sound screening methods, this issue remarks the importance of having computationally efficient procedures for the screening of heart sounds. It is also mentioned that the Morlet (also known as Gabor) wavelet has been widely investigated for the modeling of PGC signals.

The methodology adopted in this thesis led to more consideration given to the evaluation of features extraction, feature selection and classification methods for heart sounds screening rather than to the design of a sophisticated classification benchmark of a good performance. We also consider to assess the use of methods to equalize the classes of the data.

Chapter 4. Development of a heart sounds classification benchmark

This Chapter contrasts the methods for features selection and balancing stages of the classification of heart sounds. Two main approaches will be considered for features extraction: *feature averaging* and *cycle averaging*. On the other hand, we will analyze the use of the Synthetic Minority Oversampling Technique (SMOTE) to equalize the number of elements of normal and abnormal classes when creating synthetic samples (Chawla *et al.*, 2002). The features sets will be constructed using the model described in Chapter 2 and for the assessment a 10-fold cross validation stratified test will be conducted. The performance of classification will be evaluated using the classifiers defined in Chapter 3.

4.1 Features settings

According to the model presented in Chapter 2, the MP decomposition of a PCG cycle can be accurately represented by Gabor atoms. Since these waveforms are modulated cosines by a Gaussian window, they have the following parameters:

- Amplitude
- Modulation frequency
- Time shifting or position
- Window length or time scale
- Phase of the cosine waveform

We found from the proposed model that, in order to represent more than 90% of the signal energy, it requires at least 15 atoms. In this way we start having $15 \times 5 = 75$ parameters to represent a cycle by using this approach. On the other hand, we analyzed from Figure 12, that a number of $p = 15$ LPC coefficients is enough to reach a reasonable level of flatness in the spectrum. In such a way, adding the MP and LPC parameters we have 90 features to represent each cycle of the signal.

The recording lengths in the Physionet/CinC challenge database vary with a spread of more than 10 seconds, as a consequence, the number of cycles is also variable. However, in the classification task, we defined $\mathbf{X}$ as the input matrix with a number of columns equal to the number of instances (in this case, PCG recordings) and a number of rows equal to the number of features or attributes (extracted parameters from each PCG cycle). Thus, the number of features for all the instances must be the same.

To establish an equal number of features for each PCG recording instance, we propose two approaches: *feature averaging* and *cycles averaging*. Let **A** and **B** be the features sets which came out from each one of this approaches. Both methods share the following stages:

- **Cycle segmentation**: It is basically the time delineation of cycles from each heart sound. For this purpose we took the corrected annotations from the Physionet/CinC challenge database. Original annotations came as an output from Springer's method (Springer *et al.*, 2016) and then, they were hand-corrected by the sponsors.
- **Matching pursuit (MP) decomposition**: It performs the time-frequency representation of a cycle or a FHS (Mallat and Zhang, 1993). Features from MP are given by the parameters of the selected atoms from MP algorithm when using a Gabor dictionary.
- **Linear predictive coding (LPC)**: Since the MP reconstruction did not provided us a perfect reconstruction of the PCG signal, we considered to model the residual between the MP reconstructed signal and the original PCG, called the residual. LPC models this residual in a compact form.

Fig. 17 describes the dataset **A** using the *feature averaging* approach. The output feature set **A** is actually obtained from the mean value of the MP and LPC parameters for the N segmented PCG cycles, getting a total of 90 features for representing each recording.

All the revised classification benchmarks for the Physionet/CinC challenge used this method. Mean and standard deviation values of time, frequency and time-frequency features were considered as inputs for the classifier. A main difference between the

conducted methodology and the Physionet challenge entries is that we are not considering the *diastolic* states, which are the segments of time after the S2 sound and before the next S1 when there is an absence of a FHS. However, the *systolic* state between S1 and S2 is considered within the heart cycle division [1].

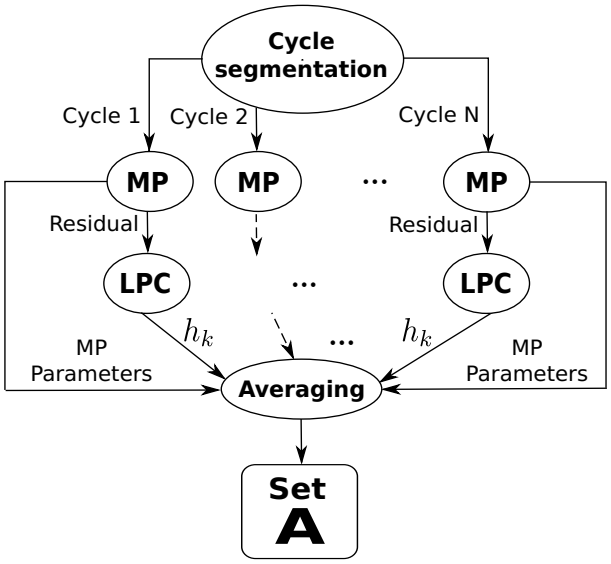

Figure 17. Methodology to obtain the feature averaging set **A**.

On the other hand, Figure 18 depicts the diagram of the method which uses a *cycle averaging* approach. MP and LPC parameters are extracted from an averaged cycle. By contrast with the first method (in which parameters are directly extracted from the whole cycle), in this second method, the averaged cycle is split into the two FHS. MP is then applied on each FHS separately, and two frequency parameters $F1$ and $F2$ are kept for each of them. This feature set arises by conducting a previous methodology described in (Wang *et al.*, 2001, 2004).

4.1.1. Pre-processing of signals

The PCG recordings were filtered by using a band-pass Butterworth sixth-order filter with cut-off frequencies of 25 Hz and 600 Hz. This filtering stage was executed in order to remove non informative frequencies from the input PCG recording. The original sampling rate of 2 kHz of the recordings was preserved. Figure 19 depicts the plot for the band-pass filter frequenct response.

[1] According to the segmentation section in the Phyisionet database paper (Liu *et al.*, 2016) four states are considered: S1, systole, S2 and diastole.

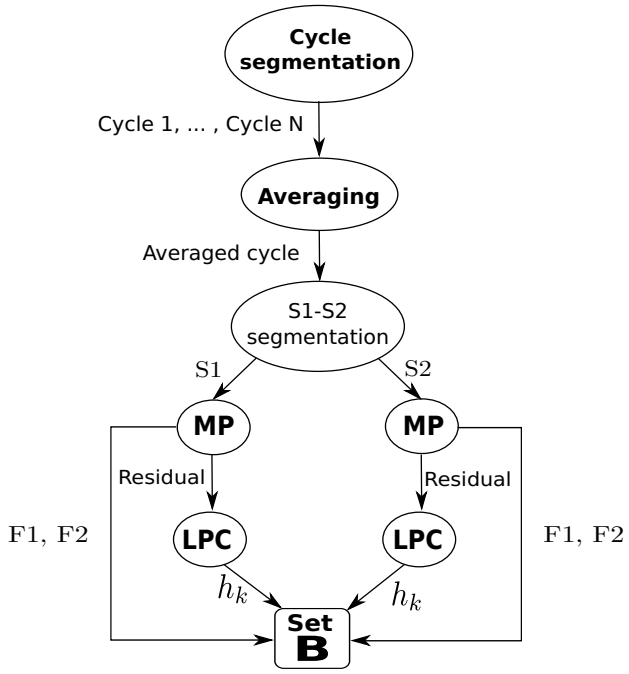

Figure 18. Methodology to obtain the feature averaging set **B**.

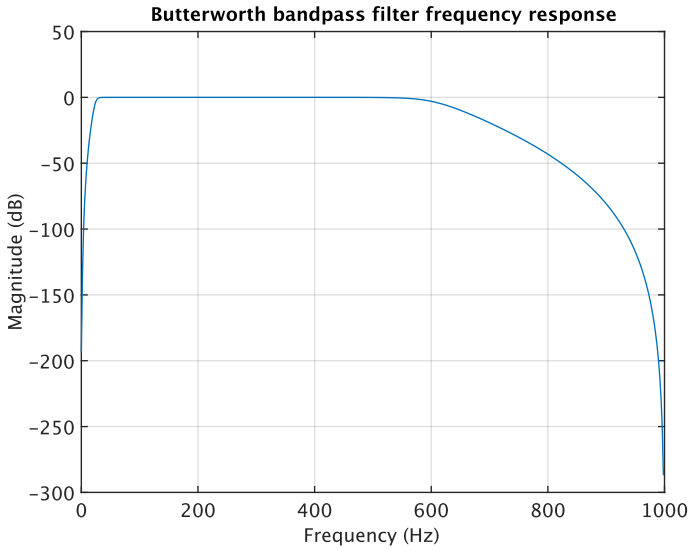

Figure 19. Frequency response of the Butterworth band-pass filter for the PCG signals pre-processing.

4.1.2. MP settings

The MP decomposition of PCG signals was configured by using the Matching Pursuit Toolkit library (MPTK) designed for MATLAB (Krstulovic and Gribonval, 2006). We used a Gabor dictionary of $J = 5$ blocks, each block corresponding to a common atom (or window) lengths L_j of 32, 64, 128, 256 and 512 samples. The selected number of atoms was $M = 15$ in order to reach almost a 99% of the energy to reconstruct a PCG

cycle (Ibarra, 2014). The atom parameters of frequency, amplitude, length, position (shift) and phase were extracted to be considered as features in features set **A**, while in features set **B** just two atom frequencies were considered, as in Wang's methodology (Wang *et al.*, 2001, 2004).

