Ph.D. in Electronic and Computer Engineering
Dept. of Electrical and Electronic Engineering
University of Cagliari, Italy

An Intelligent Diagnostic System for Screening Newborns

Amir Mohammad Amiri

Advisors: Prof. Giuliano Armano;
Curriculum: ING-INF/04 Informatica

XXVII Cycle
April 2015

Ph.D. in Electronic and Computer Engineering
Dept. of Electrical and Electronic Engineering
University of Cagliari, Italy

An Intelligent Diagnostic System for Screening Newborns

Amir Mohammad Amiri

Advisors:Prof. Giuliano Armano;
Curriculum: ING-INF/04 Informatica

XXVII Cycle
April 2015

I dedicate this dissertation to my lovely parents

Acknowledgments

Working as a Ph.D. student in Intelligent Agents & Soft-Computing (IASC) group at University of Cagliari was a magnificent as well as challenging experience to me. In all these years, many people were instrumental directly and indirectly in shaping up my academic career. It was hardly possible for me to thrive in my doctoral work without the precious support of these people. Here is a small tribute all.

Above all, I wish to express deepest gratitude to my supervisor Prof. Giuliano Armano for his full support, expert guidance, understanding and encouragement my study and research. Without his incredible patience and timely wisdom and counsel, it would not have been possible to write this doctoral thesis. Giuliano is someone you will instantly love and never forget once you meet him. He is very friendly and one of the smartest people I know. I hope that I could be as lively, enthusiastic, and energetic as Giuliano and to someday be able to command an audience as well as he can.

A very special word of thanks goes for my parents, Alireza and Hakimeh who have given me their unequivocal support throughout, as always, for which my mere expression of thanks likewise does not suffice.

I will forever be thankful of my colleagues with whom I have shared moments of deep anxiety but also of big excitement. Their presence was very important in a process that is often felt as tremendously solitaire. A warm word for my colleagues and great friends: Carla Orru, my first and one of the best Italian friend "no problem and don't worry Carla for your work", Alessandro Giuliani and Federico Palla, who were more than a friend or colleague for me and as a funny teacher for learning Italian "sorry I cannot write here what you taught me", Cristina Monni, with your active and funny attitude, we always had topic to talk or joking, who was always looking for floating things in the tea, "now you are adult, stop to drink milk", Mario Locci, I have many things to say but just, you are a real friend who always was supporting me, "wake up Mario", Martina Matta, great and kind friend who never gave up to finding a best friend as "Russelo" for me, Emanuele Tamponi, crazy and lovely friend "shave your beard", Matteo Orru, I always enjoyed to talk with you "I will watching you" and Francesca Funni, sweet and very sympathetic "bad choice but best one".

I am very grateful to all my friends Simona, Gabriele, Giampaolo, Marianna, Laura, Caterina, Ilaria, Erica, Simone, Anna, Floriana, Heike, Julio, Carolina, Matteo, Andrea.

I would like to give my deep appreciation to Mehran Zareh, Edwin Wall and Pamella Sulis thank you for being able to share happy and unhappy moments from each other.

I must also thank my close friends, Reza Farmani, Mansour Ahmadi, Farideh Tavazoei and Bahram Lavi.

"Ciao Ragazzi"

Abstract

The goal of this research is to devise and develop an intelligent system for analyzing heart sound signals, able to support physicians in the diagnosis of heart diseases in newborns. Many studies have been conducted in recent years to automatically differentiate normal heart sounds from heart sounds with pathological murmurs using audio signal processing in newborns. Serious cardiac pathology may exist without symptoms. Since heart murmurs are the first signs of heart disease, we screen newborns for normal (innocent) and pathological murmurs. This thesis presents a variety of techniques in time-frequency domain such as Cepstrum, Shannon energy, Bispectrum, and Wigner Bispectrum for feature extraction. A comparison of these techniques is considered to feature selection which has been used to reduce the size of the feature vector. In the final step, different classifiers and techniques, e.g., Multi layer perceptron (MLP), decision tree, Classification and Regression Trees (CART) and ensemble of decision trees, are applied on data in order to achieve highest performance. High classification accuracy, sensitivity, and specificity have been obtained on the given data by CART. The validation process has been performed on a balanced dataset of 116 heart sound signals taken from healthy and unhealthy medical cases.

Contents

List of Figures

List of Tables

1

Introduction

Cardiac auscultation is one of the most useful investigative tools that the physician can use as primary diagnosis tool at the bedside to detect alterations in cardiovascular anatomy and physiology. Despite remarkable advances in imaging technologies for heart diagnosis, clinical evaluation of cardiac defects by auscultation is still a main diagnostic method for discovering heart disease. In experienced hands, this method is effective, reliable, and cheap. Beside, tools for objective analysis and adequate documentation of the findings are not effective yet.

Heart sound contains information which cannot be perceived by the human ear. Recent advances in data recording technology and digital signal processing have made it possible to record and analyze the sound signals from the heart.

Heart diseases are a major cause of death. However, most risk factors can be dramatically reduced by diagnosing them in an early stage of life. In particular, some heart diseases can cause life-threatening symptoms and require intervention within the first days or weeks of life. Fortunately, even critical congenital heart diseases are often treatable in the event they are detected early [51]. Most patients with significant valvular heart disease are first diagnosed based upon the finding of a murmur.

The PCG signal discloses information about cardiac function through vibrations caused by the working heart. Perception of heart sounds is influenced by their production and transmission as well as the capability of the human auditive sensory system in recognizing correct amplitude and frequency of each sound. The human ear is not equally responsive to sound in all frequency ranges, and has relative perception about loudness and softness of a sound. Two sounds with the same intensity at different frequencies are perceived differently.

The auscultation of cardiac sound signals through either a conventional acoustic or an electronic stethoscope needs a long-term practice and experience. Namely, note that it can take years to acquire them. Although the stethoscope is the symbol of physicians, primary care physicians are documented to have poor auscultatory skill in actuality. The need for the primary care physicians to improve the cardiac auscultation skill is still very strong in the primary screening examination and becomes stronger for the general users to perform the auscultation at home (Jiang and Choi, 2006; Reed, Reed, Fritzson, 2004).

An innocent heart murmur still requires an echocardiogram for reassurance, even though the cost of an echocardiogram is not negligible. The result of this practice is a misallocation of healthcare funds. While it is clearly important to prevent healthy newborn being sent for echocardiogram, it is also important to avoid that a newborn that has a pathological heart murmur is sent home without proper treatment[23].

In particular the structure of the thesis should be clearly stated to the design and implementation of the proposed decision support system is described according to the actual steps performed in pipeline by the system: i) pre-processing, ii) feature extraction/selection and iii) classification. Figure 1-1 shows the main stages used by the proposed diagnostic system to classify heart murmurs.

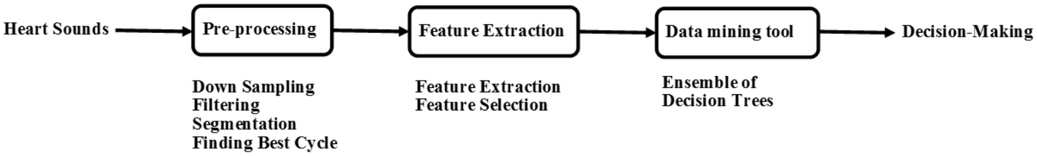


Figure 1-1: Main stages used by the proposed diagnostic system to classify heart murmurs.

2

Background

2.1 Heart

Heart is a muscular organ in humans and many animals that receives blood from the veins and pumps it through the arteries to other parts of the body that is located on the left of the chest, the heart is about 15 cm and a weight of around 325 grams. It consists of four main parts: the left atrium, left ventricle, right atrium and right ventricle. The heart works as a kind of natural pomp. Propelling the blood throughout the body, the heart beats are thus responsible for the bloodstream.

The heart has four chambers. The upper chambers are called the left and right atria, and the lower chambers are called the left and right ventricles. A wall of muscle called the septum separates the left and right atria and the left and right ventricles. The left ventricle is the largest and strongest chamber in your heart. The left ventricle's chamber walls are only about a half-inch thick, but they have enough force to push blood through the aortic valve and into your body. (See figure 2-1)

Right Ventricle

The lower right chamber of the heart. During the normal cardiac cycle, the right ventricle receives deoxygenated blood as the right atrium contracts. During this process the pulmonary valve is closed, allowing the right ventricle to fill. Once both

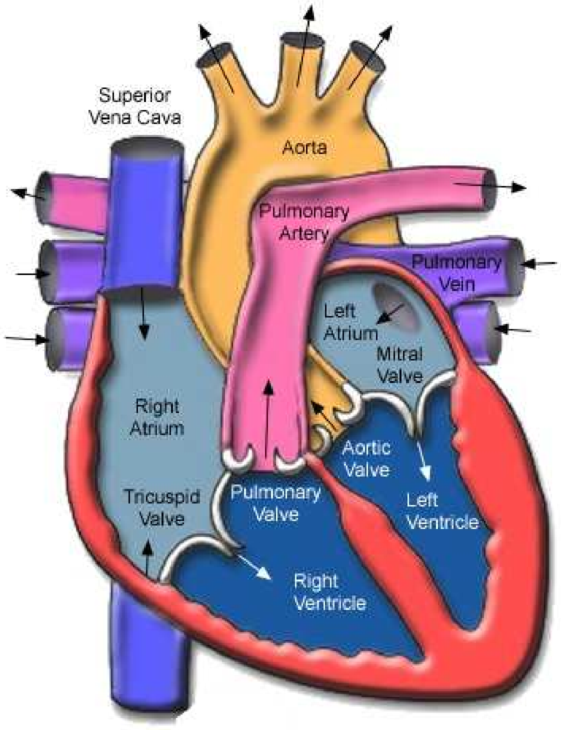


Figure 2-1: Anatomy of heart.

ventricles are full, they contract. As the right ventricle contracts, the tricuspid valve closes and the pulmonary valve opens. The closure of the tricuspid valve prevents blood from returning to the right atrium, and the opening of the pulmonary valve allows the blood to flow into the pulmonary artery toward the lungs for oxygenation of the blood The right and left ventricles contract simultaneously; however, because the right ventricle is thinner than the left, it produces a lower pressure than the left when contracting. This lower pressure is sufficient to pump the deoxygenated blood the short distance to the lungs.

Left Ventricle

The lower left chamber of the heart. During the normal cardiac cycle, the left ventricle receives oxygenated blood through the mitral valve from the left atrium as it contracts. At the same time, the aortic valve leading to the aorta is closed, allowing the ventricle to fill with blood. Once both ventricles are full, they contract. As the left ventricle contracts, the mitral valve closes and the aortic valve opens. The closure of the mitral valve prevents blood from returning to the left atrium, and the opening of the aortic valve allows the blood to flow into the aorta and from there throughout the body. The left and right ventricles contract simultaneously; however, because the left ventricle is thicker than the right, it produces a higher pressure than the right when contracting. This higher pressure is necessary to pump the oxygenated blood throughout the body.

Right Atrium

The upper right chamber of the heart. During the normal cardiac cycle, the right atrium receives deoxygenated blood from the body (blood from the head and upper body arrives through the superior vena cava, while blood from the legs and lower torso arrives through the inferior vena cava). Once both atria are full, they contract, and the deoxygenated blood from the right atrium flows into the right ventricle through the open tricuspid valve.

Left Atrium

The upper left chamber of the heart. During the normal cardiac cycle, the left atrium receives oxygenated blood from the lungs through the pulmonary veins. Once both atria are full, they contract, and the oxygenated blood from the left atrium flows into the left ventricle through the open mitral valve.

2.2 Heart murmurs in children

A baby's heart develops between the third and seventh weeks of pregnancy. Hearts start as a hollow tube that grows. As the tube becomes longer, it is forced to bend and rotate. The left and right atria form at the entry end of the tube, and the right and left ventricles form from the middle section. Walls divide the chambers and the valves form. The exit end of the original hollow tube divides into two channels, which become the pulmonary artery and the aorta.

A congenital heart defect is a heart problem that a baby is born with. It can include abnormal development of the heart, the heart valves, major arteries, or a combination of these problems. Congenital heart disease is much less prevalent than innocent murmur, occurring in only about 0.8% of live births [59], but the natural history of many common congenital cardiac defects can be one of progressive limitation and premature death. The primary care physician, therefore, very frequently faces the challenge of distinguishing between the relatively rare but important pathologic murmur and the ubiquitous innocent murmur.

Failure to diagnose heart disease is unacceptable because current treatments can dramatically improve outcomes [46]. Congenital heart defects are caused by a problem in the heart's development during the first few weeks of pregnancy. Usually the exact cause of the problem is not known, but often it is just a chance event in the complex development of the baby's heart. Sometimes infections and drugs cause a heart defect. For example, German measles (Rubella) and other viruses can damage the heart as it develops. If a woman takes certain medicines, smokes or drinks too much alcohol early in pregnancy, this can also cause heart and other problems.

Although a heart murmur is an important presenting feature of a cardiac disorder in infancy and childhood, innocent murmurs are very common, occurring in up to 80% of children at some time or other. These murmurs are frequently detected during a febrile illness and are also exacerbated by nervousness or on exercise. It is important to distinguish between innocent and pathological murmurs and to arrange more detailed evaluation of the child if there is any doubt. Children should be routinely screened for heart murmurs and other evidence of cardiac disorder between 6 and 8 weeks of age and at subsequent examinations during childhood. Serious cardiac pathology may exist without symptoms.

Innocent Murmurs

The commonest innocent murmur in children (usually heard at age 3-6 years, although also occasionally in infants) is the parasternal vibratory ejection systolic murmur (Still's murmur) which has a very characteristic low-frequency 'twanging' or musical quality. It is localized to the left mid-sternal border or midway between the apex and left lower sternal border, is of short duration, low intensity and is loudest when the child is supine often varying markedly with posture. It can be made to disappear on hyperextension of the back and neck (Scott's manoeuvre). The venous hum is a superficial continuous murmur heard beneath the clavicles and in the neck which can be abolished by head movements, by compression of the ipsilateral jugular vein or by lying the child supine. The innocent right ventricular outflow tract murmur (pulmonary flow murmur) is a soft early to midsystolic ejection murmur heard at the right upper sternal border but does not radiate to the back. In the premature and newborn infant an innocent pulmonary flow murmur may be audible radiating to the axillae and to both lungs at the back. Innocent carotid bruits common in normal children.