4.1.3. LPC settings

For the LPC representation of the residual signal output by MP, we used the MATLAB code from UCLA available on-line (Ozun *et al.*, 2017). The selected order of the filter was $p = 15$. Processing was performed in frames of 32 ms.

4.1.4. Number of features

Both feature sets **A** and **B** do not have the same number of samples. Combining the MP and LPC methods, features set **A** consists of $N_{features} = 90$ resulting from 75 parameters from MP (15 atoms with 5 parameters each one) and 15 from LPC. This choice is based on our previous work (Ibarra *et al.*, 2015). On the other hand, features set **B** contains $N_{features} = 19$, among which 4 are provided by MP (frequencies of the first two selected atoms for each FHS). This setting, in addition to the rest of set **B** extraction scheme (FHS segmentation and feature averaging), is in line with state-of-the-art (Wang *et al.*, 2001, 2004). Thus, set **B** acts as a comparison baseline for set **A**, whose performance in heart sound classification has not been assessed so far.

4.2 Classifiers settings

We tested the classification state of our proposed pipeline using seven different methods. Table 6 presents a brief description of each one by its name, acronym and the main parameters employed. Classifiers were configured according to the settings depicted in Table 6. Each one of these methods was implemented by using the scikit-learn toolbox of Python (Pedregosa *et al.*, 2011). In addition, feature sets were normalized to have zero mean and unit variance when using SVM. For the RF method we changed the number of estimators to 100 as some authors recommend in the presence of unbalanced problems (Vercio *et al.*, 2017; Mellor *et al.*, 2015).

Table 6. Brief description of the tested classifiers in this work and its parameters tuning.

Acronym	Full name	Main parameters	Value of parameters
CART	Classification and regression tree	Criterion to measure the quality of each split, spitter, strategy to choose the split at each node	Gini criterion, best split between trees.
KNN	K-nearest neighbors	Number of neighbors, weight function used in prediction, metric of distance to use for the tree	5 neighbors, uniform weight, Minkowski's distance.
LDA	Linear discriminant analysis	Name of solver to reduce data dimensions, tolerance value threshold for rank estimation in the solver	Solver: SVD, tolerance of 0.0001
LR	Logistic regression	Penalty norm, number of iterations	L2 norm penalty, 100 iterations
NB	Naive Bayes	Prior probabilities in the classes to adjust the data	No prior probabilities value
RF	Random forest	Number of estimators, criterion to measure the quality of the split, use or not bootstrap to build trees.	100 estimators, Gini criterion, bootstrap: true
SVM	Support vector machines	Penalty parameter C of the error term, kernel function used in the algorithm, kernel coefficient gamma	$C = 2.37$ for unbalanced data and $C = 1$ for balanced, radial basis function as kernel with automatic gamma parameter

4.3 Noisy recordings settings

A number of the recordings included in the Physionet/CinC database do not contain any annotation file to perform segmentation, due to their noise level[2]. To handle these files, we performed the processing of the signals in frames of length of 900 ms (with respect to the approximate duration of a cycle) and 200 ms (according to the typical length of a FHS).

4.4 The SMOTE balancing method

A number of clinical-real-world applications imply the use of highly unbalanced datasets. In our case, we have an unbalanced dataset where *normal* and *abnormal* ca-

[2]These were found as *low quality*, actually.

tegories are not equally represented. The *abnormal* samples are the minority class, covering only 21.09% of the total of recordings in the features set. The works submitted during the Physionet/CinC challenge dealed with this constrain by selecting a random subset from the majority class which has an equal amount of elements as the number of *abnormal* labeled recordings.

The SMOTE method (Chawla *et al.*, 2002) proposes an approach to the construction of classifiers from imbalanced datasets. This method reported a better classification performance when over-sampling the minority class (testing the ROC space). SMOTE proposes to randomly create synthetic minority class examples under the features space.

Depending upon the amount of over-sampling required, the k nearest neighbors are randomly chosen. Then, the creation of the synthetic samples is performed by taking the difference between an input features vector of the minority class and its nearest neighbor. This difference is multiplied by a random number between 0 and 1 and then added to the feature vector under consideration.

4.5 10-fold cross validation test

We compared the classifiers and features sets performance using a cross validation test. The procedure of this method is to make random partitions of data (folds) into complementary subsets (test and train). The classification step is executed, then, the SE, SP, ACC and MCC metrics are the resulting outputs used to evaluate the prediction at each fold. We used 10 folds in our approach. Figure 20 depicts a diagram of the 10-Fold cross validation scheme to evaluate the heart sound classification schemes.

4.6 Numerical tests

4.6.1. Features sets testing without balancing

The 10-Fold CV test was performed under scikit-learn. For the first test, data is not balanced. Table 7 shows the SE, SP, ACC and MCC average and standard deviation output metrics after evaluating each one of the enlisted classifiers combined with a features extraction set approach. The higher values of the obtained metrics are highlighted in bold-type fonts.

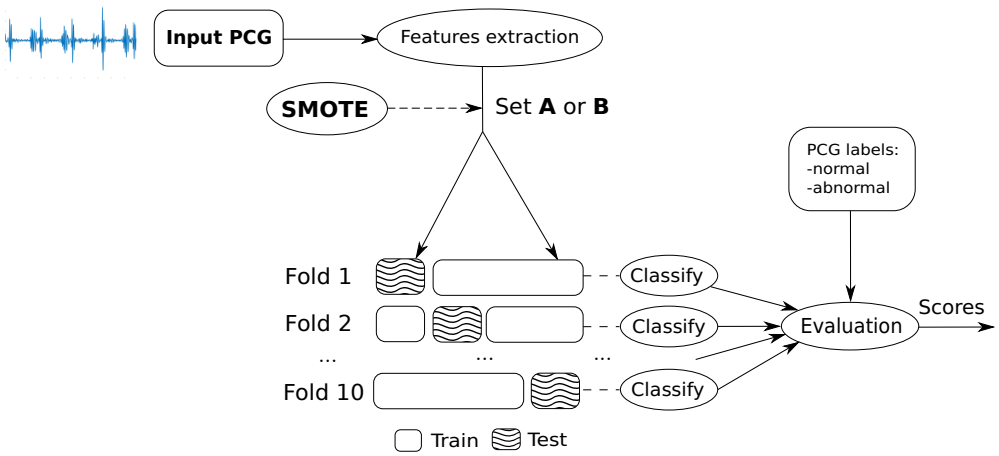

Figure 20. The 10 fold cross validation scheme to assess the quality of heart sounds classification schemes.

In terms of SP, the SVM method reached the highest score of 92.29% when using features set **A**. However, the SE score reached by this combination is the lowest, even the MCC is not the highest presented. The RF method reached the highest values for the remaining metrics when using features set **A**: $SE = 75.80$, $MCC = 0.55$ and $ACC = 0.86$. Results from Table 7 present also greater standard deviation values for SE. We observe that for all feature sets and all classifiers, SP is considerably higher than SP.

Table 7. Performance metrics resulting for the heart sounds cross validation test without balancing.

Model		Average performance			
Classifier	Features set	SE	SP	ACC	MCC
CART	A	50.19 ± 6.33	87.30 ± 1.91	0.79 ± 0.03	0.38 ± 0.08
KNN	A	48.26 ± 4.85	86.36 ± 1.21	0.78 ± 0.02	0.35 ± 0.06
LDA	A	58.32 ± 6.23	83.36 ± 0.86	0.81 ± 0.01	0.31 ± 0.05
LR	A	58.44 ± 6.97	83.36 ± 0.95	0.81 ± 0.02	0.32 ± 0.06
NB	A	49.56 ± 10.41	80.84 ± 0.68	0.79 ± 0.02	0.18 ± 0.06
RF	A	**75.80 ± 5.32**	88.03 ± 1.32	**0.86 ± 0.02**	**0.55 ± 0.06**
SVM	A	46.41 ± 1.99	**92.29 ± 0.76**	0.76 ± 0.02	0.45 ± 0.03
CART	B	42.70 ± 3.75	**84.76 ± 1.02**	0.76 ± 0.02	0.28 ± 0.05
KNN	B	60.63 ± 3.82	84.54 ± 0.85	**0.82 ± 0.01**	**0.36 ± 0.04**
LDA	B	48.11 ± 10.78	80.65 ± 0.76	0.79 ± 0.01	0.17 ± 0.07
LR	B	54.31 ± 12.73	80.43 ± 0.88	0.79 ± 0.01	0.17 ± 0.08
NB	B	58.59 ± 12.43	80.93 ± 0.68	0.80 ± 0.01	0.21 ± 0.07
RF	B	**72.74 ± 8.88**	83.17 ± 1.11	0.82 ± 0.02	0.36 ± 0.08
SVM	B	34.95 ± 4.69	81.91 ± 0.90	0.73 ± 0.02	0.16 ± 0.05

4.6.2. Features sets testing using SMOTE balancing

In a second test, we performed a balancing applying the SMOTE technique when oversampling the minority class. The SMOTE library included in the imbalanced-learn toolbox of Python (Lemaître *et al.*, 2017) was used for this purpose. Then, we conducted again the stratified 10-fold CV test to evaluate the classifiers performance. Table 8 shows the outcome average metrics of this experiment. Higher scores are highlighted in bold-type fonts. Compared to the metrics obtained without applying SMOTE, the

Table 8. Performance metrics resulting for our heart sounds cross validation test when using SMOTE balancing.

Model		Average performance			
Classifier	Features set	SE	SP	ACC	MCC
CART	A	82.61 ± 1.60	84.98 ± 1.85	0.84 ± 0.02	0.68 ± 0.03
KNN	A	70.74 ± 1.39	**97.10 ± 1.63**	0.79 ± 0.02	0.62 ± 0.03
LDA	A	76.95 ± 1.27	82.37 ± 2.14	0.79 ± 0.01	0.59 ± 0.03
LR	A	78.60 ± 1.60	82.38 ± 2.50	0.80 ± 0.02	0.61 ± 0.04
NB	A	84.54 ± 4.58	53.29 ± 0.63	0.56 ± 0.01	0.21 ± 0.03
RF	A	**91.60 ± 1.77**	92.10 ± 1.70	**0.92 ± 0.01**	**0.84 ± 0.03**
SVM	A	77.20 ± 1.09	78.77 ± 1.99	0.78 ± 0.01	0.56 ± 0.03
CART	B	80.71 ± 2.43	82.91 ± 1.82	0.82 ± 0.02	0.64 ± 0.03
KNN	B	79.24 ± 1.56	**95.87 ± 0.79**	0.86 ± 0.01	0.73 ± 0.02
LDA	B	70.55 ± 1.52	70.89 ± 2.44	0.71 ± 0.02	0.41 ± 0.04
LR	B	69.90 ± 1.79	67.97 ± 1.94	0.69 ± 0.02	0.38 ± 0.04
NB	B	82.80 ± 5.09	52.78 ± 0.66	0.55 ± 0.01	0.19 ± 0.03
RF	B	**89.10 ± 1.46**	91.55 ± 1.40	**0.90 ± 0.01**	**0.81 ± 0.02**
SVM	B	67.18 ± 4.03	53.41 ± 0.96	0.56 ± 0.02	0.15 ± 0.04

SE values showed an increment for all tested classifiers and feature sets. The highest sensitivity $SE = 91.60\%$ value was reached for the combination of the features set **A** and the RF method. This approach also reached the highest accuracy $ACC = 92\%$ and Matthews Correlation Coefficient $MCC = 0.84$. Although KNN classifier with the features set **B** presented the highest SP, the remaining scores were not the highest values. Figure 21 plots all the ACC scores obtained. An increase in the mean values is shown, except for the NB method and SVM when using as input the features set **B**.

The experiments conducted in this Chapter compared the classification performance when using two different sets of features according to the use of *feature averaging* or *cycle averaging* techniques. Seven different classification methods were also eva-

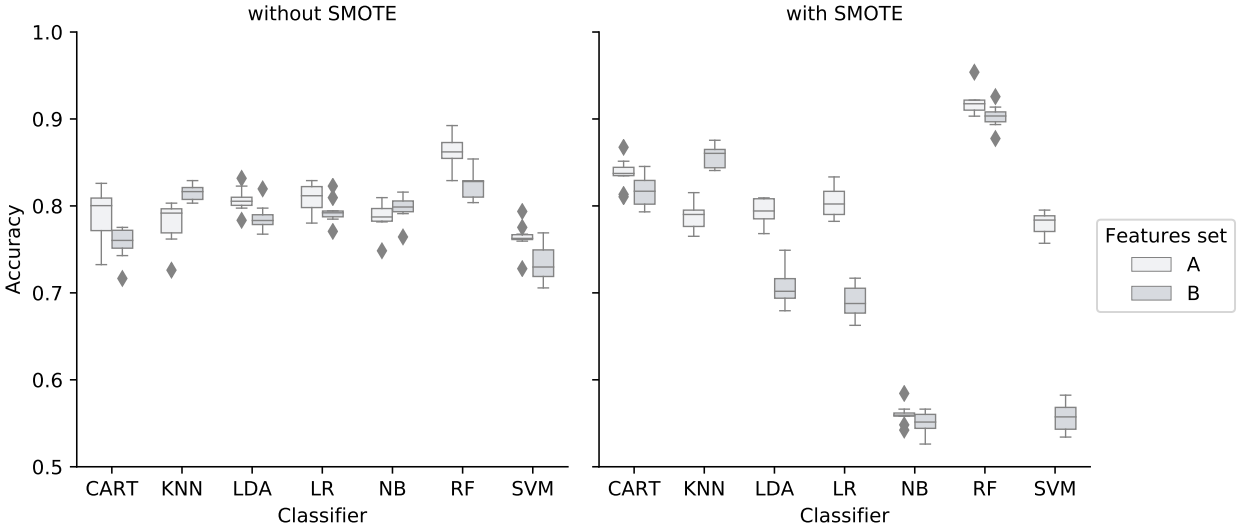

Figure 21. Accuracy scores for the conducted experiments.

luated by using the SE, SP, MCC and ACC scoring metrics. The use of the SMOTE balancing method when oversampling the minority class was also considered to be part of the classification benchmark since there was an improvement in the results obtained.

Based on the results reported in this Chapter, the feature averaging technique outperformed in terms of the resulting SE and MCC. For the classification schemes, the RF approach obtained the highest SE, ACC and MCC values. Considering the use of this classifier and the features extraction method, and the balancing of classes we will add the features selection step to the classification benchmark.

Chapter 5. Feature selection methods for the heart sounds classification benchmark

This Chapter adds the feature selection stage to the design of the heart sound signal classification benchmark. The screening tool is designed in terms of the model previously defined in Chapter 2 and the features extraction, classification scheme and balancing techniques which obtained the best evaluation results for the experiments conducted in Chapter 4. A couple of methods for features selection are compared to be added for the classification scheme. An evaluation of the complete benchmark is performed when using different sets of features. The feature extraction methods for testing the benchmark are constructed when combining the calculation of the MFCC coefficients with the parameters of the proposed reconstruction model.

5.1 Methods for features selection

A feature selection method in Machine Learning is the process of choosing a subset of relevant attributes which are the most appropriate to use for making decisions. This process makes a reduction in the number of attributes and basically gives the next advantages:

- Simplifies the model so it is easier to interpret for researchers and other users
- Training times become shorter
- Avoids the curse of dimensionality
- Reduces the overfitting

The main objective of features selection is to remove features which are either redundant or relevant without incurring much in loss of information. Two principal techniques for features selection are the correlation feature selection and information gain.

5.1.1. Correlation Features Selection (CFS)

The Correlation Features Selection method arises from the hypothesis that “good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other” (Hall, 1999). Thus, the highly correlated features between them

are removed selecting a subset in the following way:

$$CFS = \max_{s_k} \left[\frac{r_{cf_1} + r_{cf2} + \cdots + r_{cf_k}}{\sqrt{k + 2(r_{f_1 f_2} + \cdots + r_{f_i f_i} + \cdots + r_{f_k f_1})}} \right],$$

where s is the selected subset of k features which have r_{cf} correlation with the classification and r_{ff} correlation between other features. These values are not necessarily calculated using the Pearson or Spearman correlation coefficients, the minimum descriptor length and symmetrical uncertainity methods have been also used.

5.1.2. Information Gain (IG)

The Information Gain based method relies on the calculation of mutual information between features, defined by two continuous random variables X and Y as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right),$$

where $p(x,y)$ is the joint probability and $p(x)$, $p(y)$ the marginal probability distribution of the variables X and Y respectively. The following steps are commonly made in the IG method:

- Calculation of the mutual information as the score between all the features $f_i \in F$ with a target class c.
- Select the feature with the largest score $\text{argmax}_{f_i \in F}(I(f_i, c))$

The features of the set are ranked in descending order according to the mutual information score obtained.

5.2 Construction of the benchmark

With a classification method defined, the benchmark design will be evaluated with the combination of CFS and IG features selection methods. For the IG implementation, features with a mutual information equal to zero were removed.

The designed setup is depicted in Figure 22, where the train-test split was proposed to be a 80%-20% percentages of the data to be used for this subsets. The RF classification method was chosen with the parameters settings described in Table 6.

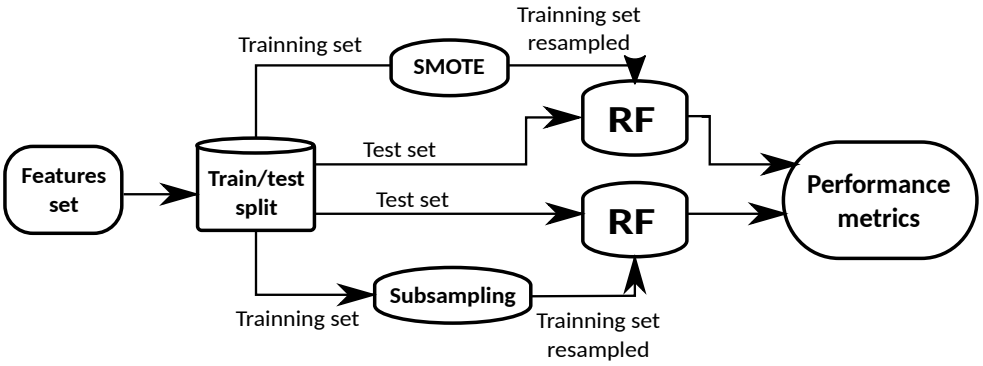

Figure 22. Classification setup diagram.

Scoring metrics for the evaluation of the selected classifier will be also the SE, SP, ACC and MCC. The scheme will be implemented in the scikit-learn (Pedregosa *et al.*, 2011) toolbox of Python and the Weka software (Hall *et al.*, 2009). For balancing the dataset the SMOTE oversampling strategy will be considered. An under sampling approach will be also tested by selecting a random subset which equalize the class samples. Both strategies were implemented using the the imbalanced learn library of Python (Lemaître *et al.*, 2017).

5.2.1. Construction of features sets

Table 9 shows the sets of features to be compared during the developed experiments in this Chapter when including the features selection methods. For the features extraction method we take both strategies considered in Chapter 4, feature averaging (set **A**) and cycle averaging (set **B**).