What is not innocent ? In addition to listening for murmurs careful attention should be paid to the presence of other evidence of cardiac pathology. Certain features indicate that a murmur is likely to be pathological and that prompt expert evaluation is needed:

1. Cyanosis or clubbing

2. Abnormal cardiac impulse

3. Abnormal breathing (tachypnoea, intercostal recession)

4. Thrill over precordium or suprasternal notch

5. Cardiac failure

6. Abnormal heart sounds

7. Failure to thrive

8. Presence of click

9. Abnormal pulses - diminished or absent femorals

10. Radiation of murmur to the back

11. Arrhythmia

12. Murmur which is purely diastolic

The most common cause of murmurs in newborns is when a specific condition called Patent Ductus Arteriosus (PDA) occurs, which is often detected shortly after birth, most commonly in premature newborns. This is a potentially serious condition in which blood circulates abnormally throughout the ductus arteriosus. In most cases, the only symptom of PDA is a heart murmur, which lasts until the ductus closes on its own, for healthy newborns typically shortly after birth. Sometimes, especially in premature newborns, it may not close on its own, or it may be large and permit too much blood to pass through the lungs, which can place extra strain on the heart, forcing it to work harder and causing a rise in blood pressure in the arteries of the lungs. If this is the case, a medication or, rarely, surgery may be needed to help close the PDA.

Pathological Murmurs

The pathological murmurs divided into two categorize: i) Systolic murmurs, maximal at the upper sternal borders are more likely to be ejection in type due to heart outflow abnormality or increased flow - aortic valve, subvalve or supravalve stenosis and HOCM being maximal on the right radiating to the neck whilst pulmonary valve,

subvalve or supravalve stenosis or atrial septal defect murmurs are louder on the left and radiate to the back. Those at the lower sternal border are more likely to be of regurgitant type due to ventricular septal defect, mitral or tricuspid regurgitation. Some pathological systolic murmurs are heard widely over the whole precordium and different types of murmur may coexist. Coarctation of the aorta is an important cause of a murmur over the back particularly in the interscapular region.
ii) Pathological diastolic murmurs, Diastolic murmurs should always be re g a rded as pathological. Early diastolic decrescendo murmurs are associated with incompetence of a semilunar valve - the a o rtic valve in bicuspid aortic valve or Marfan syndrome, the pulmonary valve following surgery for tetralogy of Fallot or pulmonary stenosis and more rarely in conjunction with pulmonary hypertension. Mid or late diastolic murmurs are found at the lower sternal borders in patient.

2.3 Heart Sounds

There are two major sounds: The first heart sound caused of the closing mitral and tricuspid valves. The sound produced by the closure of the mitral valve is termed M1 and the sound produced by closure of the tricuspid valve is termed T1. That mean first heart sound content of two components (M1 and T1). The M1 sound is much louder than the T1 sound due to higher pressures in the left side of the heart, thus M1 radiated to all cardiac listening posts (loudest at the apex) and T1 is usually only heard at the left lower sternal border. The M1 sound is thus the main component of S1.
The first heart sound (S1) is a relatively low frequency sound usually described as a 'lub'. It marks the beginning of mechanical systole and therefore starts some time shortly after the beginning of the QRS complex of the ECG. The word murmur describes a swishing sound made as the blood flows through any of the heart's chambers or valves. It makes a sound like water rushing through a pipe. A heart murmur is a continuous sound that is audible with a common stethoscope, produced when blood

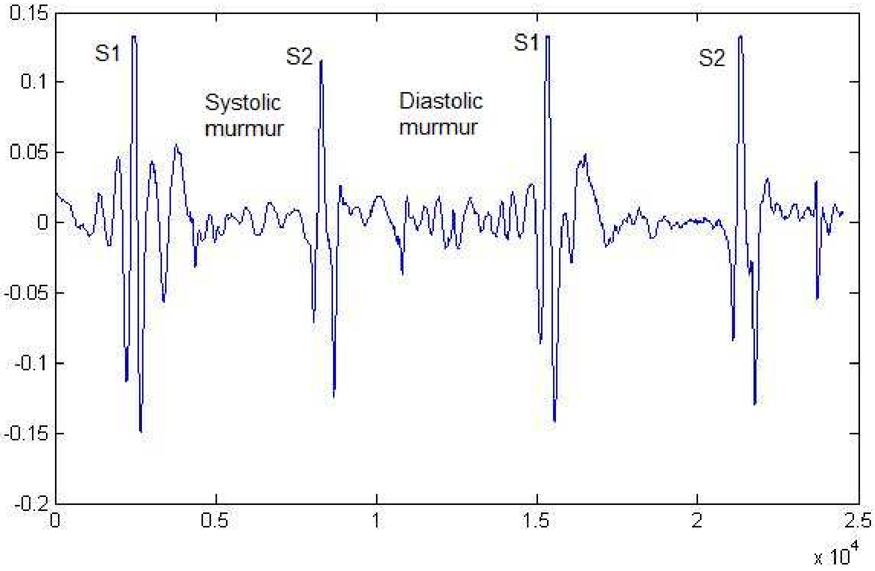


Figure 2-2: Sample of two cycle heart sound, where components S1, S2 and heart murmurs are highlighted.

passes through particular areas of the heart. Systolic murmurs occur between S1 and S2 (first and second heart sounds) and therefore are associated with mechanical systolic and ventricular ejection.

The second heart sound is created by the closing of the aortic and pulmonic valves. The aortic component of S2 produced by the closure of the aortic valve is termed A2 and the sound produced by the closure of the pulmonic valve is termed P2. The A2 sound is normally much louder than the P2 due to higher pressures in the left side of the heart, thus A2 radiates to all cardiac listening posts (loudest at the right upper sternal border) and P2 is usually only heard at the left upper sternal border. The A2 sound is thus the main component of S2.

Diastolic murmurs occur after S2 and before S1; they are therefore associated with ventricular relaxation and filling. Diastolic murmurs include aortic and pulmonic regurgitation (early diastolic), and mitral or tricuspid stenosis (mid-late diastolic). Tricuspid stenosis is very rare and is discussed further in the valvular heart disease section. As in Figure 2-2 are shown.

Basic cardiac sound signals are mostly comprised of four sound classes: two outstanding sounds named as the first heart sound (S1) and the second heart sound

(S2), and two weak sounds named as the third (S3) and the fourth heart sounds (S4). These four sounds may be audible by the auscultation of heart and occur in the frequency range of 20-200 Hz. However, most researches will restrict the S1 and S2 because S3 and S4 appear at very low amplitudes with low frequency components and are difficult to be caught in usual auscultation. As for heart defect, in the meanwhile, the unitary murmurs as a systolic ejection murmur (e.g., aortic stenosis) and a pansystolic murmur (e.g., mitral regurgitation) mostly appear between the S1 and S2 with different noise patterns like the diamond and rectangular shapes.

Heart murmurs are often the first sign of pathological changes of heart valves, and they are usually found during auscultation in primary health care. Heart murmurs are an important feature to identify cardiac disorders in childhood, infancy, and especially in newborns. Unrecognized heart disease in newborns carries a serious risk of avoidable mortality, morbidity and handicap [1]. The main advantages for early recognizing a cardiac disease are that newborns will be seen and assessed earlier and in better clinical conditions.

Cardiac murmurs occur frequently in healthy children, but it can also be a feature associated to many forms of congenital heart disease, including regurgitation, stenosis of heart valves, left to right shunt lesions at the atrial, ventricular, or great arterial levels. Careful examination reveals innocent systolic murmurs in about 72% of school-age children. A high prevalence of innocent murmur also has been documented in infant and neonates. Seven types of innocent heart murmurs are reported in children, i.e. pulmonary flow murmur, innocent pulmonary branch murmur of infancy, supraclavicular bruit, venous hum, mammary souffle, and cardiorespiratory murmur. Generally, clinical history and physical examination are diagnostic for these murmurs. Traditionally, heart auscultation is a screening method for early diagnosis of heart diseases and it is still a main diagnostic method for discovering them in clinical evaluation. When a physician visits a newborn with phonocardiography (PCG), heart murmurs are the most common abnormal auscultation finding. When a murmur is detected, the physician must decide whether to classify it as pathological or innocent. Heart murmurs are the most important feature for detecting a cardiac disorder, as they are often the first sign of pathological changes of heart valves. They are an extra and swishing sound caused by blood flowing through any of the heart's chambers or valves.

A heart murmur is a swishing sound heard when there is turbulent or abnormal blood flow across the heart valve that a heart murmur is an extra or unusual sound heard during a heartbeat. Murmurs range from very faint to very loud. Sometimes they sound like a whooshing or swishing noise. Generally, heart murmurs are divided in two categories: physiological and pathological murmurs.

We used two categories for this work normal or innocent murmurs that are very common. innocent murmurs are very common, occurring in up to 80% of children at some time or other. These murmurs are frequently detected during a febrile illness and are also exacerbated by nervousness or on exercise. It is important to distinguish between innocent and pathological murmurs and to arrange more detailed evaluation of the child if there is any doubt.

Many conditions may cause the blood to flow with turbulence, leading to a heart murmur on auscultation. All these conditions do not necessary indicate abnormality and cause no ill effect on health. Some of the main conditions causing an innocent heart murmur are: Small blood vessels to the lungs, this is because while they were in their mothers' uterus, there was very little blood flow to the lungs since babies do not breathe prior to birth. This will cause the blood vessels to the lungs to be small. Once the child is born, blood flow increases tremendously to the lungs, this will cause blood to be turbulent as it crosses these relatively small blood vessels, this turbulence will produce a heart murmur. Or Blood flow through the aortic valve and pulmonary valve (Physiologic pulmonary flow murmur): Blood flow across these two valves is audible in some newborn. This is not because there is anything wrong with these valves, but it may be due to the fact that newborns have a faster heart rate, which means that blood normally travels with a higher speed causing noise, resulting in the heart murmur. Also, newborns have a thinner chest wall, which allows sounds to be more readily audible. These innocent murmurs will disappears with growing heart.

Another category that we used in this work includes pathological systolic murmurs. Systolic murmurs maximal at the upper sternal borders are more likely to be ejection in type due to heart outflow abnormality or increased flow - aortic valve, sub valve or supravalve stenosis and HOCM being maximal on the right radiating to the neck whilst pulmonary valve, sub valve or supravalve stenosis or atrial septal defect murmurs are louder on the left and radiate to the back. Those at the lower sternal border are more likely to be of regurgitate type due to ventricular septal defect, mitral or tricuspid regurgitation. Some pathological systolic murmurs are heard widely over the whole precordium and different types of murmur may coexist. Coarctation of the aorta is an important cause of a murmur over the back, particularly in the inter scapular region.

Also called functional murmur, the former does not require follow-up visits by a cardiologist, whereas the latter can be related to serious disease conditions. Unfortunately, physiological murmurs are often similar to pathological ones. Among typical physiological murmurs, let us recall PDA (Patent Ductus Arteriosus), often detected shortly after birth, which is a very common disease in premature newborns. In patients affected by PDA, the blood circulates abnormally between two of the major arteries near the heart, due to the failure of a blood vessel (the ductus arteriosis) between these arteries to properly close. In most cases, the only symptom of PDA is a heart murmur, which continues until the ductus closes on its own, usually shortly after birth.

Major problems in newborns may depend on congenital heart diseases such as PDA, Atrial Septal Defect (ASD) and Ventricular Septal Defect (VSD). All these defects are expected to simply resolve on their own, as the child grows. However, especially in premature newborns, sometimes they may not. For instance, the ductus arteriosus in newborns may not close on its own, or it may be large and permit too much blood to pass through the lungs, which can place extra strain on the heart, forcing it to work harder and causing a rise in blood pressure in the arteries of the lungs. If this is the case, a medication or, rarely, surgery may be needed to close the PDA.

Although an innocent heart murmur does not entail a disease condition, a physician

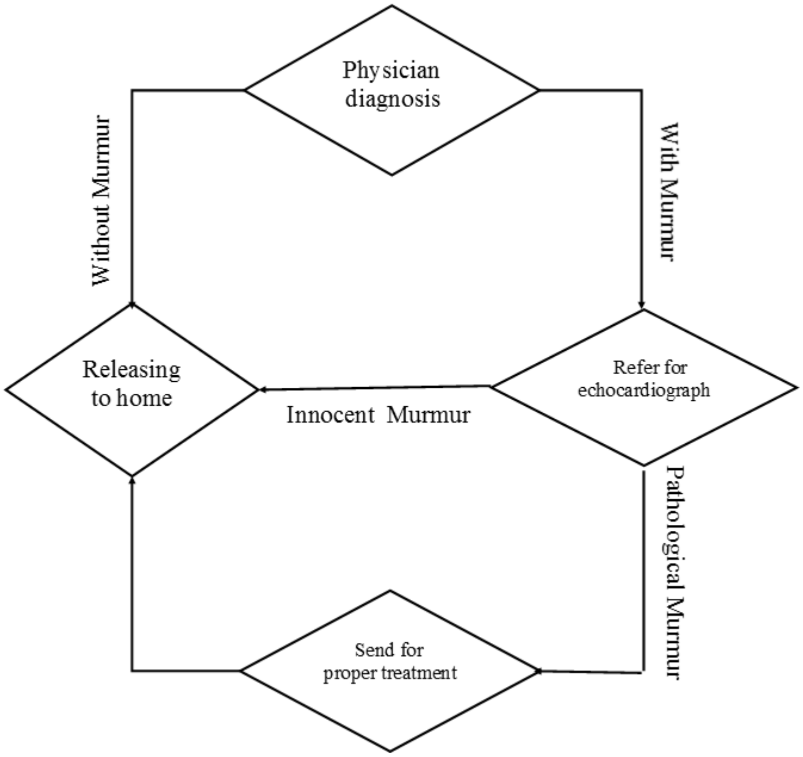


Figure 2-3: Discharge procedures for newborns from the hospital.

assuming that a newborn is healthy typically orders an echocardiogram for reassurance, although its cost may be not negligible. Overall, the result of this practice is a misallocation of health care funds. Indeed, while it is clearly important to avoid *type-I* errors, i.e. healthy newborn sent for echocardiogram, it is also important to avoid *type-II* errors, i.e. newborns having a pathological heart murmur sent home without proper treatment (on this matter see also Figure 2-3, which describes the typical discharge procedure from a hospital related to hearth murmurs).

2.4 Data Acquisition

Phonocardiography, diagnostic technique that creates a graphic record, or phonocardiogram, of the sounds and murmurs produced by the contracting heart, including its valves and associated great vessels. The phonocardiogram is obtained either with a chest microphone or with a miniature sensor in the tip of a small tubular instrument

Figure 2-4: An electronic stethoscope for recording heart sound.

that is introduced via the blood vessels into one of the heart chambers.

In 1816, the French physician Rene Laennec invented the first stethoscope using a long, rolled paper tube to funnel the sound. The word stethoscope is derived from the two Greek words, stethos (chest) and scopos (examination). Apart from listening to the heart and chest sounds, it is also used to hear bowel sounds and blood flow noises in arteries and veins. Throughout the 20th century many minor improvements were made to these iconic devices to reduce weight, improve acoustic quality, and filter out external noise to aid in the process of auscultation. Electronic versions of the stethoscope were introduced to further amplify sound. Stethoscopes are now available in a wide array of styles, with designs available for virtually every branch of medicine.

An electronic stethoscope has been used to record heart sounds which was connected to a voice recorder (see Figure 2-4). Since newborn's heart is small, a single point recording data is proposed for study. The data recorded for 12 seconds from newborns and all data is labeled by a cardiologist after checking newborn by echocardiography.