Three new feature sets: **C**, **D** and **E** are included for the benchmark assessment. These sets of attributes were constructed when combining the MP and LPC parameters with the MCC coefficients as feature extraction methods. In set **C**, we include the MP-LPC for feature extraction including MFCC coefficients of the heart cycles. On the other hand, features set **D** includes only the calculation of MFCC parameters whi-

Table 9. Combination of feature extraction and feature selection methods to be evaluated.

Label	Feature extraction methods	Feature selection methods	Number of features
A	MP and LPC, feature averaging	IG and CFS	90
B	MP and LPC, cycle averaging		19
C	MP, LPC and MFCC, feature averaging	CFS	146
D	MFCC feature averaging		56
E	MP and MFCC feature averaging		131

le set **E** was constructed as the fusion of the MP and MFCC values. Table 9 details this information. Diverse researchers have used MCC coefficients to efficiently extract the heart sound cycles features reporting promising results (Wang *et al.*, 2007; Potes *et al.*, 2016; Abdollahpur *et al.*, 2017). The MFCC C_n coefficients are calculated as in the following equation:

$$C_n = \sum_{m=1}^{M} D_m \cos\left(\frac{m-0.5}{M}\right)\pi n, \tag{20}$$

where D_m is the output of the k-th filter bank channel, M is the number of filter bank channels and $m = 1,2,\cdots M$. The filter bank used is triangular and the bandwidth of each channel is defined in a Mel-scale, which is constructed using the following conversion:

$$\text{mel}(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right). \tag{21}$$

The parameters used in this work were a number of M =14 filters, low and high frequencies of f_{low} =20 Hz and f_{high} =900 Hz respectively. Each heart sound event (considering S1, systole, S2 and diastole as events) was divided into frames of 20 ms with and overlap of 5 ms. To compute the MFCCs, we used the same settings/parameters as proposed in (Abdollahpur *et al.*, 2017) in order to compare the performance of these coefficients with respect to the model attributes obtained in this work for the PCG signals classification task. Figure 23 illustrates the resulting triangular filter bank for the MFCC calculation.

5.3 Numerical results

In this section we implemented the classification scheme described in Figure 22 using the features extraction approaches described in Table 9. According to including or not the feature selection stage, the next section describes the different scenarios.

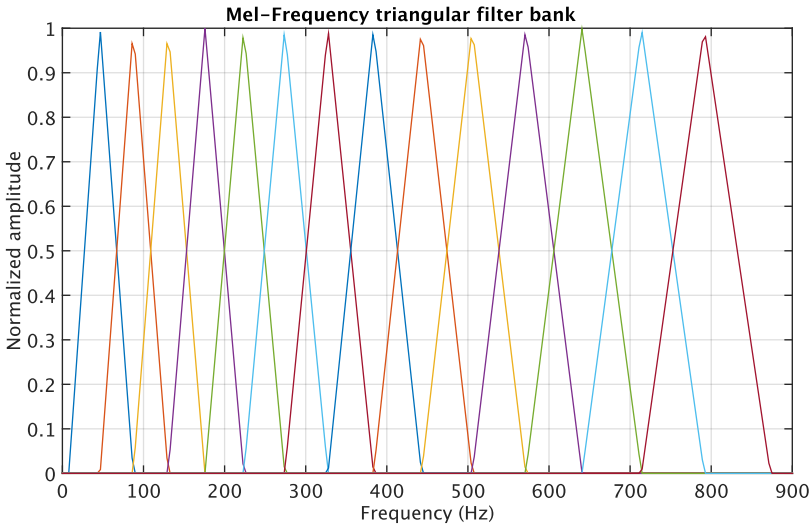

Figure 23. Triangular Mel-scale filter bank to calculate the MFCC coefficients.

5.3.1. Implementation with no feature selection stage

In a first experiment, we implemented the classification designed scheme in scikit-learn to compare the performance metrics of feature sets **A**, **B**, **C** and **D**. In this case the feature selection approach is not considered. Table 10 details the outcome results.

Table 10. Performance evaluation of the classification benchmark without using a feature selection stage.

Features set	Resampling	SE	SP	ACC	MCC
A	over	71.07	94.07	88.28	0.68
B	over	27.05	95.56	78.29	0.33
C	over	81.14	**95.98**	**92.24**	**0.80**
D	over	77.99	92.59	88.91	0.71
E	over	79.25	95.56	91.45	0.77
A	under	88.06	79.03	81.30	0.61
B	under	74.22	70.13	71.16	0.40
C	under	93.72	83.27	85.90	0.70
D	under	91.20	82.00	84.32	0.67
E	under	**94.34**	84.12	86.69	0.72

The graphical representation of the results given in Table 10 for the undersampling case is shown in Figure 24.

On the other hand, Figure 25 illustrates the bar chart from the results in Table 10 when the SMOTE oversampling method was applied. Since only one value was represented for each metric the error bars cannot be included in all the charts presented in this section.

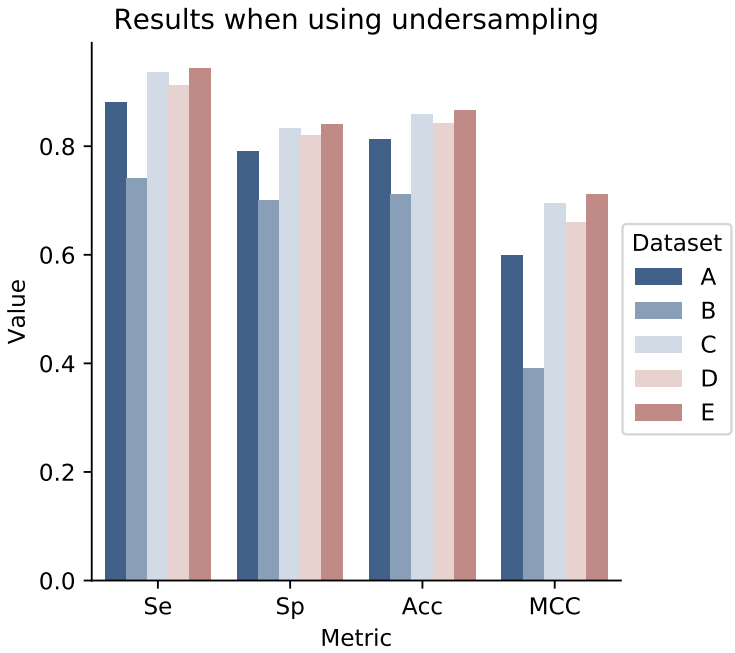

Figure 24. Bar chart of the output metrics for the classification test when no feature selection stage was implemented. Undersampling was applied.

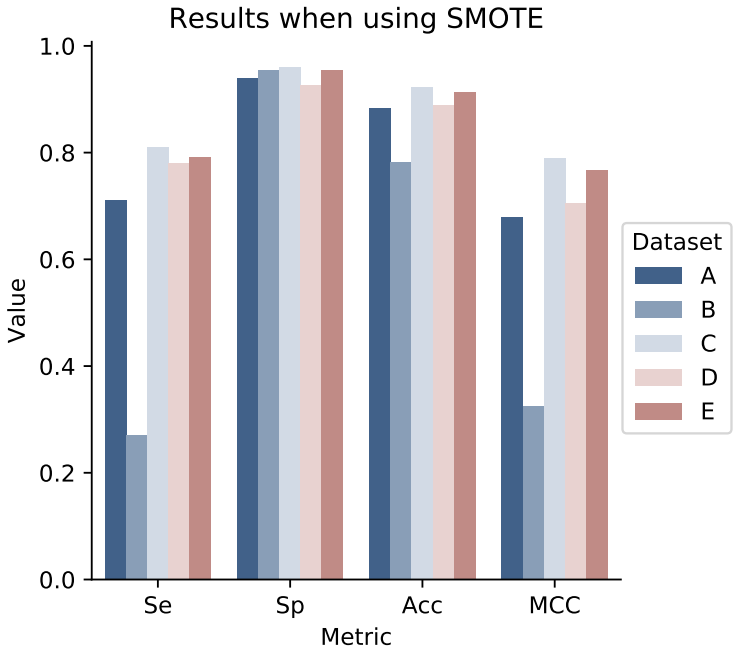

Figure 25. Bar chart of the output metrics for the classification test when no feature selection stage was implemented. The oversampling technique was applied.

From the results shown in Table 10, Figure 24 and Figure 25 we observe that the highest Sensitivity SE =94.34 was obtained when using random under sampling in the features set **E**. The highest Specificity SP =95.88 and proportionally the highest Accuracy ACC =92.24 were obtained by features set **C** when using SMOTE. The use of

this oversampling method also improves the MCC score, in fact, features set **C** also obtained the highest value of MCC =0.80.

5.3.2. Implementation including the feature selection stage

A second experiment was conducted when including the IG and CFS features selection methods to the classification benchmark. The feature selection methods were implemented in Weka. On the other hand the classification scheme was applied in scikit-learn. Results are shown in Table 11.

Table 11. Performance evaluation of the classification benchmark using a features selection stage.

Features Set	Features selection	Resampling	SE	SP	ACC	MCC
A	CFS	over	71.07	89.83	85.10	0.61
B	CFS	over	12.58	**97.88**	76.39	0.21
A	IG	over	67.92	93.64	**87.16**	**0.65**
B	IG	over	32.70	93.01	77.81	0.33
A	CFS	under	**88.05**	78.60	80.98	0.60
B	CFS	under	70.44	66.74	67.67	0.33
A	IG	under	86.79	79.02	80.98	0.59
B	IG	under	75.47	71.82	72.74	0.41

Figure 26 illustrates the output metrics of Table 11 when applying the undersampling technique as balancing method.

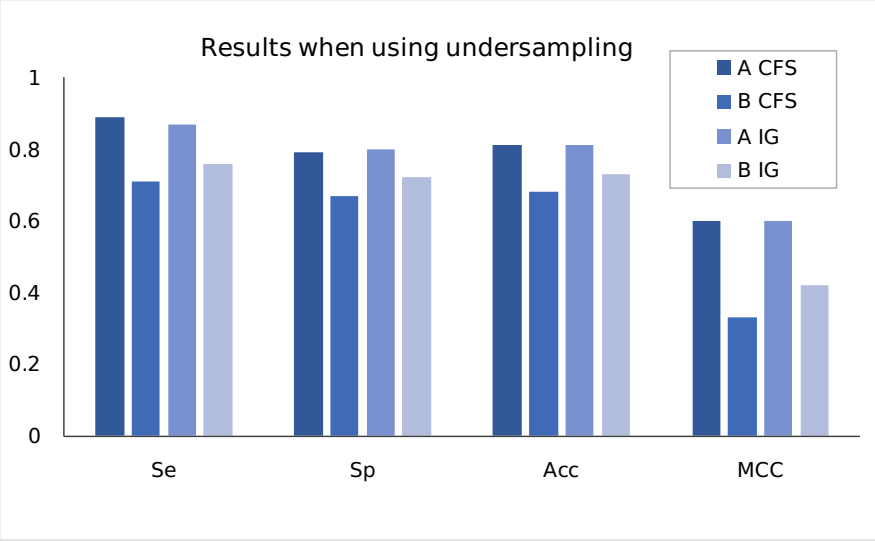

Figure 26. Bar chart showing the output metrics of the classification test applying CFS and IG. The undersampling was applied for balancing.

For the oversampling case, Figure 27 illustrates the output metrics of Table 11 when applying SMOTE to equalize the classes.

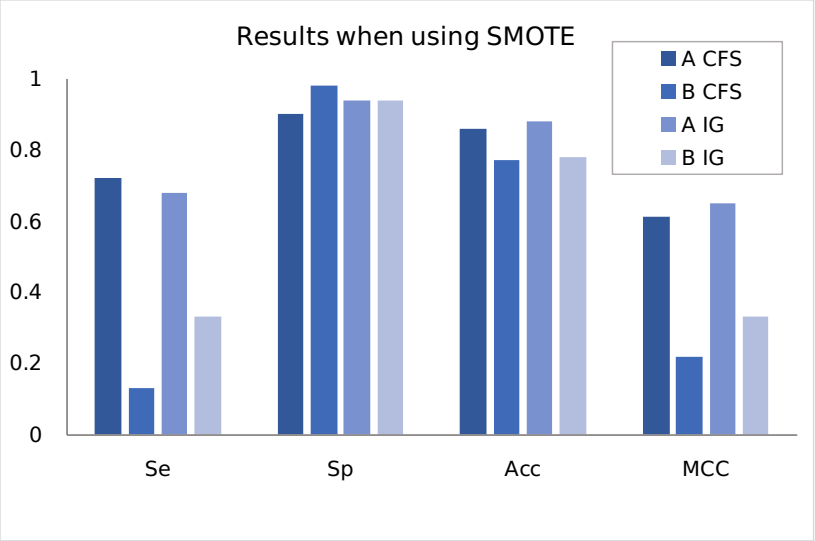

Figure 27. Bar chart showing the output metrics of the classification test applying CFS and IG. SMOTE was applied for balancing.

The outcome results of this experiment showed that the features set **A** obtained the highest $SE = 88.05$ value when using IG and undersampling as balancing method. The rest of the scores were higher when using the SMOTE oversampling. In terms of specificity $SE = 97.88$ was the highest score obtained by features set **B** when using CFS. The higest $ACC = 87.16$ and $MCC = 0.65$ scores correspond to the features set **A** when using the IG technique. In addition, features set **A** was reduced from 90 to 22 features when using CFS and from 90 to 66 when using IG. On the other hand, features set **B** presented the following reductions: in case of CFS it was reduced from 19 to 18 features while for IG none of the features was eliminated, preserving the whole set of features (19).

5.3.3. Implementation including the feature selection stage for the MFCC-based feature sets

A last experiment was conducted to assess the classification of PCG signals comparing the feature sets **C**, **D** and **E**, which contain the MFCC's calculation as feature extraction method. Hence when applying the CFS method higher scores where obtained, it was implemented to select the most relevant features of the sets. CFS method also produced a higher reduction in the number of features $N_{features}$. Table 12 presents the original and reduced number of features for the MFCC-based feature sets when implementing the CFS feature selection method in Weka.

Table 12. Original and reduced number of attributes for the MFCC-based features sets.

Features set	Original $N_{features}$	Reduced $N_{features}$ applying CFS
C	146	30
D	56	12
E	132	23

The CFS reduction effect in the number of features that displays Table 12 preserves among 15% to 30% of the original number of attributes. Set **E** obtained the highest reduction, conserving just 17% of the features as relevant.

The compressed sets of attributes were tested for classification of PCGs in Weka and scikit-learn. Table 13 displays the obtained results for the implementation in Weka. Figure 28 provides a radar plot of the outcomes depicted in Table 13. The highest

Table 13. Results from Weka for the classification PCGs using the MFCC-based feature sets including feature selection.

Dataset	Balancing	Feature selection	SE	SP	ACC	MCC
C	under	CFS	90.20	80.50	82.60	0.61
D	under	CFS	89.70	79.00	81.30	0.59
E	under	CFS	**90.60**	80.90	83.00	0.62
C	over	CFS	82.90	**90.50**	**88.80**	**0.70**
D	over	CFS	82.50	85.80	85.10	0.62
E	over	CFS	85.00	89.50	88.50	**0.70**

SE=90.60 was obtained by set **E** using undersampling. It can be seen that the resulting SP is directly related with the output ACC and the highest scores of this metrics were obtained when using SMOTE and the feature extraction provided by **C**. The MCC obtained better results when using SMOTE, and the highest score was obtained for both **C** and **E** sets. In this implementation, features set **D** did not obtain any higher score, showing that better performance is presented when adding any other features to the MFCC approach. The classification benchmark was also conducted in scikit-learn using as input attributes the features given by sets **C**, **D** and **E**. Results are shown in Table 14.

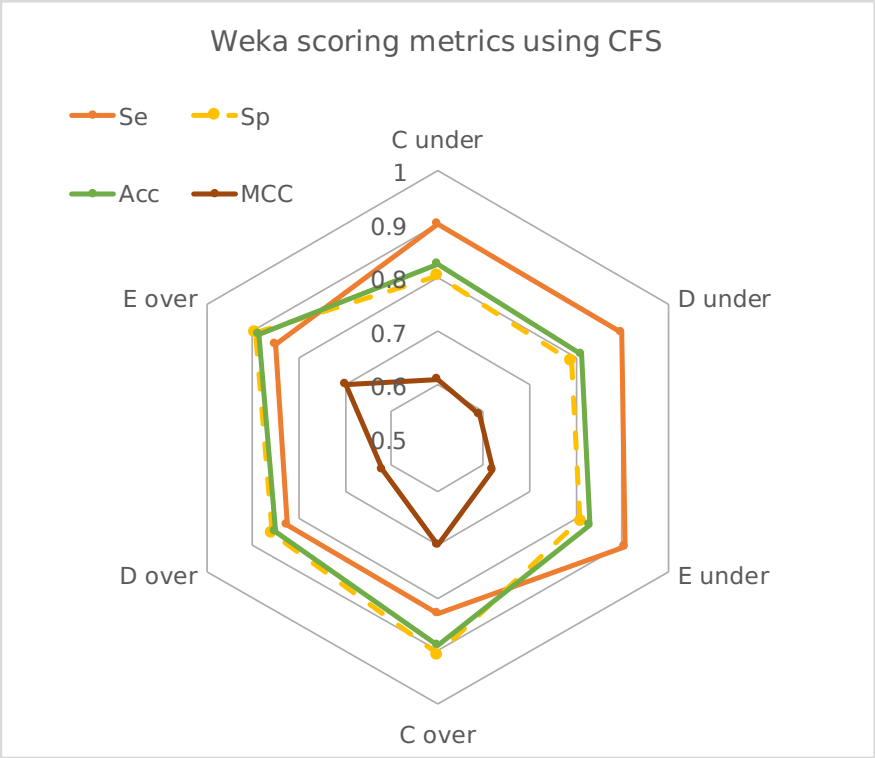

Figure 28. Radar plot of the results from the scikit-learn implementation using the MFCC-based feature sets and CFS.

Table 14. Results for the classification of PCGs using the MFCC-based feature sets including feature selection under scikit-learn.

Dataset	Balancing	Se	Sp	Acc	MCC
C	under	**91.83**	82.63	84.95	0.68
D	under	88.68	82.84	84.32	0.66
E	under	89.31	83.27	84.79	0.67
C	over	80.51	**92.59**	**89.55**	**0.73**
D	over	76.11	89.62	86.22	0.65
E	over	82.39	91.11	88.91	0.72

Figure 29 provides a visual representation in radar plot for the results depicted in Table 14. In this implementation, set **E** obtained the highest SE=91.83, using under-sampling for balancing. The rest of the metrics obtained the best results when applying feature sets **C**. Outcomes obtained from both implementations are related because SP and ACC are highly associated too, they are improved when using SMOTE and MCC is also higher when using this oversampling approach. The features set **D** did not obtain

any high score, showing also that the inclusion of MP and LPC features improves the classification results.