2.5 Previous work

Cardiac auscultation was used by physicians since early nineteenth century and before that heart sounds could be listened by applying their ear directly to the chest, although human ears are poorly for cardiac auscultation. After the stethoscope was introduced by Laennec in 1816,"mediate auscultation" became possible, introducing an exciting and practical new method of bedside examination. Over the past 2 centuries, many illustrious physicians have contributed to the understanding of cardiac auscultation by providing an explanation for the sounds and noises that are heard in the normal and diseased heart [36]. Auscultation remains a low cost, though sophisticated procedure that intimately connects the physician to the patient and transfers that all-important clinical power known as "the laying on of the hands". When used with skill, it may correctly determine whether more expensive testing should be ordered. In this way, the stethoscope deserves our continued respect and more attention as an indispensable aid for the evaluation of our patients.

Prior works performed on heart murmur are concerned with various stages of life and approaches in feature extraction (signal processing) and classification techniques. Different tools are used for feature extraction and classification of heart sounds Christer Ahlstrom[1] use Phonocardiographic signals that were acquired from 36 patients with aortic valve stenosis, mitral insufficiency or physiological murmurs, and the data were analyzed with the aim to find a suitable feature subset for automatic classification of heart murmurs. Techniques such as Shannon energy, wavelets, fractal dimensions and recurrence quantification Analysis were used to extract 207 features. 157 of these features have not previously been used in heart murmur classification. A multi-domain subset consisting of 14, both old and new, features was derived using Pupil's sequential floating forward selection method. This subset was compared with several single domain feature sets. Using neural network classification, the selected multi-domain subset gave the best results; 86% correct classifications compared to 68% for the first runner-up. In conclusion, the derived feature set was superior to the comparative sets, and seems rather robust to noisy data.

Curt G. DeGroff [23] Used an electronic stethoscope to record heart sounds from 69 patients (37 pathological and 32 innocent murmurs). Sound samples were processed using digital signal analysis and fed into a custom ANN. With optimal settings, sensitivities and specificities of 100% were obtained on the data collected with the ANN classification system developed. For future unknowns, our results suggest the generalization would improve with better representation of all classes in the training data. This work demonstrated that ANNs show significant potential in their use as an accurate diagnostic tool for the classification of heart sound data into innocent and pathological classes. This technology offers great promise for the development of a device for high-volume screening of children for heart disease.

A diagnostic system designed by S. L. Strunic [65] based on Artificial Neural Networks (ANN) that can be used in the detection and classification of heart murmurs. Segmentation and alignment algorithms serve as important pre-processing steps before heart sounds are applied to the ANN structure. The system enables users to create a classifier that can be trained to detect virtually any desired target set of heart sounds. The output of the system is the classification of the sound as either normal or a type of heart murmur. The ultimate goal of this research is to implement a heart sounds diagnostic system. Testing has been conducted using both simulated and recorded patient heart sounds as input. The system was able to classify with up to 85.4% accuracy and 95.8% sensitivity.

Cota Navin Gupta [34] present a novel method for segmentation of heart sounds (HSs) into single cardiac cycle (S1-Systole-S2-Diastole) using homomorphic filtering and K-means clustering. Feature vectors were formed after segmentation by using Daubechies-2 wavelet detail coefficients at the second decomposition level. These feature vectors were then used as input to the neural networks. Grow and Learn (GAL) and Multilayer perceptron Back propagation (MLP-BP) neural networks were used for classification of three different heart sounds (Normal, Systolic murmur and Diastolic murmur). It was observed that the classification performance of GAL was similar to MLP-BP. However, the training and testing times of GAL were lower as compared to MLP-BP. The proposed framework could be a potential solution for automatic analysis of heart sounds that may be implemented in real time for classification of HSs.

3

Data Pre-processing

We describe various preprocessing techniques in this chapter. The aim of signal preprocessing is to remove noise and prepare the raw signal for further processing. Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure or data preprocessing is a data mining technique that involves transforming raw data into an understandable format.

Commonly used as a preliminary data mining practice, data preprocessing transforms data into a format that will be more easily and effectively processed for the purpose of the user for example, in a neural network. There are a number of different tools and methods used for preprocessing, including: sampling, which selects a representative subset from a large population of data; transformation, which manipulates raw data to produce a single input; denoising, which removes noise from data; normalization, which organizes data for more efficient access; and feature extraction, which pulls out specified data that is significant in some particular context.

3.1 Down Sampling

Pre-processing occurs in two steps: The first step of signal processing is filtering heart sounds, with the goal of removing the unwanted noise. The recording of PCG usually has a sampling frequency higher than 8000Hz. In the event that the recording

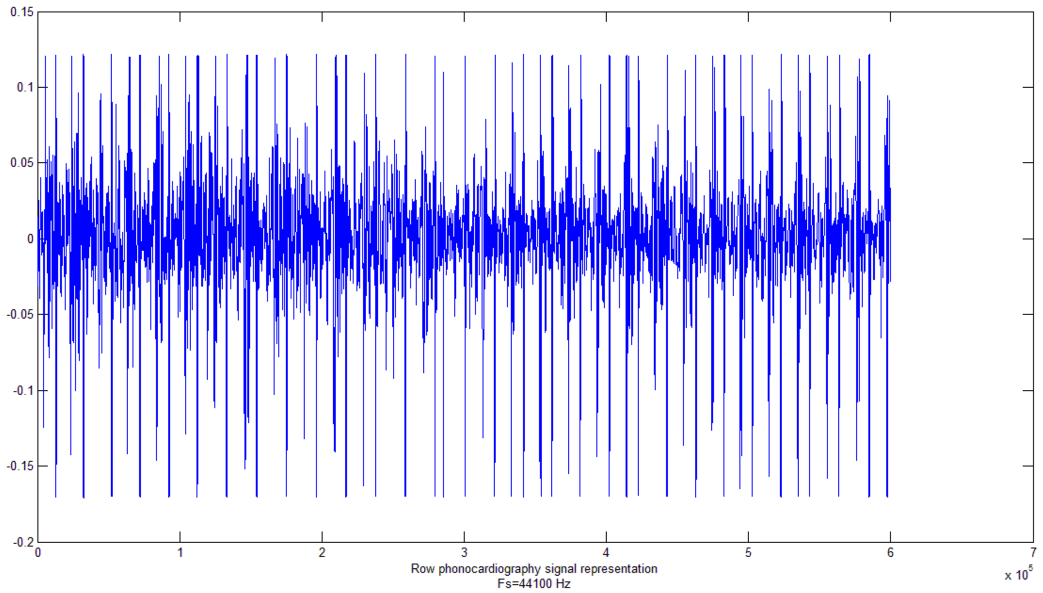


Figure 3-1: Row signal of a heart sound.

environment cannot be controlled enough, noise is coupled into the PCG. To avoid unpredictable effects brought by noise, filtering becomes important for later processing. An electronic stethoscope has been used to record heart sounds, giving rise to a dataset at 44k Hz which is shown in Figure 3-1

Heart sounds data recorded with 44100 Hz frequency sample rate. Feature extraction (signal processing) in this sample rate due to increasing a number of computations. A useful component in DSP systems is the down sampler, which can be used to lower the effective sampling rate at which a signal has been sampled. There are a variety of reasons why such a device can be useful. As a first example, CD quality audio recording is typically sampled at 441000 Hz. However, if the sampled signal is a human voice, a perfectly intelligible signal can be reconstructed from an 8 kHz sampled signal. The downsampling is a tool to make a digital audio signal smaller by lowering its sampling rate or sample size (bits per sample). Downsampling is done to decrease the bit rate when transmitting over a limited bandwidth or to convert to a more limited audio format. Consider down-sampling a signal $x(n)$ and reducing the sampling rate by a factor M as shown in Figure 3-2 and the output is defined as:

$$y(m) = x(mM) \tag{3.1}$$

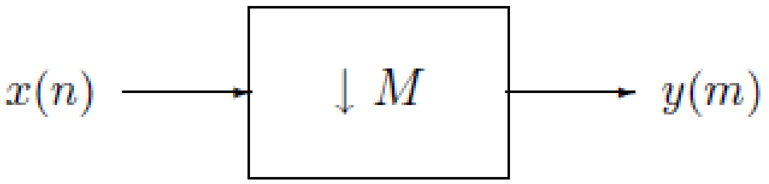


Figure 3-2: Down sampling rate by the factor M.

i.e., it consists of every M^{th} element of the input signal. It is clear that the decimated signal y does not in general contain all information about the original signal x. Therefore, decimation is usually applied in filter banks and preceded by filters which extract the relevant frequency bands.

In order to analyze the frequency domain characteristics of a multirate processing system with decimation, we need to study the relation between the Fourier transforms, or the z-transforms, of the signals x and y. For simplicity we consider the case $M = 2$ only. Then the decimated signal y is given by

$$y(m) = x(2M) \tag{3.2}$$

$$or \qquad y(m) = x(0), x(2), x(4), ... \tag{3.3}$$

$$\tag{3.4}$$

Given the z-transform of $\{x(n)\}$,

$$\hat{x}(z) = x(0) + x(1)z^{-1} + x(2)z^{-2} + ... + x(n)z^{-n} + ... \tag{3.5}$$

We should like to have an expression for the z-transform of $\{y(m)\}$,

$$\begin{aligned} \hat{y}(z) &= y(0) + y(1)z^{-1} + y(2)z^{-2} + ... + x(n)z^{-m} + ... \\ &= x(0) + x(2)z^{-1} + x(4)z^{-1} + ... + x(2m)z^{-1} + ... \end{aligned} \tag{3.6}$$

In order to derive an expression for the z-transform of y, it is convenient to represent the decimation in two stages as follows. First, define the signal:

$$v(n) = x(0), 0, x(2), 0, x(4), 0, ... \tag{3.7}$$

which has the z-transform

$$\hat{v}(z) = x(0) + x(2)z^{-2} + x(4)z^{-4} + ... + x(2m)z^{-2m} + ... \tag{3.8}$$

As

$$\hat{x}(-z) = x(0) - x(2)z^{-2} + x(4)z^{-4} - ... + x(2m)z^{-2m} + ... \tag{3.9}$$

it follows that

$$\hat{v}(z) = \frac{1}{2}\left(\hat{x}\left(z\right) + \hat{x}\left(-z\right)\right) \tag{3.10}$$

By (3.6) and (3.8), $\hat{y}(z) = \hat{v}(z^{1/2})$. Hence, we have obtained the relation

$$\hat{y}(z) = \frac{1}{2}(\hat{x}(z^{1/2}) + \hat{x}(-z^{1/2})) \tag{3.11}$$

In order to determine the frequency domain characteristics of the decimated signal $\{y(m)\}$, recall that the Fourier transform is related to the z-transform according to:

$$Y(\omega) = \hat{y}(z)|_{z=e^{j\omega}} \tag{3.12}$$

Hence, we have from (3.11),

$$Y(\omega) = \frac{1}{2}(\hat{x}(e^{j\omega/2}) + \hat{x}(-e^{j\omega/2})) \tag{3.13}$$

Noting that$-1 = e^{j\pi}$

$$\begin{aligned} Y(\omega) &= \frac{1}{2}(\hat{x}(e^{j\omega/2}) + \hat{x}(e^{j\omega/2+\pi}) \\ &= \frac{1}{2}(X(\omega/2) + X(\omega/2+\pi)) \end{aligned} \tag{3.14}$$

where X is the Fourier-transform of the sequence $x(n)$. But from the properties of the Fourier transform (periodicity and symmetry) it follows that $X(\omega/2 + \pi) =$

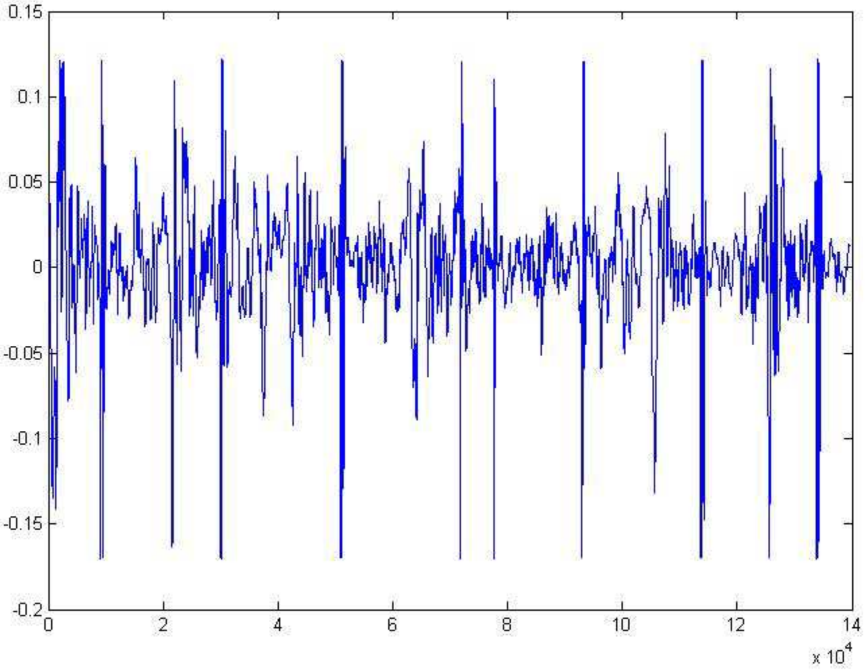


Figure 3-3: Heart sound signal with frequency sample rate 44100

$X(\omega/2-\pi) = X(\pi-\omega/2)^*$. Hence

$$Y(\omega) = \frac{1}{2}(x(\omega/2) + x(\pi-\omega/2)^*) \tag{3.15}$$

The Fourier-transform of $\{y(m)\}$ thus cannot distinguish between the frequencies $\omega/2$ and $\pi-\omega/2$ of $\{x(n)\}$. This is equivalent to the frequency folding phenomenon occurring when sampling a continuous-time signal.

Hence, while the signal $\{x(n)\}$ consists of frequencies in $[0,\pi]$, the frequency contents of the decimated signal $\{y(m)\}$ are restricted to the range $[0,\pi/2]$. Moreover, after decimation of the signal $\{x(n)\}$, its frequency components in $[0,\pi/2]$ cannot be distinguished from the frequency components in the range $[\pi/2,\pi]$.

The row heart sound signal is shown in figure 3-3 which the sampling rate is 441000 Hz. The above sequence of numbers represent the indicts of the samples of a signal prior to down sampling, the bottom sequences of numbers represent the resultant indicts of the signal after subjecting it to downsampling operation. Figure 3-4 shows

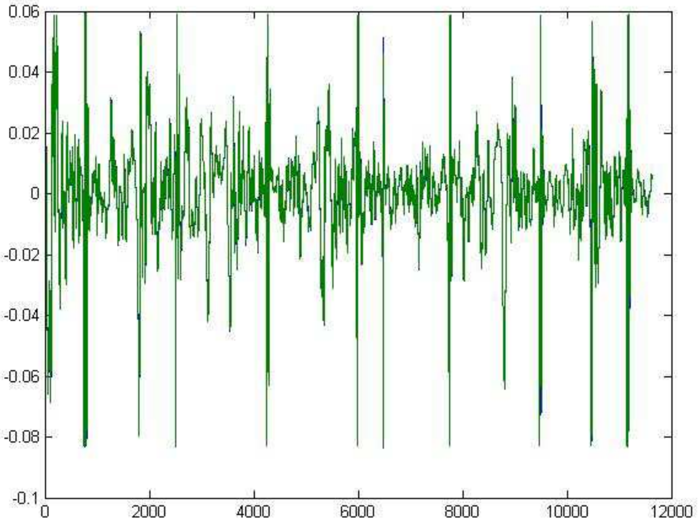


Figure 3-4: Heart sound signal after downsampling.

how to use down sample to obtain the phases of a signal. Downsampling a signal by $M = 11$ can produce M unique phases. To avoid unpredictable effects brought by noise, filtering out the unwanted noise becomes important for later processing.

3.2 Noises and Filtering

In many situations, the PCG is recorded in hospital that the signal is corrupted by different types of noise, sometimes originating from another physiological process of the body such as respiratory sound. Hence, noise reduction represents another important objective of PCG signal processing; in fact, the waveforms of interest are sometimes so heavily masked by noise that their presence can only be revealed once appropriate signal processing has first been applied.

Removal and measurement of the noises divided in three categories: baseline wander (changes in the baseline signal, principally due to respiration and microphone movement artifacts); power-line (due to the power distribution network) and residual noise (including noise arising from the myoelectric potentials of skeletal muscles due to patient movement).

The main spectrum of first and second (s1 and s2 respectively) heart sound occurs

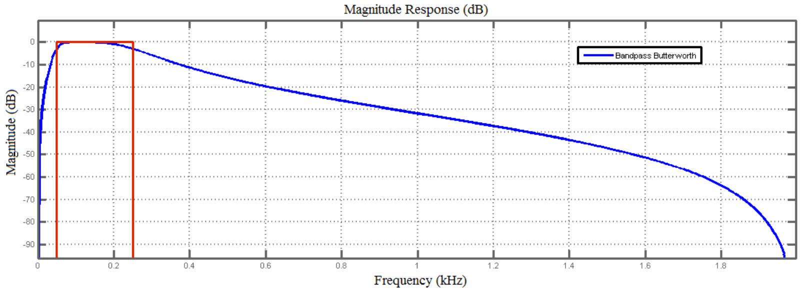


Figure 3-5: Illustration of 4th order a band-pass Butterworth filter.

within the range of 50 and 250Hz, hence a band pass filter is used. A band-pass filter is a device that passes frequencies within a certain range and rejects (attenuates) frequencies outside that range. Butterworth filter is implied to reducing noises. The Butterworth filter is a type of signal processing filter designed to have as flat frequency response as possible (no ripples) in the pass-band and zero roll off response in the stop-band. The Butterworth filters are one of the most commonly used digital filters in motion analysis and in audio circuits. They are fast and simple to use. Since they are frequency-based, the effect of filtering can be easily understood and predicted. Proposed filter is a band pass filter with fourth order as shown in figure 3-5.

3.3 Heart Sound Segmentation

The third step of preprocessing for detecting systolic murmurs uses a segmentation algorithm aimed at identifying the heart sound components S1 and S2. The detection can be manual and automatic. In this study we'll consider both.

The manual method is based on the timing between high amplitude components. The fact that the time interval that occurs between S1 and S2 (systole) is always less than the one between S2 and S1 (diastole) is the basis for this process.

Typically, heart sounds consist of two regularly repeated thuds, known as S1 and S2 and appearing one after the other for every heartbeat. The time interval between S1 and S2 is the systole, while the gap between S2 and the next S1 corresponds to the diastole. Currently, the detection is performed manually; however, we are planning

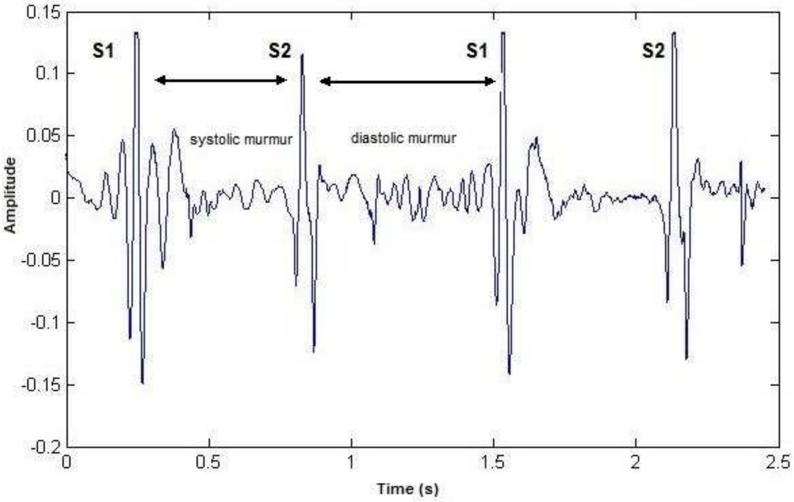


Figure 3-6: Heart sound segmentation manually.

to identify of S1 and S2 with an automatic procedure in the next future. The segmentation method is based on the timing between those high amplitude components. The fact that the time for systole is always less than the time for diastole is the basis for this process.
The largest section of systolic and diastolic murmur, common to all database samples, was chosen for analysis (represented with 400 and 700 points, for systolic and diastolic murmurs respectively). The main characteristics of heart sounds, such as their timing relationships and components, frequency contents, their occurrence in the cardiac cycle, and the envelope shape of murmurs can be quantified by means of advanced digital signal processing techniques. figure 3-7 shows the interval distance between two major sounds (S1 and S2).

The second method to detect and segment heart sounds that is performed by automatically algorithm. The segmentation method is based on the timing between high amplitude components. The basis for this process is that the time interval that occurs between S1 and S2 (systole) is always smaller than the one between S2 and S1 (diastole). Even after per-processing, the actual heart sound signal still has very complicated patterns with numerous small spikes that have little impact on diagnosis but may influence the location of S1 and S2. Peak conditioning was performed for the

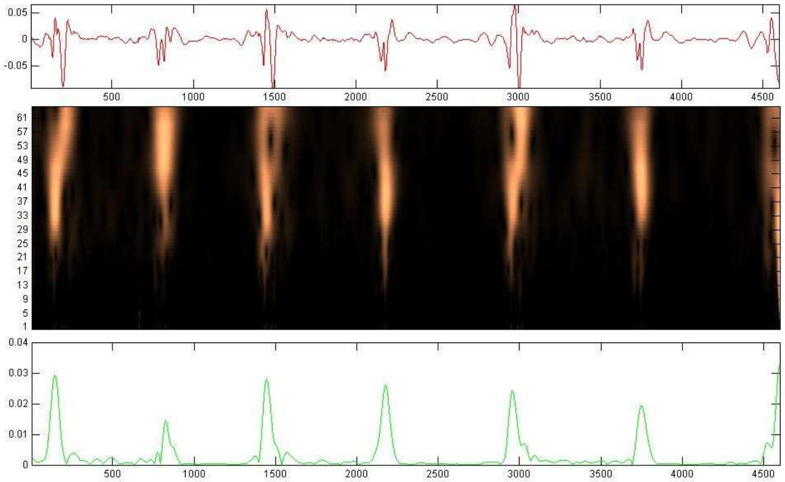


Figure 3-7: Illustration original signal (top), wavelet coefficients scale colored (middle), Coefficients line (bottom).

obtained peaks using wavelet transform, which enabled the cycle detection process. There is a wavelet transfer introduce for detecting peaks. To find peak location we used the Complex Morlet (Gabor) Wavelet (CMW) transfer which are very popular in biomedical data analysis for time-frequency decomposition. To this end, we used the Gabor Wavelet for peak detection (see [35]), which can be formally described as follows:

$$\Psi(t) = C \cdot e^{-jwt} \cdot e^{-t^2} \tag{3.16}$$

where $e^{-jwt} \cdot e^{-t^2}$ is the complex Gaussian function and C is a normalizing constant. The threshold was used to identify the peaks. A threshold value was set to 0.1 for wavelet scale coefficients (see Figure 3-9). We identified cardiac cycle peaks using K-mean clustering. K-means is a non-hierarchical partitioning and simplest unsupervised learning algorithms which method is partitions the observations in the data into K mutually exclusive clusters, and returns a vector of indicating to which of the K clusters it has assigned each observation. It uses an iterative method that minimizes the sum of distances from each object to its cluster centroid, over all clusters.

The main idea is to define K centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes varies result. So, the better choice is to place them as much as possible far away from each other.

The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group-age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these K new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.
Each class of heart murmurs contains distinctive information in time and frequency domains. This stage involves the extraction of each cardiac cycle of the PCG signal, after the peak detection and peak conditioning stages.
Systolic (S1-S2) and diastolic (S2-S1) murmurs occur within the time intervals that were calculated by the peak conditioning process. These time intervals were clustered into two clusters. Cluster 1 and cluster 2 occur consecutively and indicate a single cardiac cycle. The cycle smaller time interval was then identified as systole while the other interval was identified as diastole. Consecutive occurrence of cluster 1 or cluster 2 might be due to loss of peak, extra peak or equal systolic and diastolic intervals [34]. We extracted each single cycle of PCG signals using clusters, as shown in Figure 3-8 for normal heart sound case, heart sound with systolic and diastolic murmur, respectively.

In other side, there is another way to segment heart sounds after peak detection is using a threshold. Zero-crossing is used to find the spots where peaks occur, the number of zero crossings per segment being also an equivalent representation of the dominant component of a signal segment. The algorithm calculates the size of the intervals in which the value of the function is zero. Let us recall that systolic (S1-S2) and diastolic (S2-S1) murmurs occur respectively in the smaller and bigger time interval as in Figure 3-9 shows.

After detecting cardiac cycles, it is important to identify which cycle shows more signs of heart disease –as detecting the most informative cycle can optimize the performance of the classification process. For each patient, recorded data include several

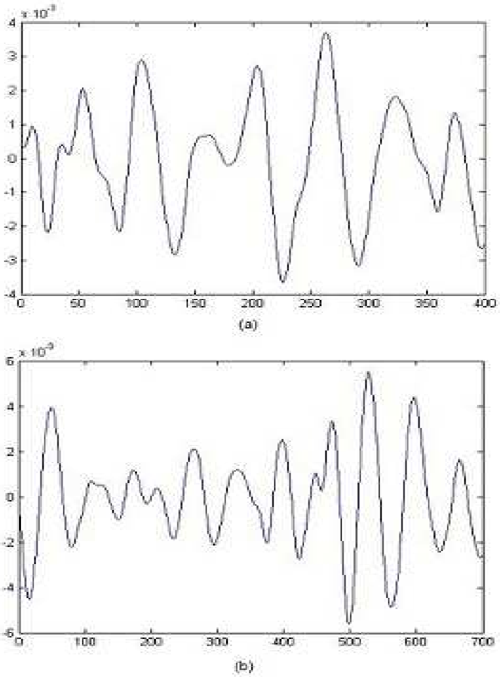


Figure 3-8: a) Systolic murmur (b) Diastolic murmur.

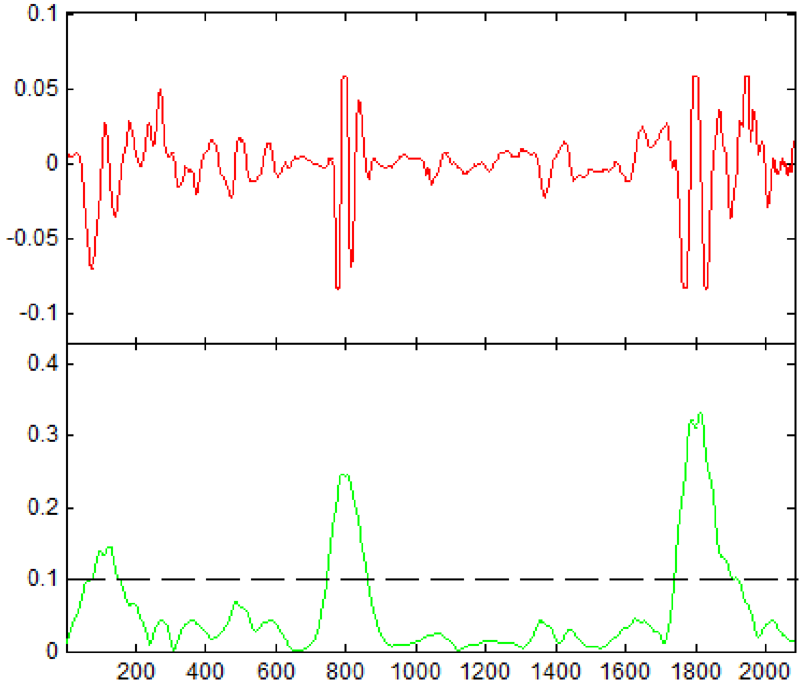


Figure 3-9: Peak detection using Complex Gabor Wavelet and thresholded by zero-crossing.

cardiac cycles on a time span of few seconds. Despite the fact that filtering has been implemented to remove noise, the residual noise may be part of the heart sound signal –such as respiratory sound, artifact noise or newborn voices. Pearson's correlation coefficient has been used to select the cycle with minimum noise and most properties of the whole signal. Pearson's correlation coefficient between two signals X and Y (cardiac cycles, in this context) is defined as follows:

$$r_{XY} = \frac{SS_{XY}}{\sqrt{(SS_{XX})(SS_{YY})}} \tag{3.17}$$

where:[1]

- $SS_{XY} = \sum XY - \frac{1}{n} \cdot \sum X \cdot \sum Y$
- $SS_{XX} = \sum X^2 - \frac{1}{n} \cdot (\sum X)^2$
- $SS_{YY} = \sum Y^2 - \frac{1}{n} \cdot (\sum Y)^2$

The Pearson correlation coefficient is calculated between pairs of signals, each signal including several cycles. An example of correlation is is shown in Figure 3-10.

The overall correlation for each cardiac cycle $C_{i,j}$ is obtained through the following formula:

$$r_{C_i} = \frac{1}{n-1} \sum_{i \neq j}^{j} r_{C_{i.j}} \tag{3.18}$$

where n is the number of cardiac cycles. In the given example C_4 is selected according to Equation (3.18) as the cycle with minimum noise (and hence with the most informative content for the whole signal).

[1] n is number of data pairs for each sample; in this case 1500.