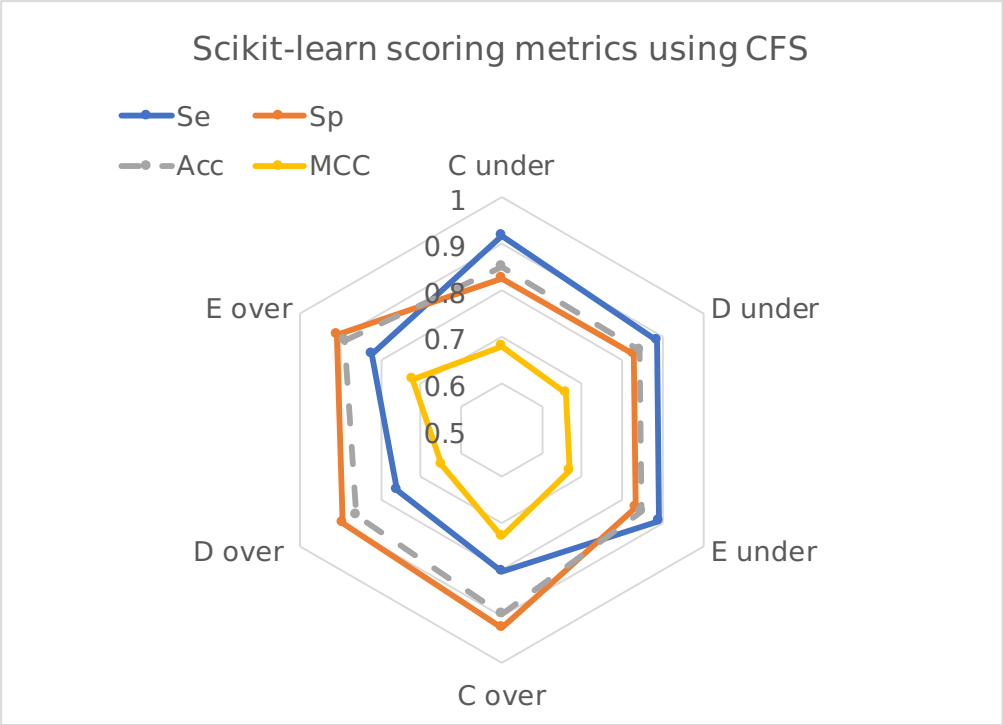

Figure 29. Radar plot of the results from the Weka implementation using the MFCC-based feature sets and CFS.

This Chapter introduced the concept of features selection to avoid the use of relevant features. We analyzed the effects of adding a features extraction step to the design of a classification of heart sounds benchmark. For this step, the Information Gain and Correlation Feature Selection methods were evaluated. The CFS method outperforms when it was applied to feature sets **A** (feature averaging) and **B** (cycle averaging) in terms of SE, SP, ACC and MCC metrics.

The conducted experiments also considered the inclusion of a random under sampling approach for balancing. We contrast the classification performance results when using or not this method compared to oversampling the minority class with SMOTE.

The Chapter includes also an evaluation of the performance of the PCG signals classification including the calculation of the MFCC coefficients as a method for feature extraction. Tree new sets of features **C**, **D** and **E** were created when combining the obtained LPC and MP based attributes with the MFCC parameters. The complete

benchmark was evaluated under Weka and scikit-learn platforms.

Outcomes of the test showed a better performance of feature sets **C** and **E** in terms of SE when using the under sampling method. However, the SP, ACC and MCC obtained the highest values with the application of SMOTE. It was presented a high relationship between the SP and ACC values for the different scenarios.

The random undersampling method has been an efficient method for class imbalanced learning. It has been used among the participants of the Physionet/CinC challenge. However, since many majority class samples are ignored, the training set may potentially neglect useful information (for instance, the quality of the signal) (Liu *et al.*, 2009).

Chapter 6. Conclusions

We presented in this work a methodology for the automated analysis of cardiac sound signals focusing on the modeling (feature extraction) and classification stages. The main goal was to evaluate different methods for feature extraction, classification and selection of the most important signal attributes to design a screening tool for the automated detection of heart diseases from cardiac sound signals. This Chapter highlights the main contributions of the research work produced by this thesis.

6.1 Heart sounds reconstruction model

To reconstruct the PCG, a model was designed under the hypothesis that it is a signal composed by the addition of a deterministic plus a stochastic part given by the expression $x(t) = x_h(t) + x_n(t)$. The deterministic part $x_h(t)$ was modeled by using a Matching Pursuit decomposition with Gabor atoms. These are waveforms that represent the time-frequency components of the signal. Since when using the MP algorithm it is not possible to represent the 100% of the energy contained in the signal, we attain an energy amount close to 90%. The numerical tests conducted showed that it is possible to attain this percentage when using an average number of $M = 15$ atoms per cardiac cycle. For the stochastic part of the model $x_n(t)$, we consider to represent the residual signal, which is actually the subtraction between the MP representation and the original signal. This residual signal has low correlation with the set of time-frequency atoms selected by MP to represent $x_h(t)$. As a result, we opted for the LPC coding technique to represent the residual as an autoregressive process.

Effectiveness of the proposed reconstruction model was evaluated under objective and subjective tests. For the objective test, we evaluate the percentage of root-mean distortion PRD to examine the quality of the representation. A set of 38 different PCG signals containing different murmurs and pathological states were evaluated (eGeneral Medical, n.d.). Numerical results displayed an average PRD =5% when using a 99% of the energy to represent the deterministic part and the remaining 1% for the residual. For the subjective test, we asked health experts to participate in a MUSHRA test (ITU-R, 2001; Mason, 2002). Participants were required to evaluate the quality of 10 reconstructed PCG signals using our proposed model. It is important to highlight that the physicians who took the test considered that the PCG reconstructed signals can

be used to diagnose an unhealthy cardiac state. We are not aware of any other model for heart sounds representation rigorously evaluated such as the one considered in this thesis. As a conclusion, numerical results from the subjective and objective tests reveal that our representation model provides an accurate reconstruction of diverse pathological heart sounds.

6.2 Heart sounds classification benchmark

Motivated by the results of the model evaluation, we designed a benchmark for the classification of heart sounds. The Physionet/CinC 2016 database was used to evaluate the performance of the algorithms. Two sets of time-frequency features were extracted from each heart sound, based on the designed reconstruction model (Ibarra *et al.*, 2015; Ibarra-Hernández *et al.*, 2017). Feature sets considered are based on the feature averaging (called features set **A**) and cardiac cycle averaging (called features set **B**) approaches. A number of different methods for binary classifications were evaluated when conducting a 10-fold cross validation test. Among all conducted experiments, the Random Forest (RF) classifier combined with SMOTE technique outperformed all other configurations, reaching a competitive score for accuracy (92%), with a good sensitivity-specificity trade-off. Detailed insights from the experiments also validate the newest of the two used feature sets in the considered classification task.

It must be noted that metrics used in the Physionet/Challenge are slightly different from ours, so direct comparisons and interpretations should be considered with attention. Specifically, we did not weight SP and SE according to the data labels as indicated in the challenge. However, we consider that only little discrepancies can be expected from this difference in the evaluation criteria. On the other hand, the average $MAcc$ is different from the ACC criterion, which puts more importance on the SE/SP trade-off, and they cannot be directly compared. The highest ranked system (Potes *et al.*, 2016), based on a convolutional neural network, reached a mean accuracy of $MAcc = 86.02\%$ with $Se' = 94.24\%$ and $Sp' = 77.81\%$. Although we obtained a SE score which is not as high as the best ranked method in the Physionet/Challenge, our research evaluates the feature extraction, classification and balancing methods performance before selecting any of these settings for the screening task.

The calculated features sets, reported scoring results tables and graphic charts are

available at `https://github.com/roilhi/PaperSIPAIM`.

6.3 Classification setup

The final stage of this research is the evaluation of the PCG signals classification scheme when adding the use of a features selection method. For this purpose, we evaluated the performance of the Correlation Feature Selection (CFS) and Information Gain methods. The CFS method obtained better performance than IG method since the highest SE, SP, ACC scores were obtained when using this feature selection method. Numerical test showed also that the number of relevant features selected was less when using CF than IG, giving as a result that only 17% of the original attributes can be considered as relevant to be used effectively for the classification task.

Calculation of the MFCC coefficients has been a frequently selected method to derive PCG signal features because they exhibit acceptable results to efficiently represent the relevant aspects of the signal (Wang *et al.*, 2007). In this research we compared the use of MFCC coefficients as feature selection method by creating datasets when combining this set of coefficients with the MP and LPC parameters of the proposed reconstruction model. Numerical tests displayed that the classification results improved when feature extraction methods are associated. As a matter of fact, feature set **C** created as junction of all feature extraction methods, outperforms in terms of SE=94.34, SP=95.98%, ACC=92.24% and MCC=0.80% when not using a feature selection stage and scores of SE=91.83, SP=92.59%, ACC=89.55% and MCC=0.73% scores when using CFS (keeping just a 17% of the attributes). The features set **D**, provided by only the MFCC coefficients as features did not get any high score.

Considering a special processing for noisy recordings represents a challenge. Hence, PCG signals recorded in noisy environments significantly affected the classification results. The segmentation step is also affected if the signal has been recorded with low quality. However, the proposed research work in this thesis worth on the use of simplest classifiers and the evaluation of the classification performance alternative feature extraction and features selection methods. The results obtained can lead to a future exploration of different time-frequency approaches for noisy recordings.