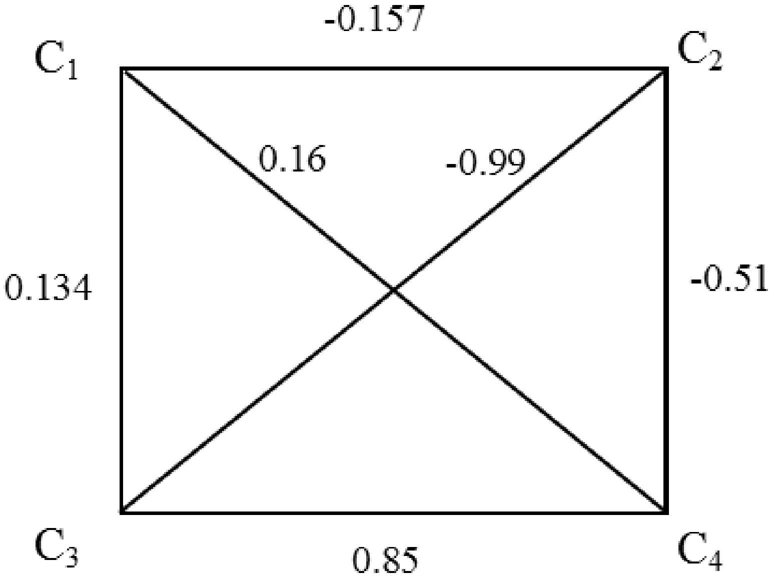


Figure 3-10: Correlation among cardiac cycles

4

Signal Processing Techniques and Analysis

This chapter gives an overview of some of the signal processing techniques used in this thesis. Feature extraction is introduced and a technique for feature extraction (which presents the theoretical background of time-frequency methods) used to analyze phonocardiographic data. Localizing information in both time and frequency domains is the main propose of time-frequency analysis. The purpose of this phase is developing an automatic heart sound signal analysis system and select feature to using in intelligent systems which able to support the physician in the diagnosing of heart murmurs at early stage of life.

Heart murmurs are the first signs of heart disease. We will able to screen newborns for normal (innocent) and pathological murmurs. This chapter presents an analysis and comparisons of signal processing techniques and also, extracting and selecting signal features able to highlight significant properties of the PCG signal. These features have also undergone a selection process, aimed at identifying the most appropriate for classification purposes. As for feature extraction, several metrics have been taken into account, including Maximum value amplitude, Peak to Peak, Variance, Absolute negative area, Shannon energy, Bispectrum and Wigner Bispectrum [9]. Feature selection has then been used to reduce the size of the feature vector.

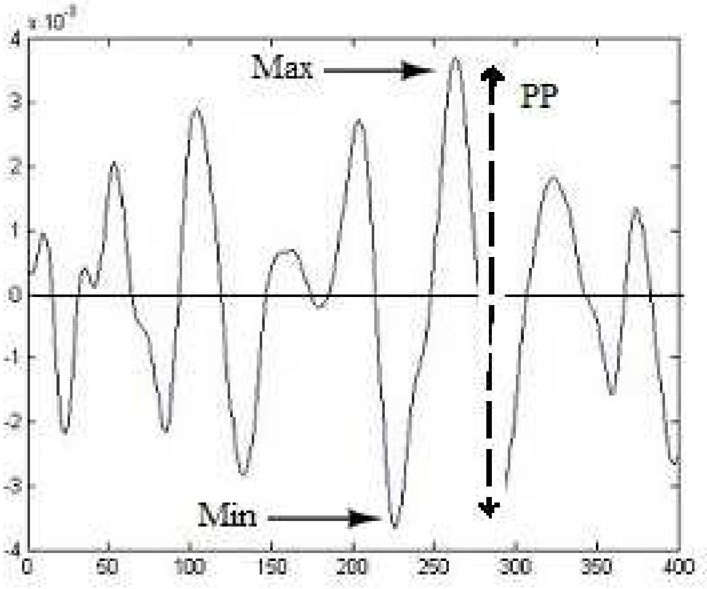


Figure 4-1: Maximum and minimum value amplitude of a heart sound signal.

4.1 Maximum and Minimum of Value Amplitude

The amplitude of a periodic variable is a measure of its change over a single period (such as time or spatial period). There are various definitions of amplitude which are all functions of the magnitude of the difference between the variable's extreme values. The maximum displacement of a vibrating particle of the medium from its mean position is called Amplitude. Here in the sound wave, amplitude represents the loudness of the sound which is opposite for minimum. Figure 4-1 shows maximum and minimum amplitude of a sample heart sound signal (figure 3-8).

4.2 Peak to Peak Amplitude

Peak-to-peak amplitude is a pretty simple concept which is the change between peak (highest amplitude value) and trough (lowest amplitude value, which can be negative). With appropriate circuitry, peak-to-peak (PP) amplitudes of electric oscillations can be measured by meters or by viewing the waveform on an oscilloscope. Peak-to-peak is a straightforward measurement on an oscilloscope, the peaks of the waveform being easily identified and measured against the graticule. This remains a common way of specifying amplitude, but sometimes other measures of amplitude are more appropriate. peak to peak amplitude of a heart sound signal is shown in figure 4-1.

4.3 Variance

A measurement of the spread between numbers in a data set. The variance measures how far each number in the set is from the mean. Variance is calculated by taking the differences between each number in the set and the mean, squaring the differences (to make them positive) and dividing the sum of the squares by the number of values in the set. A variance of zero indicates that all the values are identical. Variance is always non-negative: a small variance indicates that the data points tend to be very close to the mean (expected value) and hence to each other, while a high variance indicates that the data points are very spread out around the mean and from each other. We use the following formula to compute variance which we denote by σ^2 is defined as:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} \tag{4.1}$$

where μ is the mean, N is the number of data values, and X stands for each data value in turn.

4.4 Shannon Energy

The Shannon Energy is calculated on signal segments. Here we segment data each of 0.02-second and with 0.01-second signal segment overlapping throughout the signal. The normalized envelope of Shannon Energy is calculated to delimit the beginning and end of each of heart sounds using a fixed threshold from the maximum value of the envelope which allows to determinate their average duration. The average Shannon energy $E_s(t)$ for a frame t (see for example [1]) can be calculated on signal

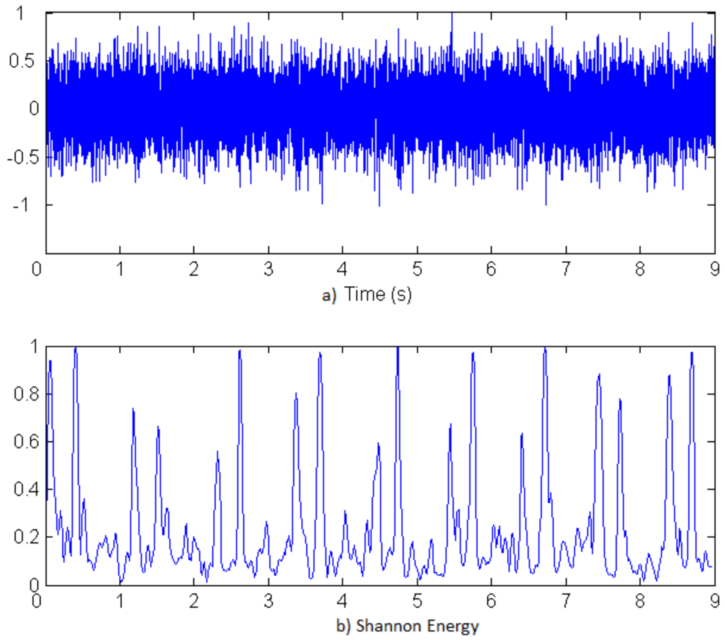


Figure 4-2: Shannon energy of PCG signal with additive Gaussian noise.

segments as follows:

$$E_s\left(t\right) = \frac{1}{N}\sum_{j=1}^{n} x_n^2 \log x_n^2 \tag{4.2}$$

where x_n is a normalized input signal, n the length of data, and N the signal length.

With $M(E_s(t))$ denoting the mean value of $E_s(t)$ and $S(E_s(t))$ the standard deviation of $E_s(t)$, the normalized average Shannon energy $N_{se}(t)$, also called Shannon envelope, is then calculated as follows:

$$N_{se}\left(t\right) = \frac{E_s\left(t\right) - M(E_s\left(t\right))}{S(E_s\left(t\right))} \tag{4.3}$$

In et al Ali Moukadem [53] presented a module for heart sounds segmentation based on time frequency analysis (S-Transform). The goal of this study is to develop a generic tool, suitable for clinical and home monitoring use, robust to noise, and applicable to diverse pathological and normal heart sound signals without the necessity of any previous information about the subject [3]. The proposed segmentation Shannon energy is used to detection of the localized heart sounds and classification block to distinguish between S1 and S2.

Figure 4-2 shows the robustness of Shannon energy method against white additive noise. In this study, the advantage of performing a time-frequency analysis which makes methods more robust against noise is proposed.

4.5 Spectrum

Techniques have recently been developed and demonstrated that allow the tracking of spectral parameters as time elapses. Approaches of this type have also been called time-variant spectral analysis or time-frequency analysis. For a detailed description of the algorithms and methodologies proposed and for some experimental studies, see references Basano et al. (1995), Bianchi et al. (1993), Cerutti et al. (1989), Keselbrener & Akselrod (1996), Lee & Nehorai (1992), Mainardi et al. (1994, 1995) and Novak et al. (1993). The advantages of these methodologies are associated mainly with reducing the influence of non-stationaries and monitoring transient cardiac events occuring in long-term recordings. In Cerutti et al. (1989), a procedure of compressed spectral arrays (CSA) was implemented which can reduce the spectral data obtained from 24 hour ambulatory ECG recordings. The method was based on the calculation of AR spectral estimates for successive RR interval segments, and checking whether a new spectrum differs significantly from the preceding one.

Spectral analysis, especially presentation of the dominant frequency, is a technique which is much more cost-effective than echocardiography and magnetic resonance imaging. The magnitude spectrum of an audio signal describes the distribution of magnitudes with frequency i.e. what frequencies (of pure tones) are present and at what amplitudes. The phase spectrum can display in what way the phase relationship between two signals varies with frequency. With the advent of miniaturized and powerful technology for data acquisition, display and digital signal processing, the possibilities for detecting cardiac pathology by signal analysis have increased. Permanent records permit objective comparisons of the acoustic findings [69, 24]. Many advanced methods for signal processing and analysis (e.g. sound spectral averaging techniques, artificial neural networks, time-frequency analysis and wavelet analysis) have been reported to be effective for presentation and analysis of cardiac acoustic signals [66]. Using electronic mail, it is possible to transmit the records to remote sites for further, sophisticated analysis and conclusions.

Let us define Short Time Fourier Transform (STFT) in the continuous-time case, the function to be transformed is multiplied by a window function which is nonzero for only a short period of time. The Fourier transform (a one-dimensional function) of the resulting signal is taken as the window is slid along the time axis, resulting in a two-dimensional representation of the signal. Mathematically, this is written as:

$$STFT\ x(\tau,\omega) = \int_{\infty}^{-\infty} x(t)\omega(t-\tau)e^{-j\omega t}dt \tag{4.4}$$

where $w(t)$ is the window function, commonly a Hann window or Gaussian window bell centered around zero, and $x(t)$ is the signal to be transformed. (Note the difference between w and ω) $X(\tau,\omega)$ is essentially the Fourier Transform of $x(t)w(t-\tau)$, a complex function representing the phase and magnitude of the signal over time and frequency. Often phase unwrapping is employed along either or both the time axis, τ, and frequency axis, ω, to suppress any jump discontinuity of the phase result of the STFT. The time index τ is normally considered to be "slow" time and usually not expressed in as high resolution as time t. In et. al S.M. Debbal [22], The paper is concerned with a synthesis study of the fast Fourier transform, the short-time Fourier transform, the Wigner Distribution (WD) and the wavelet transform in analyzing the phonocardiogram signal. It is shown that these transforms provide enough features of the PCG signals that will help clinics to obtain qualitative and quantitative measurements of the time-frequency PCG signal characteristics and consequently aid diagnosis. Similarly, it is shown that the frequency content of such a signal can be determined by the FFT without difficulties. The studied techniques of analysis can thus be regarded as complementary in the TF analysis of the PCG signal; each will relate to a part distinct from the analysis in question. The magnitudes of the Fourier transforms of S1 and S2 for these two cases are shown in Figure 4-3.

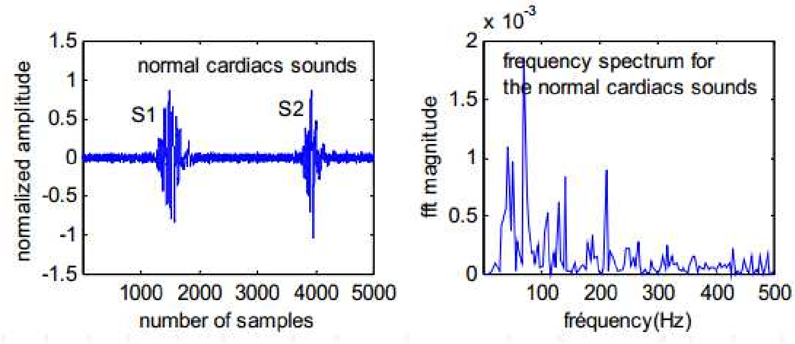


Figure 4-3: Fast Fourier Transform for the normal cardiac sounds (S1 and S2).

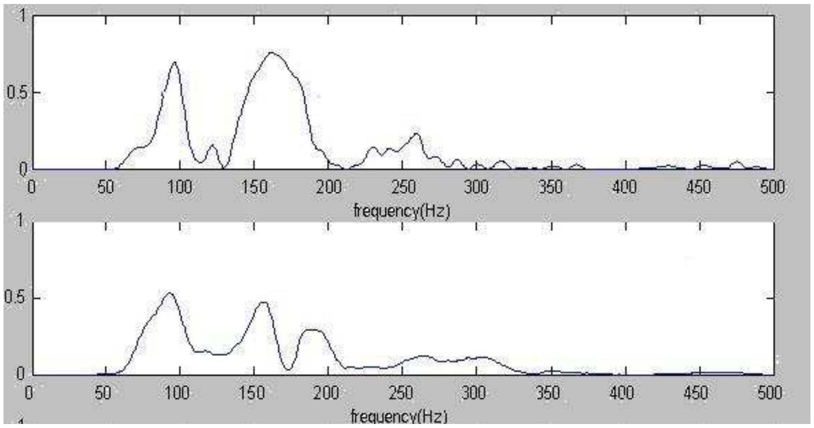


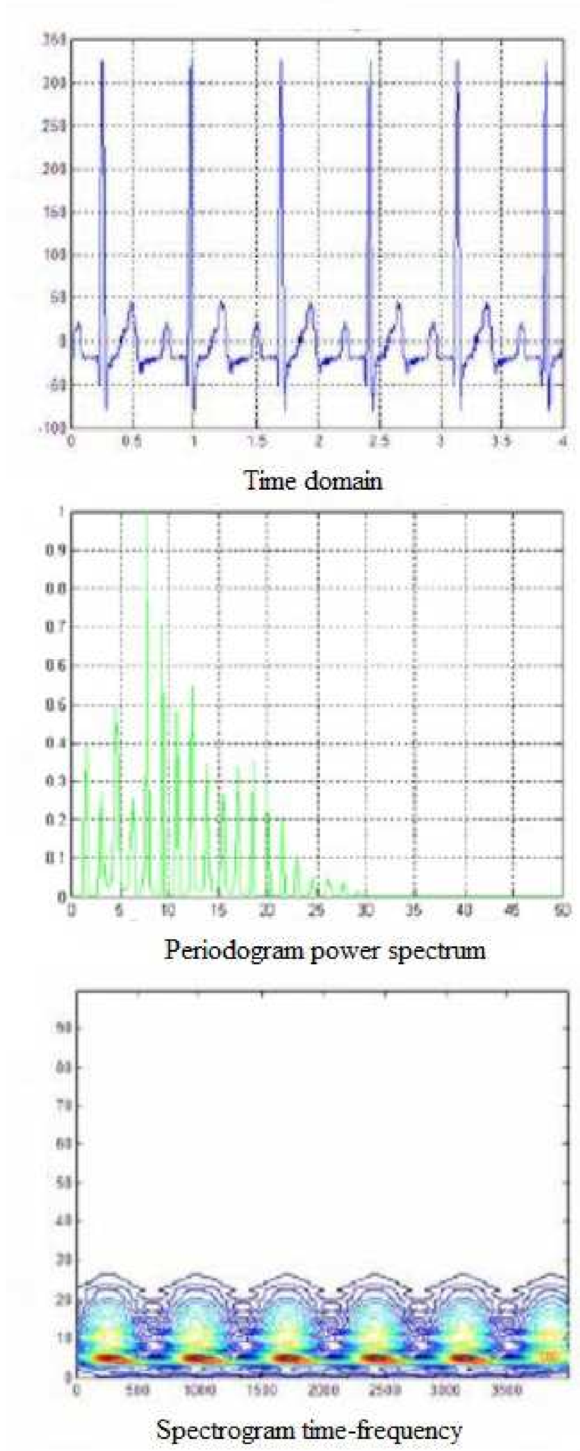
Figure 4-4: The magnitude responses of the S1 and S2 using FFT.