6.4 Contributions

From the present research arises a rigorous evaluation of methods for the automated evaluation of heart sounds, focusing on the feature selection and classification methods. The contributions of this work according to the different stages formulated are stated in the following points:

- We compared the performance of different well-known state-of-the-art classification methods to detect the presence or absence of pathologies in heart sounds. They were then tested in a 10 fold stratified cross validation technique.
- An oversampling technique, SMOTE, was also added in order to compensate for unbalanced classes (less recordings with an abnormal condition, which is typical in biological data), and its impact on classifiers performance was assessed. Performance was measured through sensitivity, specificity, accuracy and Matthews correlation coefficient. The MCC measure has not been considered for the classification of the Physionet database recordings assessment.
- Two alternative strategies for feature extraction were tested, feature averaging and cycle averaging. The problem of having an equal number of features is solved for all instances since we have a variable duration in the recordings.
- A couple of commonly used feature selection methods were tested, in order to know the effects of reducing the dimensionality of the dataset and reducing the time consumption for the training stage.
- The performance of the classification algorithm was tested when using different sets of features based on the MP and LPC parameters of the model and the MFCC coefficients, which have been widely used for PCG signals classification.

6.4.1. Future work

As the identified scheme shows promising results for detecting abnormal conditions in PCG signals, future work will target performance and computational efficiency improvements through a refined preprocessing (denoising) and Principal Component Analysis (PCA) methods for dimensionality reduction. It is important to remark the challenging task of classifying recordings under noisy environments (Ismail *et al.*, 2018).

The thesis focused on the simplest classification methods that do not require large amounts of training data. This approach motivates for the design of computationally effective solutions in real time of PCG signals murmurs detection based on the presented scheme. The execution times and computational resources used for the thesis were reasonably low. The implemented algorithms do not consume more than 10 seconds when executed over a Mac-book pro 2014 Intel Core i5 2.6 GHz processor over the Jupyter notebook of Python.

Although excluded from our benchmark due to their high training data requirements, neural network based classifiers cannot be ignored nowadays and should also be tested against the proposed feature set. However, the correct performance of these classification methods relies on the use of features sets of high dimensionality, using large sets of labeled data to learn features directly from data instead of extracting them manually.

The use of deep learning techniques to extract and train features can be also considered for the heart sounds classification task, however, it requires a large number of labeled samples for the training stage (millions of sounds or images, generally) (LeCun *et al.*, 2015). In addition, there is still a lack of information concerning to the heart sound pathological states since data available online does not cover the complete variety of pathologies that can be presented in clinical scenarios (Ismail *et al.*, 2018). Nevertheless, the Physionet sounds database is nowadays the only set of heart sound recordings that allows the comparison of screening tools for researchers.

One solution for the lack of information would be a *multi-modal* classification task. This scheme consists of adding different cardiac signals for diseases detection, such as the ECG, MRI and echocardiography. The classification might be stregthened for each patient, and could also be not binary for the detection of a specific pathology.

6.5 Productivity

The findings of this study produced the following publications. One journal publication and four conference papers were published.

Journals

- Ibarra-Hernández, R. F., Alonso-Arévalo, M. A., Cruz-Gutiérrez, A., Licona-Chávez, A. L., and Villarreal-Reyes, S. (2017). Design and evaluation of a parametric model for cardiac sounds. *Computers in biology and medicine, 89*, 170-180.

Conference proceedings

- Ibarra-Hernández, R. F., Alonso, M. A., Villarreal, S., and Nieblas, C. I. (2015, November). A parametric model for heart sounds. In *Signals, Systems and Computers, 2015 49th Asilomar Conference on* (pp. 765-769). IEEE.
- Ibarra-Hernández, R. F., Bertin, N., Alonso-Arévalo, M. A., and Guillén-Ramírez, H. A. (2018, December). A benchmark of heart sound classification systems based on sparse decompositions. In *14th International Symposium on Medical Information Processing and Analysis* (Vol. 10975, p. 1097505). International Society for Optics and Photonics (SPIE).
- Ibarra-Hernández, R. F., Alonso-Arévalo, M. A., Cruz-Gutiérrez, Villarreal-Reyes S. and Conte R. (2015, October). Desarrollo de un Codificador para audio cardiaco. In *Congreso Nacional de Ingeniería Biomédica (CNIB)* (pp. 157-160). Sociedad Mexicana de Ingeniería Biomédica (SOMIB).
- Ibarra-Hernández, R. F., Alonso-Arévalo, M. A., Cruz-Gutiérrez, Villarreal-Reyes S. and Conte R. (2015, September). Reconstrucción de señales de audio cardiaco mediante Matching Pursuit. In *Mexican Conference on Computer Science*. Sociedad Mexicana de Ciencias de la Computación (SMCC).

Cited bibliography

Abbas, A. K. and Bassam, R. (2009). Phonocardiography signal processing. *Synthesis Lectures on Biomedical Engineering*, **4**(1): 1–194.

Abdollahpur, M., Ghiasi, S., Mollakazemi, M. J., and Ghaffari, A. (2016). Cycle selection and neuro-voting system for classifying heart sound recordings. En: *Computing in Cardiology Conference (CinC), 2016*. IEEE, pp. 1–4.

Abdollahpur, M., Ghaffari, A., Ghiasi, S., and Mollakazemi, M. J. (2017). Detection of pathological heart sounds. *Physiological measurement*, **38**(8): 1616.

Agostihno, M. C. and Souza, M. N. (1997). A New Heart Sound Simulation Technique. *IEEE 19th International Conference - IEEE/EMBS*, **00**(C): 323–326.

Audi, R. (1999). The Cambridge dictionary of Phylosophy.

Ball, N. M., Brunner, R. J., Myers, A. D., and Tcheng, D. (2006). Robust machine learning applied to astronomical data sets. i. star-galaxy classification of the sloan digital sky survey dr3 using decision trees. *The Astrophysical Journal*, **650**(1): 497.

Bentley, P., Nordehn, G., Coimbral, M., Mannor, S., and Getz, R. (2011). The PASCAL classifying heart sounds challenge 2011 (CHS2011). `https://peterjbentley.com/heartchallenge/index.html`.

Breiman, L. (2001). Random forests. *Machine learning*, **45**(1): 5–32.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Chan, T. F., Golub, G. H., and LeVeque, R. J. (1982). Updating formulae and a pairwise algorithm for computing sample variances. En: *COMPSTAT 1982 5th Symposium held at Toulouse 1982*. Springer, pp. 30–41.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, **16**: 321–357.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM review*, **43**(1): 129–159.

Clifford, G. D., Liu, C., Moody, B., Springer, D., Silva, I., Li, Q., and Mark, R. G. (2016). Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016. En: *Computing in Cardiology Conference (CinC), 2016*. IEEE, pp. 609–612.

Clifford, G. D., Liu, C., Moody, B., Millet, J., Schmidt, S., Li, Q., Silva, I., and Mark, R. G. (2017). Recent advances in heart sound analysis. *Physiological measurement*, **38**(8): E10–E25.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, **20**(3): 273–297.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**(1): 21–27.

Cruz, J. A. and Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, **2**: 117693510600200030.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., *et al.* (2011). A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, **43**(5): 491.

Djebbari, A. and Bereksi-Reguig, F. (2011). A new chirp–based wavelet for heart sounds time–frequency analysis. *Int J Commun Antenna Propagation (IRECAP)*, **1**: 92–102.

Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., and Binder, M. (2001). A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *Journal of biomedical informatics*, **34**(1): 28–36.

eGeneral Medical (n.d.). eGeneral Medical Cardiac Auscultation of Heart Murmurs. `http://www.egeneralmedical.com/listohearmur.html`.

Elad, M. (2010). *Sparse and Redundant Representations. From theory to Applications in Signal and Image Processing*. Springer.

Eldar, Y. C. and Kutyniok, G. (2012). *Compressed sensing: theory and applications*. Cambridge University Press.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, **7**(2): 179–188.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, Vol. 1. Springer series in statistics New York, NY, USA:.

Fuchs, G., Helmrich, C. R., Marković, G., Neusinger, M., Ravelli, E., and Moriya, T. (2015). Low delay lpc and mdct-based audio coding in the evs codec. En: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, pp. 5723–5727.

Gribonval, R. (2001). Fast matching pursuit with a multiscale dictionary of Gaussian chirps. *IEEE Transactions on Signal Processing*, **49**(5): 994–1001.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, **11**(1): 10–18.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. Tesis de doctorado, Department of Computer Science University of Waikato Hamilton.

Huiying, L., Sakari, L., Iiro, H., and Processing, T. S. (1997). A Heart Sound Segmentation Algorithm Using Wavelet Decomposition and Reconstruction. En: *Proceedings - 19th international Conference - IEEE/EMBS Oct. 30 - Nov. 2, 1997 Chicago, IL. USA*, Chicago, IL. USA. Vol. 1630, pp. 1630–1633.

Ibarra, R. (2014). *Desarrollo de un códec para la transmisión de audio cardiaco sobre redes de bajas tasas de datos*. Tesis de maestría, Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE).

Ibarra, R. F., Alonso, M. A., Villarreal, S., and Nieblas, C. I. (2015). A parametric model for heart sounds. En: *Signals, Systems and Computers, 2015 49th Asilomar Conference on*. IEEE, pp. 765–769.

Ibarra-Hernández, R. F., Alonso-Arévalo, M. A., Cruz-Gutiérrez, A., Licona-Chávez, A. L., and Villarreal-Reyes, S. (2017). Design and evaluation of a parametric model for cardiac sounds. *Computers in biology and medicine*, **89**: 170–180.