While sufficient study reveals that the basic features shown in the transfer functions exist in the Fourier transforms of the signals themselves (as they of course must), the distinguishing features between the cases are much more difficult to identify. As seen in Figure 4-4, FFT appeared two major component M1 and T1 for the sound S1 and A2 and P2 for the sound S2. Figure 4-4 illustrates the average FFT spectrum of a normal heart sound (first and second heart sound).

The spectrogram is defined as the squared modulus of Short Time Fourier Transform of a given signal $x(t)$. This transform is a liner projection combined with a quadratic operation which provides an energy estimation of the analyzed signal. As define:

$$Spectrogram\ x(\tau,\omega) \equiv |X(\tau,\omega)|^2 \tag{4.5}$$

The spectral density of a signal characterizes the distribution of the signal's energy or power in the frequency domain. This concept is particularly important when considering filtering in communication systems. We need to be able to evaluate the signal and noise at the filter output. The energy spectral density (ESD) or the power spectral density (PSD) is used in the evaluation. For instance, in et. al Norhashimah Mohd Saad [58], the authors discuss how to use digital signal processing approach for the detection of heart blocks in electrocardiogram (ECG) signals. Signal analysis techniques such as the periodogram power spectrum and spectrogram time-frequency analysis are employed to analyze ECG variations. Seven subjects are identified: normal, first degree heart block, second degree heart block type I, second degree heart block type II, Third degree heart block, right bundle branch block and left bundle branch block. Analysis results revealed that normal ECG subject is able to maintain higher peak frequency range (8 Hz), while heart block subjects revealed a significant low peak frequency range ($<$4 Hz). The results revealed that the periodogram power spectrum can be used to differentiate between normal and heart block subjects, while the spectrogram time-frequency analysis is used to give better characterization of ECG parameters. These analyses can be used to construct ECG monitoring and analyzing system for heart blocks detection, As in figure 4-5 is shown.
Figure 4-5 shows the per-processing simulation results for normal ECG and third degree heart block subject. For the normal subject, the power spectrum shows that the signal frequency is 8 Hz, while the spectrogram shows that the signal frequency lies at all times during the observation interval. For the third degree heart block subject, the power spectrum shows that the signal frequency is 3.9 Hz, while the spectrogram represents that the signal frequency is only appears within the duration of 500 ms periodically for every 2000 ms. It is not shown on the power spectrum representations.



Time domain

Periodogram power spectrum

Spectrogram time-frequency

Figure 4-5: a) ECG normal signal, b) Periodogram power spectrum, c) Spectrogram time-frequency.

4.6 Smoothed Spectral Estimation via Cepstrum Thresholding

We use varies techniques for analyze data and selecting feature. One of those technique is cepstrum thresholding, named SThresh, which is shown to be an effective, yet simple, way of obtaining a smoothed non-parametric spectrum estimate of a stationary signal.

Cepstrum thresholding is shown to be an effective way for obtaining a smoothed nonparametric estimate of the spectrum of an audio signal, such as heart sound. Introducing the cepstrum thresholding-based spectral estimator for non-stationary signal is of interest to researchers in spectral analysis and allied topics, such as audio signal processing.

The cepstrum of $y(t)$ can be defined as follows [33]:

$$c_k = \frac{1}{N} \sum_{l=0}^{N-1} \ln(\phi_l) e^{i\omega_l k} \qquad (4.6)$$

$$K = 0, 1, ..., \ N - 1$$

Let us consider a stationary, real valued signal, real valued signal $\{y(t)_{t=0}^{t=N-1}\}$ its periodogram estimate $\hat{\phi}_p$ for $p = 0, 1, ..., \frac{N}{2}$ is given by:

$$\hat{\phi}_p(\omega) = \frac{1}{N} \left| \sum_{t=0}^{N-1} y(t) e^{-j2\pi f t} \right|^2 \qquad (4.7)$$

where it is assumed that $\phi > 0, \forall p$ The cepstral coefficients have several interesting features, one of which is mirror symmetry, defined as:

$$c_{N-k} = C_k \qquad (4.8)$$

$$k = 0, 1, ..., \frac{N}{2}$$

In other words, only half of the sequence, $c_0, c_1, ..., c_{(N/2)-1}$ is distinct and the other half is obtained from eq 4.8.

In $c1, ..., c_{(N/2)-1}$ using the periodogram estimate in eq 4.7, a common estimate of the cepstral coefficients is obtained by replacing $\phi(\omega)$ in $\hat{\phi}(\omega)$ in the next equation, which gives [63]:

$$\hat{c}_k = \frac{1}{N}\sum_{l=0}^{N-1} \ln[\hat{\phi}(\omega_l)]e^{j\omega_l k} + \gamma\delta_{k,0} \qquad (4.9)$$

$$k = 0, ..., \frac{N}{2}$$

where

$$\delta_{k,0} = \begin{cases} 1 & \text{if } k = 1 \\ 0 & \text{otherwise} \end{cases}$$

and $\gamma = 0.577$ (Euler's constant) It can be shown (see, e.g.[63]) that with large samples the estimated cepstral coefficients $\{\hat{c}_k\}_{k=0}^{N/2}$ are independent normally distributed random variables. In symbols $\hat{c}_k \cong N(C_k, S_K^2)$ with:

$$S_k^2 = \begin{cases} \frac{\pi^2}{3N} & \text{if } k = 0, \frac{N}{2} \\ \frac{\pi^2}{6N} & \text{if } k = 1, ..., \frac{N}{2} - 1 \end{cases}$$

Keeping in mind the above equations, the idea behind cepstrum thresholding is straightforward. Let $\hat{c}_k$ be a new estimate of C_k and note that $\hat{c}_k = 0$ has a mean squared error (MSE) equal to C_k^2. This estimate is preferred to $\hat{c}_k$ as long as $C_k^2 \leq S_k^2$ as now let:

$$S = \{K \in \left[0, \frac{N}{2}\right] \ |c_k^2 \leq s_k^2\} \qquad (4.10)$$

And let S be an estimate of S . Thresholding $\{\hat{c}_k\}_{k \in S}$ gives the following new estimate of c_k:

$$\tilde{c}_k = \begin{cases} 0 & \text{if } k \in \tilde{S} \\ \hat{c}_k & \text{Otherwise} \end{cases}$$

$k = 0, ..., \frac{N}{2}$ A good estimate of S is given by (see [64] for details):

$$\tilde{S} = \{K \in \left[0, \frac{N}{2}\right] \mid |c_k^2| \leq \mu s_k\} \tag{4.11}$$

Where the parameter controls the risk of concluding that $|c_k^2|$ is significant" while this is not true, the so called "false alarm probability". The following values of μ are recommended in [63, 33] for $N \in (128, 2048)$ and $\mu = \mu_0 + \frac{N-12}{1920}$. For sample lengths $N < 500$, which are most commonly encountered in applications, we recommend $\mu_0 = 2 \; and \; 4$ for narrow band and broadband signals respectively, whereas for $N \geq 500$ we suggest $\mu_0 = 3 \; and \; 5$ for narrow band and broadband signals respectively.
This implies that μ will belong to the interval. $(\mu_0, \mu + 1)$ for other intervals of the sample length, N. Similar rules can be given. The smoothed spectral estimate corresponding to $\{\hat{C}_k\}$ is given by:

$$\phi_{cep}(\omega_l) = exp\left[\sum_{k=0}^{N-1} \hat{c} e^{-j\omega_l k}\right] \qquad l = 0, ..., N-1 \tag{4.12}$$

where the subscript cep signifies its cepstrum dependence. The final scaled spectrum estimate $\hat{\phi_{cep}}(\omega)$ is then given by $\hat{\phi}_{cep}(\omega_l) = \hat{\alpha}\tilde{\phi}_{cep}(\omega_l)$ l=0,...,N-1.

The proposed nonparametric spectral estimate is obtained by a simple scaling

$$\hat{\alpha} = \frac{\sum_{l=0}^{N-1} \hat{\phi}_{cep}(\omega_l)\tilde{\phi}_{cep}(\omega_l)}{\sum_{l=o}^{N-1} \tilde{\phi}^2_{cep}(\omega_l)} \tag{4.13}$$

In et al, Prabhu Babu [11] proposed a fully automatic method for variance reduction of spectrum estimates which used the technique of cepstrum thresholding, named SThresh. The method is shown to be an effective, yet simple, way of obtaining a smoothed non-parametric spectrum estimate of a stationary signal. The study obtained the threshold via a cross-validation scheme and the results are shown to be in agreement with those obtained when the spectrum is fully known. Smoothed nonparametric spectral estimation via cepstrum thresholding is shown in Figure 4-6.

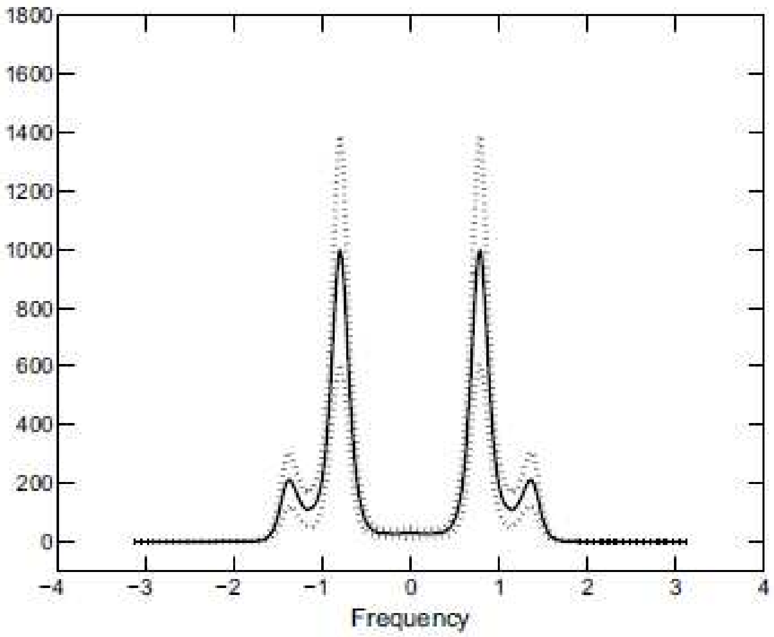


Figure 4-6: Smoothed spectrum of simulated narrowband ARMA signal (N=512).

The above figure 4-6 depict the results related to spectrum estimate via cepstrum thresholding, whereas the one reported below illustrate the result.
Amplitude changes are observed. This method is nonparametric and capable of producing smoother and better cepstrum estimates without imposing any parametric model. In figure ??, we clearly see that the variance of the smoothed spectrum is significantly smaller than that of the peridogram.

4.7 Bispectrum

A potential tool for future feature extraction or appear heart sound's components may be the estimation of the bispectrum. The power spectrum is based on the second order statistics of the time series, but the bispectrum make use of third order statistics. By definition, a gaussian random process has a zero higher-order spectrum of order two (bispectrum) (Nikias & Petropulu 1993), which allows the study of the deviation from the gaussianity or to suppress gaussian noise. The method also contains information about the phase character of the signal, which is failed with the methods based on the second order statistics. Moreover, the bispectrum estimation can be used in detection and characterization of the nonlinearities by analysis of quadratic phase coupling in the frequency domain. A preliminary study on quadratic phase locking in HRV can be found in Calcagnini et al. (1996).

The third-order spectrum, called bispectrum, is a particular example of higher-order spectrum (HOS), which is defined as the Fourier transform of third-order cumulant sequence. The power spectrum is member of the class of higher-order spectra. HOS are the extension to higher orders of the concept of the power spectrum. In cases where the process is non-Gaussian or is generated by nonlinear mechanisms, HOS provide information which can not be obtained from the conventional spectrum.

For the sake of completeness, let us recall that second order statistics such as autocorrelation and power spectrum provide important information in analysis of Gaussian, stationary and linear processes.

$$m_x = E(x) \tag{4.14}$$

$$m_x^2(i) = E\left\{X(n)\, X(n+i)\right\} \tag{4.15}$$

Higher order statistics, used in the analysis of Gaussian, stationary and non-linear processes, typically allow to obtain important results.

Higher order statistics are calculated upon higher order moments (HOM) such as m3 and m4, i.e., third and fourth order moment, defined as follows:

$$m_x^3(i,j) = E\left\{X(n)\, X(n+i)\, X(n+j)\right\} \tag{4.16}$$

It is worth noting that moments give more accurate results in the analysis of deterministic signals, while cumulants give more accurate results in the analysis of random signals[2]. Power spectrum of random signals are defined by DFT, i.e., Discrete Fourier transform, (see also Equation 4):

$$B^x = (f_1, f_2) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} C_3^x(m,n) e^{-j2\pi(mf_1+nf_2)} \tag{4.17}$$

In the event that the signal is a stationary random process with real values, we can write:

$$B(w_1, w_2) = X(w_1)\ X(w_2)\ X^*(w_1, w_2) \tag{4.18}$$

A diagonal slice of a single variable bispectrum for a special situation at which frequencies are equal can be defined as follows:

$$B(w) = X(w)\ X^*(2w) \tag{4.19}$$

As bispectrum analysis is not easy to calculate, this slice of spectrum obtained from a bispectrum is used for giving an idea in the analysis of data that do not exhibit nonlinear or Gaussian distribution in the signal. In et al AimÃľ Lay-Ekuakille [44] presented an original implementation of EEG signal processing using filter diagonalization method to build a bispectrum and contour representation to discover possible abnormalities hidden in the signal for aided-diagnosis. The detection of neurophysiological features by means of electroencephalogram (EEG) is one of the most recurrent medical exams to be performed on human beings. Two different electroencephalogram EEG signals are used for this scope. EEG signals are acquired simultaneously with electrocardiograms (ECG) and ergospirometric ones. ECG signals are also processed along with EEGs. A comparison is made with high order spectra approach. All experimental data regarding EEG, ECG, and ergospirometry are acquired during suspected-patient walking along a path of 32 for verifying the impact of fatigue on neurophysiological processes and vice versa figure 4-7.