Institute, O. U. H. (2019). Heart valve viseases. Ottawa University Heart Institute: `https://www.ottawaheart.ca/heart-condition/heart-valve-disease`. Accessed 2018-06-04.

Ismail, S., Siddiqi, I., and Akram, U. (2018). Localization and classification of heart beats in phonocardiography signals—a comprehensive review. *EURASIP Journal on Advances in Signal Processing*, **2018**: 1–27.

ITU-R (2001). Bs. 1534-1. method for the subjective assessment of intermediate sound quality (mushra). *International Telecommunications Union, Geneva*.

Jabbari, S. and Ghassemian, H. (2011). Modeling of heart systolic murmurs based on multivariate matching pursuit for diagnosis of valvular disorders. *Computers in Biology and Medicine*, **41**(9): 802–811.

Jayant, N. S. and Noll, P. (1984). Digital coding of waveforms: principles and applications to speech and video. *Englewood Cliffs, NJ*, pp. 115–251.

Joo, T. H., McClellan, J. H., Foale, R. A., Myers, G. S., and Lees, R. S. (1984). Pole-Zero Modeling and Classification of Phonocardiograms. *IEEE Transactions on Biomedical Engineering*, **BME-31**(6): 473–474.

Kay, E. and Agarwal, A. (2016). Dropconnected neural network trained with diverse features for classifying heart sounds. En: *Computing in Cardiology Conference (CinC), 2016*. IEEE, pp. 617–620.

Kay, E. and Agarwal, A. (2017). Dropconnected neural networks trained on time-frequency and inter-beat features for classifying heart sounds. *Physiological measurement*, **38**(8): 1645.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, **13**: 8–17.

Köymen, H., Altay, B. K., and Ziya, I. Y. (1987). A Study of Prosthetic Heart Valve Sounds. *Biomedical Engineering, IEEE Transactions on*, **BME-43**(11): 853–863.

Krstulovic, S. and Gribonval, R. (2006). MPTK: Matching pursuit made tractable. En: *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, Vol. 3, pp. III–III.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, **521**(7553): 436.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, **18**(17): 1–5.

Leung, T., White, P., Cook, J., Collis, W., Brown, E., and Salmon, A. (1998). Analysis of the second heart sound for diagnosis of paediatric heart disease. *IEE Proceedings-Science, measurement and technology*, **145**(6): 285–290.

Liu, C., Springer, D., Li, Q., Moody, B., Juan, R. A., Chorro, F. J., Castells, F., Roig, J. M., Silva, I., Johnson, A. E., *et al.* (2016). An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, **37**(12): 2181.

Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **39**(2): 539–550.

Mahnke, C. B. (2009). Automated heartsound analysis/computer-aided auscultation: A cardiologist's perspective and suggestions for future development. En: *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, pp. 3115–3118.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, **63**(4): 561–580.

Mallat, S. (1999). *A wavelet tour of signal processing*. Elsevier.

Mallat, S. and Zhang, Z. (1993). Matching Pursuits With Time-Frequency Dictionaries. *IEEE Transactions on Signal Processing*, **41**(12): 3415–3997.

Manikandan, M. S. and Dandapat, S. (2007). Wavelet energy based compression of phonocardiogram (PCG) signal for telecardiology. *IET Seminar Digest*, **2007**(2): 650–654.

Martínez-Alajarín, J. and Ruiz-Merino, R. (2004). Wavelet and wavelet packet compression of phonocardiograms. *Electronics Letters*, **40**(17): 1040–1041.

Mason, A. (2002). The MUSHRA audio subjective test method. *BBC R&D White Paper WHP*, **38**.

Mellor, A., Boukir, S., Haywood, A., and Jones, S. (2015). Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin. *ISPRS Journal of Photogrammetry and Remote Sensing*, **105**: 155–168.

Müller, A. C., Guido, S., *et al.* (2016). *Introduction to machine learning with Python: a guide for data scientists*. .º'Reilly Media, Inc.".

Nieblas, C. I., Alonso, M. a., Conte, R., and Villarreal, S. (2013). High performance heart sound segmentation algorithm based on Matching Pursuit. *2013 IEEE Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)*, pp. 96–100.

Ozun, O., Steurer, P., and Thell, D. (2017). Wideband speech coding with linear predictive coding (lpc). UCLA Electrical Engineering Digital Speech Processing: `http://www.seas.ucla.edu/spapl/projects/ee214aW2002/1/report.html`. Accessed: 2018-05-17.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**: 2825–2830.

Polat, K., Güneş, S., and Arslan, A. (2008). A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. *Expert systems with applications*, **34**(1): 482–487.

Potes, C., Parvaneh, S., Rahman, A., and Conroy, B. (2016). Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. En: *Computing in Cardiology Conference (CinC), 2016*. IEEE, pp. 621–624.

Quatieri, T. F. (2006). *Discrete-time speech signal processing: principles and practice*. Pearson Education India.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital processing of speech signals*, Vol. 100. Prentice-hall Englewood Cliffs, NJ.

Ravelli, E., Richard, G., and Daudet, L. (2008). Union of mdct bases for audio coding. *IEEE transactions on audio, speech, and language processing*, **16**(8): 1361–1372.

Redlarski, G., Gradolewski, D., and Palkowski, A. (2014). A system for heart sounds classification. *PLoS ONE*, **9**(11).

Reed, T. R., Reed, N. E., and Fritzson, P. (2004). Heart sound analysis for symptom detection and computer-aided diagnosis. *Simulation Modelling Practice and Theory*, **12**(2): 129–146.

Sava, H., Pibarot, P., and Durand, L. G. (1998). Application of the matching pursuit method for structural decomposition and averaging of phonocardiographic signals. *Medical and Biological Engineering and Computing*, **36**(3): 302–308.

Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, **162**(1-2): 83–112.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., *et al.* (2002). Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, **8**(1): 68.

Springer, D. B., Tarassenko, L., and Clifford, G. D. (2016). Logistic regression-hsmm-based heart sound segmentation. *IEEE Transactions on Biomedical Engineering*, **63**(4): 822–832.

Tang, H., Li, T., Park, Y., and Qiu, T. (2010). Separation of heart sound signal from noise in joint cycle frequency-time-frequency domain based on fuzzy detection. *IEEE Transactions on Biomedical Engineering*, **57**(10): 2438–2447.

Tang, H., Zhang, J., Sun, J., Qiu, T., and Park, Y. (2016). Phonocardiogram signal compression using sound repetition and vector quantization. *Computers in Biology and Medicine*, **71**: 24–34.

UMHS (n.d.). Michigan Heart Sound and Murmur Library. `http://www.med.umich.edu/lrc/psb_open/html/repo/primer_heartsound.html`.

Vaidyanathan, P. (2007). The theory of linear prediction. *Synthesis lectures on signal processing*, **2**(1): 1–184.

Vercio, L. L., Del Fresno, M., and Larrabide, I. (2017). Detection of morphological structures for vessel wall segmentation in ivus using random forests. En: *12th International Symposium on Medical Information Processing and Analysis*. International Society for Optics and Photonics, Vol. 10160, p. 1016012.

Wang, P., Lim, C. S., Chauhan, S., Foo, J. Y. A., and Anantharaman, V. (2007). Phonocardiographic signal analysis method using a modified hidden markov model. *Annals of Biomedical Engineering*, **35**(3): 367–374.

Wang, W., Guo, Z., Yang, J., Zhang, Y., Durand, L.-G., and Loew, M. (2001). Analysis of the first heart sound using the matching pursuit method. *Medical and Biological Engineering and Computing*, **39**(6): 644–648.

Wang, W., Pan, J., and Lian, H. (2004). Decomposition and analysis of the second heart sound based on the matching pursuit method. En: *Signal Processing, 2004. Proceedings. ICSP'04. 2004 7th International Conference on*. IEEE, Vol. 3, pp. 2229–2232.

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, **10**(Feb): 207–244.

WHO (2018). The top 10 causes of death. World Health Organization: `https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death`. Accessed: 2018-05-02.

Xie, B. and Minn, H. (2012). Real-time sleep apnea detection by classifier combination. *IEEE Transactions on Information Technology in Biomedicine*, **16**(3): 469–477.

Xu, J., Durand, L. G., and Pibarot, P. (2001a). Extraction of the aortic and pulmonary components of the second heart sound using a nonlinear transient chirp signal model. *IEEE Transactions on Biomedical Engineering*, **48**(3): 277–283.

Xu, J., Durand, L. G., and Pibarot, P. (2001b). Extraction of the aortic and pulmonary components of the second heart sound using a nonlinear transient chirp signal model. *IEEE Transactions on Biomedical Engineering*, **48**(3): 277–283.

Zabihi, M., Rad, A. B., Kiranyaz, S., Gabbouj, M., and Katsaggelos, A. K. (2016). Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. En: *Computing in Cardiology Conference (CinC), 2016*. IEEE, pp. 613–616.

Zhang, X., Durand, L.-g., Senhadji, L., Lee, H. C., and Coatrieux, J.-l. (1996). Application of the matching pursuit method for the analysis and synthesis of the phonocardiogram. En: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Amsterdam 1996 4.2.8: Time-frequency Analysis of Various Signals II*. Vol. 1, pp. 1035–1036.

Zhang, X., Durand, L. G., Senhadji, L., Lee, H. C., and Coatrieux, J. L. (1998). Analysis-Synthesis of the Phonocardiogram Based on the Matching Pursuit Method. *IEEE transactions on bio-medical engineering*, **45**(8): 962–71.