According to this study, EEG signals for the envisaged intervals show what happens; that is also correlated to the distribution of peaks on magnitude spectrum, and absorption spectra figures. This distribution allows to understand the oxygen uptake necessary for brain functioning. However HOS does not clearly point out the issue related to neurophysiological aspects.

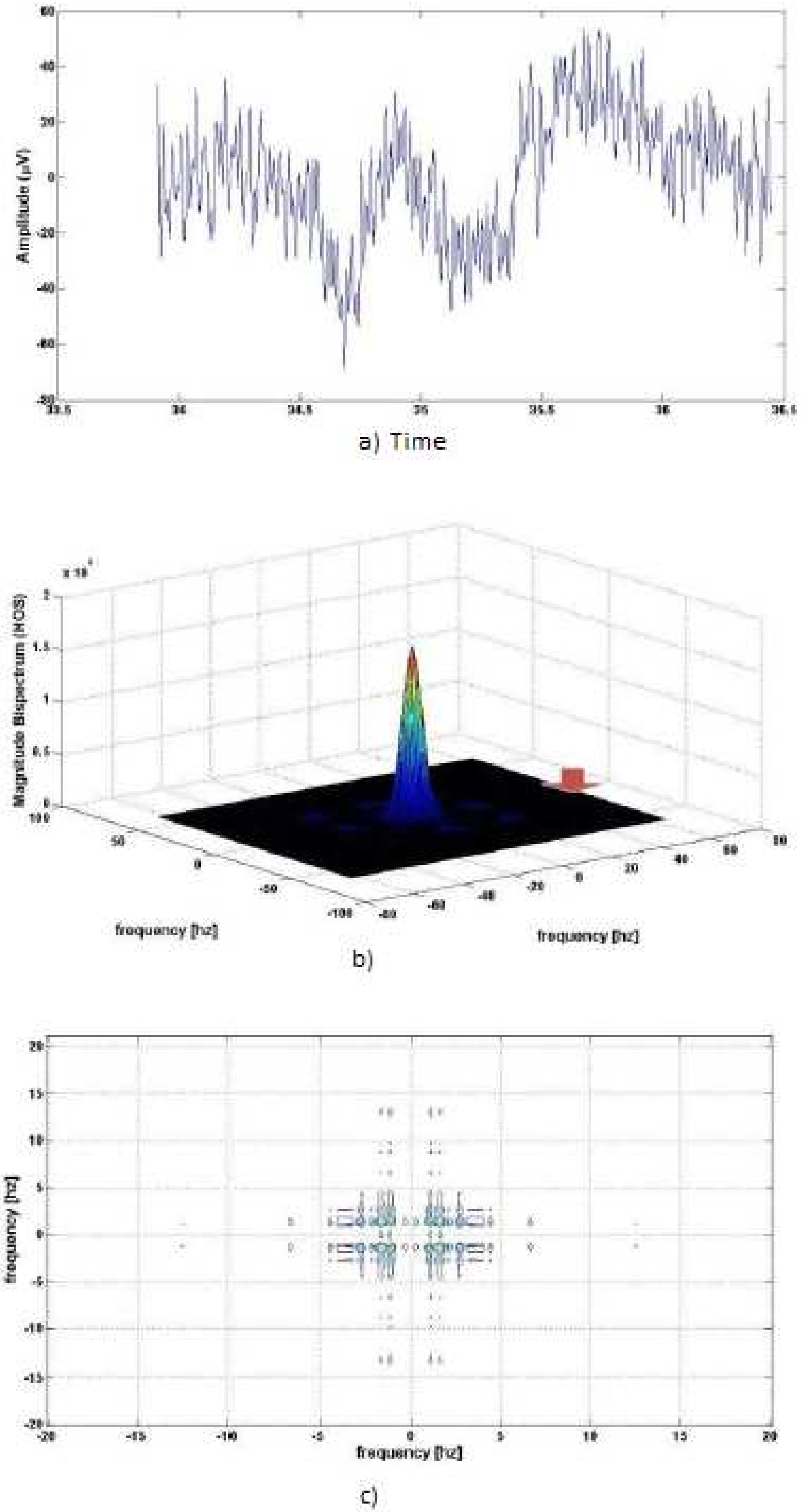


Figure 4-7: a) EEG signal of a patient, b) magnitude bispectrum with HOS, c) contour plot of bispectrum.

4.8 Wigner Distribution

The Wigner Distribution (WD) was introduced in 1932 by Wigner in the context of quantum mechanics, it's usefulness to problems in communication theory was discovered by Ville in 1948, consequently, it is often called the Wigner-Ville distribution. In order to differentiate the second-order WD from the higher order WDs, we will refer to the conventional WD as the Wigner Spectrum (WS) [21].

However Wigner Distribution (WD) and corresponding Wigner Ville Distribution (WVD) can analyses the non-stationary signal properly. This ability comes from the fact that WD can separate the signal in both time and frequency directions. The advantage of WD over STFT is that it has no time frequency trade-off problem, but its disadvantage is that in its response it has a cross-term. Nonlinear behavior of the WD is the main cause of the cross-term. To remove the cross-term it is necessary to smooth the time frequency plane but it decreases the time frequency resolution.

Given a signal $x(t)$, the corresponding Wigner distribution is defined by [60]

$$W(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} x^*(t-\frac{1}{2}\tau)x(t+\frac{1}{2}\tau)e^{-j\tau\omega}d\tau \tag{4.20}$$

or, given the associated spectrum $X(\omega)$ of the signal $x(t)$,

$$W(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X^*(\omega+\frac{1}{2}\theta)X(\omega-\frac{1}{2}\theta)e^{-jt\theta}d\theta \tag{4.21}$$

The two definitions can be easily proven to be equivalent by substituting $x(t)$ with its expression in terms of the spectrum. Figure 4-8 is a contour plot of Wigner Distribution from a sinus signal.

4.9 Wigner Bispectrum

Time-frequency distribution are transform transformations that attempt to describe how the spectral content of a signal is changing with time. They are known as distributions because somehow they describe the energy or intensity of a signal in time

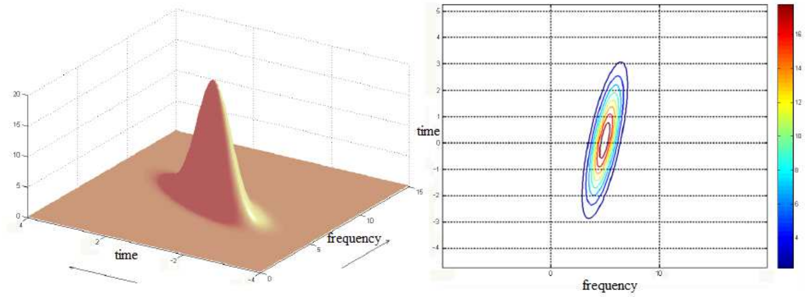


Figure 4-8: Wigner distribution of the above signal, represented as a surface and as contour curves.

and in frequency simultaneously. Nevertheless, they are not distributions in a probabilistic sense since positivity can not usually be ensured. An infinite number of time-frequency distributions can be generated from Cohen's general class formulation [18]. Special cases of this general class include the spectrogram, Rihaczek, Page, Wigner-Ville and Choi-Williams distributions [18].

Cohen's class of distributions are bilinear expressions, i.e., they are based on the second-order moments of the signal. A definition of a general class of time-frequency distributions in terms of higher order moments could contribute to the understanding of time-varying higher order moment spectra (HOMS)[55] in the same way that Cohen's general class does for the time-varying second-order spectra. The definition of a general class requires, however, the formulation of the basic representation, i.e., the representation for which the kernel is equal to unity.

In analogy with Cohen's general class, the Wigner higher order moment spectra (WHOS) are chosen as the basic representation. For every time instant t, the WHOS expresses the varying HOMS in the same way that the WD does for the instantaneous power spectrum. The third-order Wigner distribution was originally introduced by Gerr [31]. This definition has been carefully conceived to preserve the properties of the WD. In particular, the properties related to the instantaneous power and spectral density function in the WD are now related to the instantaneous $(k + 1)$ th-order moment and $(k + 1)$ th-order HOMS. The properties of this higher order moment spectra derivation can differ substantially from a derivation based on higher order cumulant spectra. The study Javier et. al 1993 has proved that under low SNR circumstances, the Wigner Bispectrum is better than the Wigner-Ville distribution [21]. The High-Order Spectra of Wigner- Ville Distribution of signal $x(t)$ is defined as follows [29]:

$$
\begin{aligned}
W(t, f_1, f_2, \cdots, f_k) = \\
& \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x^*(t - \frac{1}{k+1} \sum_{m=1}^{k} \tau_m). \\
& \prod_{i=1}^{k} x(t + \frac{k}{k+1}\tau_i - \frac{1}{k+1} \sum_{j=1, j\neq 1}^{k} \tau_i) \\
& exp(-2j\pi f_i \tau_i) d\tau_i.
\end{aligned}
\tag{4.22}
$$

Where $W(t, f_1, f_2, f_k)$ represents the kth order Fourier transform of a k-dimensional local function. Let us define R_{kt} as follows:

$$
R_{kt}(\tau_1, \tau_2, \ldots \tau_k) = x^*(t - \alpha) \prod_{i=1}^{k} x(t + \tau_i - \alpha) \tag{4.23}
$$

Where α, is the delay of time. Note that $R_t(\tau_1, \tau_2, \ldots \tau_k)$ is defined such that one of the factors in the product is a delayed version of the conjugate of $x(t)$ and the rest are delayed versions of $x(t)$.

To fulfill the three basic properties of time-frequency distributions, in a higher order moment spectrum domain, the value of α should be chosen properly. In particular, to attain the instantaneous frequency as the mean frequency in the multi frequency space at a given time, it will be shown that $R_t(\tau_1, \tau_2, \ldots \tau_k)$ should be centered at time instant t, in such a way that [61]

$$
\frac{1}{k+1}\left((t - \alpha) + \sum_{i=1}^{k}(t + \tau_i - \alpha)\right) = t \tag{4.24}
$$

Consequently,

$$
\alpha = \frac{1}{k+1} \sum_{i=1}^{k} \tau_i \tag{4.25}
$$

and

$$\begin{aligned}R_{kt}(\tau_1, \tau_2, \ldots \tau_k) = & \\ & x^*\left(t - \frac{1}{k+1}\sum_{m=1}^{k}\tau_m\right) \\ & \cdot\prod_{i=1}^{k} x\left(t + \frac{k}{k+1}\tau_i - \frac{1}{k+1}\sum_{j=1, j\neq}^{k}\tau_j\right)\end{aligned} \tag{4.26}$$

This leads to definition 4.22. Special cases of WHOS include the Wigner bispectrum (WB) for $k = 2$,

$$\begin{aligned}W_{2x}(t, f_1, f_2) = & \\ & \int_{\tau_1}\int_{\tau_2} x^*\big(t - \frac{1}{3}\tau_1 - \frac{1}{3}\tau_2\big) \\ & x\big(t + \frac{2}{3}\tau_1 - \frac{1}{3}\tau_2\big) \\ & x\big(t + \frac{2}{3}\tau_2 - \frac{1}{3}\tau_1\big) exp(-j2\pi f_1\tau_1) \\ & exp(-j2\pi f_2\tau_2) d\tau_1 d\tau_2.\end{aligned} \tag{4.27}$$

and the Wigner trispectrum (WT) for $k = 3$,

$$\begin{aligned}W_{3x}(t, f_1, f_2, f_3) = & \\ & \int_{\tau_1}\int_{\tau_2}\int_{\tau_3} x^*\big(t - \frac{1}{4}\tau_1 - \frac{1}{4}\tau_2 - \frac{1}{4}\tau_3\big) \\ & x\ \big(t + \frac{3}{4}\tau_1 - \frac{1}{4}\tau_2 - \frac{1}{4}\tau_3\big) \\ & x\ \big(t + \frac{3}{4}\tau_2 - \frac{1}{4}\tau_1 - \frac{1}{4}\tau_3\big) \\ & x\ \big(t + \frac{3}{4}\tau_3 - \frac{1}{4}\tau_1 - \frac{1}{4}\tau_2\big) \\ & exp(-j2\pi f_1\tau_1)\ exp(-j2\pi f_2\tau_2) \\ & exp(-j2\pi f_3\tau_3) d\tau_1 d\tau_2 d\tau_3\end{aligned} \tag{4.28}$$

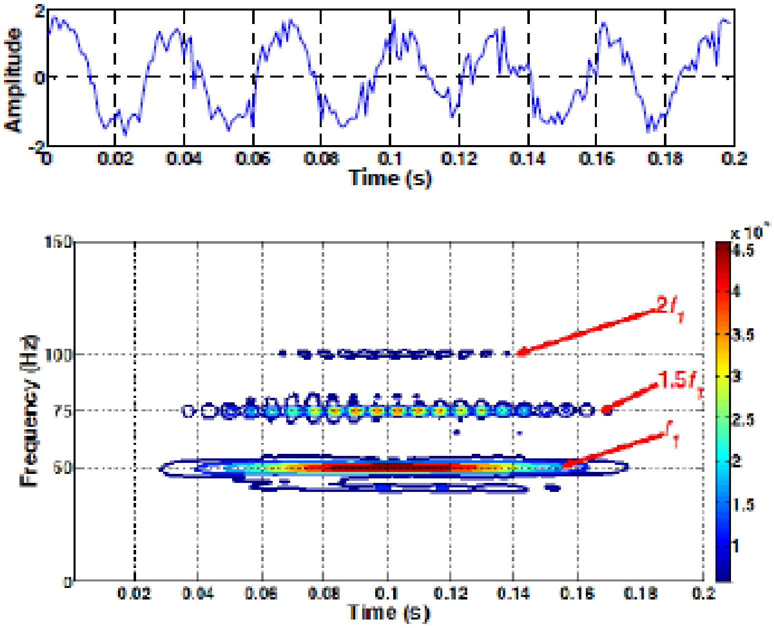


Figure 4-9: A signal sample in the time frequency and Wigner Bispectrum.

Observe that for $k = 1$ the WD follows from 4.22:

$$W_{1x}(t,f) = \int_{\tau} x^{*}(t - \frac{1}{2}\tau)\, x(t + \frac{1}{2}\tau)\, exp(-j2\pi f\tau) d_{\tau}. \tag{4.29}$$

The definition of WB of 4.22 differs from the third order Wigner distribution proposed by Gerr in [31] only in the dependence of the conjugate of the signal. Et al Zhixiong Li [45] describes and evaluates the development and application of an intelligent diagnostic technique based on the integration of the empirical mode decomposition, kernel independent component analysis, Wigner bispectrum and support vector machine. It is work on the fault detection for a diesel engine using the instantaneous angular speed. In this study, in order to solve the undetermined blind source separation (BSS) problem the combination of EMD and KICA is firstly presented to estimate IAS signals from a single-channel IAS sensor. The KICA is also applied to select distinguished features extracted by Wigner bispectrum which the Wigner bispectrum analysis is employed to extract sensitive amplitude and phasic features and assess the state of the machine figure 4-9.

4.10 Wavelet Transform

As we discussed, a Fourier transform based spectral analysis is the dominant analytical tool for frequency domain analysis. However, Fourier transform cannot provide any information of the spectrum changes with respect to time. Fourier transform assumes the signal is stationary, but PD signal is always non-stationary. To overcome this deficiency, a modified method-short time Fourier transform allows to represent the signal in both time and frequency domain through time windowing function. The window length determines a constant time and frequency resolution. Thus, a shorter time windowing is used in order to capture the transient behavior of a signal; we sacrifice the frequency resolution. an alternative mathematical tool- wavelet transform must be selected to extract the relevant 36 time-amplitude information from a signal. In the meantime, we can improve the signal to noise ratio based on prior knowledge of the signal characteristics.

In this work, we stated only some keys equations and concepts of wavelet transform, more rigorous mathematical treatment of this subject can be found in [16, 20, 38]. A continuous-time wavelet transform of f(t) is defined as:

$$CWT_{\Psi}\, f(a,b) = W_f(a,b) = |a|^{\frac{1}{2}} \int_{-\lim}^{\lim} f(t)\Psi^*(\frac{t-b}{a})dt \tag{4.30}$$

Here $a, b \in R, a \neq 0$ and they are dilating and translating coefficients, respectively. The asterisk denotes a complex conjugate. This multiplication of $|a|^{\frac{1}{2}}$ is for energy normalization purposes so that the transformed signal will have the same energy at every scale. The function $\psi(t)$, the so-called mother wavelet, is scaled by a, so a wavelet analysis is often called a time-scale analysis rather than a time-frequency analysis. The wavelet transform decomposes the signal into different scales with different levels of resolution by dilating a single prototype function, the mother wavelet. Furthermore, a mother wavelet has to satisfy that it has a zero net area, which suggest that the transformation kernel of the wavelet transform is a compactly support function (localized in time), thereby offering the potential to capture the PD spikes which normally occur in a short period of time [54]. Daubechies wavelets has used for our application, so let us talk few about Daubechies wavelets.

Based on these equation, Daubechies [19], designed a type of wavelet for a given vanishing moment p and find the minimum size discrete filter. The conclusion is that if we want the wavelet function with p vanishing moments, the minimum filter size is $2p$. The derivation define as:

$$H_\phi(e^{j\omega}) = \sqrt{2}(\frac{1+e^{-j\omega}}{2})^p R(e^{j\omega}) \tag{4.31}$$

where $H_\phi(\omega)$ is the discrete-time Fourier transform of the discrete filters and the absolute-square of this function is

$$\begin{aligned} \left|H_\phi(e^{j\omega})\right|^2 &= H_\phi(e^{j\omega})H_\phi^*(e^{j\omega}) \qquad (4.32) \\ &= 2(\frac{1+e^{-j\omega}}{2}\frac{1+e^{j\omega}}{2})^p R(e^{j\omega})R^*(e^{j\omega}) \\ &= 2(\frac{2+e^{-j\omega}+e^{j\omega}}{4})^2 p \left|R(e^{j\omega})\right|^2 \\ &= 2(cos\frac{\omega}{2})^{2p} P(sin^2\frac{\omega}{2}). \end{aligned}$$

$$\tag{4.33}$$

The last step makes $P(sin^2\frac{\omega}{2}) = \left|R(e^{j\omega})\right|^2$. we can determine the form of $P(x)$. Let $y = sin^2\frac{\omega}{2}$. We have

$$(1-y)^p P(y) + y^p P(1-y) = 1. \tag{4.34}$$

A theorem in algebra, called Bezout theorem, can solve this equation. The unique solution is

$$P(y) = \sum_{k=0}^{p-1} \begin{pmatrix} p-1+k \\ k \end{pmatrix} y^k. \tag{4.35}$$

The polynomial $P(y)$ is the minimum degree polynomial satisfying equation 4.34. Once we have $P(y)$, the polynomial $R(e^{j\omega})$ can be derived. First we decompose

$R(e^{j\omega})$ according to its roots.

$$R(e^{j\omega}) = \sum_{k=0}^{m} r_k e^{-jk\omega} = r_0 \prod_{k=0}^{m} (1 - a_k e^{-j\omega}) \tag{4.36}$$

Let $z = e^{j\omega}$, the relation between P and R is

$$P(\frac{2 - z - z^{-1}}{4}) = r_0^2 \prod_{k=0}^{m} (1 - a_k z^{-1})(1 - a_k z) \tag{4.37}$$

By solving the roots of $P(\frac{2-z-z^{-1}}{4}) = 0$, we have the roots of R, $\{a_k,\ 1/a_k\}_{k=0,1,\ldots,m}$ and $r_0 = 2^{p-1}$. Usually, we choose a_k lies in the unit circle to have minimum phase filter. Taking $p = 2$ for a example. The obtained polynomial $P(y)$ is

$$P(y) = \sum_{k=0}^{1} \begin{pmatrix} 1+k \\ k \end{pmatrix} y^k = 1 + 2y. \tag{4.38}$$

$$P(\frac{2 - z - z^{-1}}{4}) = 2 - \frac{1}{2}z - \frac{1}{2}z^{-1} \tag{4.39}$$

The roots are $2 + \sqrt{3}\ and\ 2 - \sqrt{3}$. After factorization, we have the low pass filter to be

$$H_\phi(e^{j\omega}) = \frac{\sqrt{2}+\sqrt{6}}{8}[n] + \frac{3\sqrt{2}+\sqrt{6}}{8}e^{-j\omega} + \frac{3\sqrt{2}-\sqrt{6}}{8}e^{-j2\omega} + \frac{3\sqrt{2}-\sqrt{6}}{8}e^{-j3\omega} \tag{4.40}$$

The discrete-time domain representation is

$$h_\phi[n] = \frac{\sqrt{2}+\sqrt{6}}{8}\delta[n] + \frac{3\sqrt{2}+\sqrt{6}}{8}\delta[-n] + \frac{3\sqrt{2}-\sqrt{6}}{8}\delta[n-2] + \frac{3\sqrt{2}-\sqrt{6}}{8}\delta[n-3] \tag{4.41}$$

The result is the minimum size filter with 2 vanishing moments and the corresponding filter size is 4. Recall the conclusion mentioned above, the filter size is two times the vanishing moment. Higher order Daubechies wavelets are derived at similar way. The coefficient and the plot of heart sound are shown in figure 4-10 and figure 4-11 as normal and pathological heart sounds respectively.

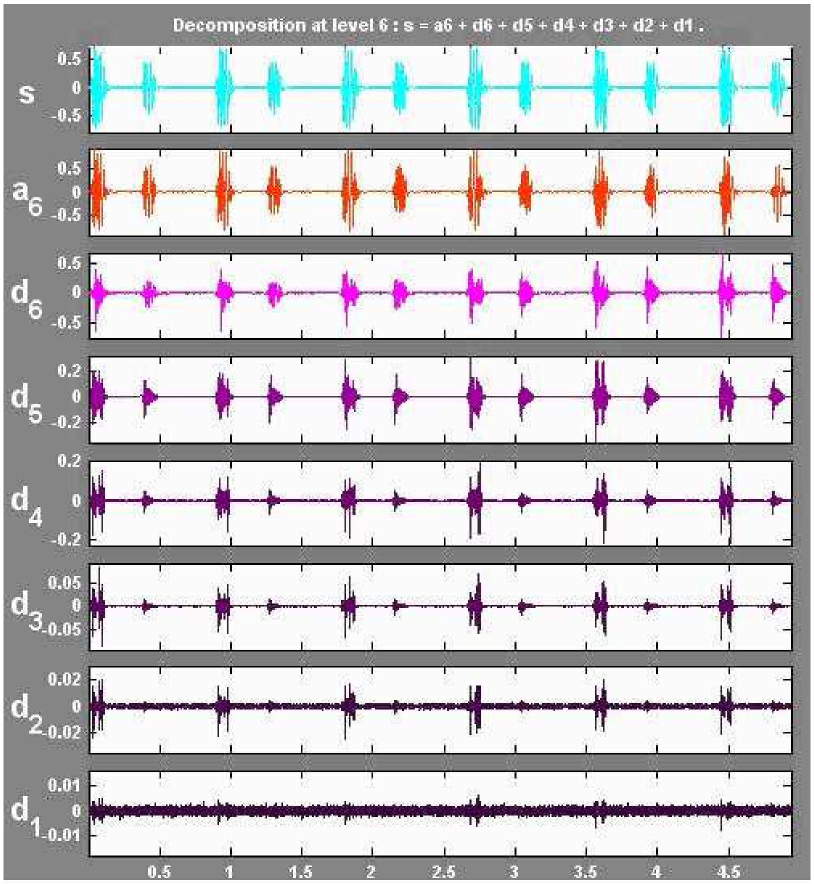


Figure 4-10: Daubechies wavelet transform of normal heart sound.

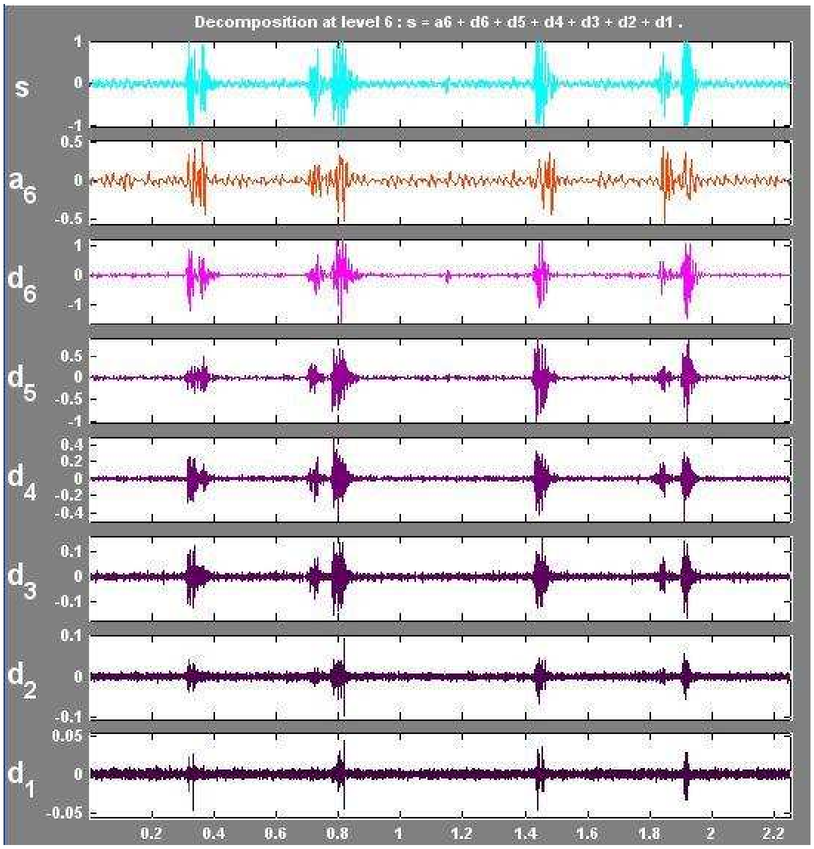


Figure 4-11: Daubechies wavelet transform of pathological heart sound.

5

Data Mining and Classification Tools

There are two forms of data analysis that can be used for extract models describing important classes or predict future data trends. These two forms are as follows: i) Classification and ii) Predication.

These data analysis help us to provide a better understanding of large data. Classification predicts categorical and prediction models predicts continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods such as neural networks or decision trees. Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction. The objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data. Finding useful patterns in data is known by different names (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing) [68].

The term "data mining" is primarily used by statisticians, database researchers, and the business communities. The term KDD (Knowledge Discovery in Databases) refers to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process [28, 32]. The steps in the KDD process, such as data preparation, data selection, data cleaning, and proper interpretation of the results of the data mining process, ensure that useful knowledge is derived from the data. Data mining is an extension of traditional data analysis and statistical approaches as it incorporates analytical techniques drawn from various disciplines like AI, machine learning, OLAP, data visualization, etc.

There are problem categories that cannot be formulated as an algorithm. Problems that depend on many subtle factors, for example the purchase price of a real estate which our brain can (approximately) calculate. Without an algorithm a computer cannot do the same. Therefore the question to be asked is: How do we learn to explore such problems? Artificial Neural Network (ANN) is one of tools that can solve it [39].

5.1 Artificial Neural Networks

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurones) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurones. This is true of ANNs as well [62].

Artificial neural networks are inspired by attempts to simulate biological neural systems. The human brain consists primarily of nerve cells called neurons, linked together with other neurons via stand of fiber called axons. Axons are used to transmit

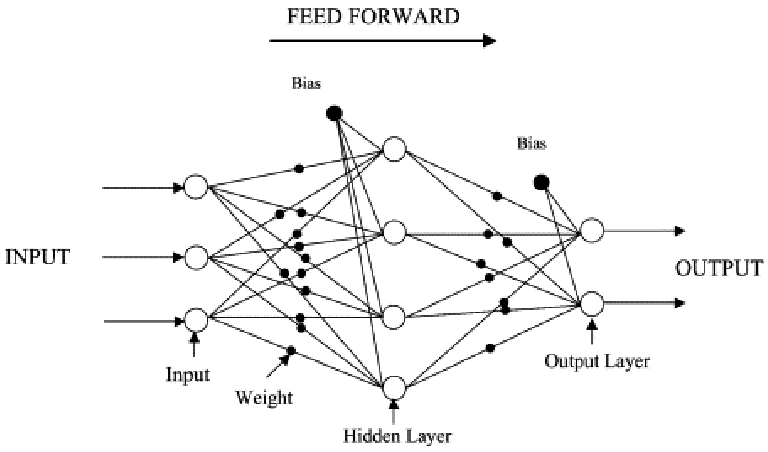


Figure 5-1: Feed-forward Neural Network Model.

nerve impulses from one neuron to another whenever the neurons are stimulated. A neuron is connected to the axons of other neurons via dendrites, which are extensions from the cell body of the neurons. The contact point between a dendrite and an axon is called a synapse [62].

Multilayer perceptron (MLP) is a feed-forward neural networks trained with the standard back-propagation algorithm. Figure 5-1 illustrates architecture of a simple feed forward neural network. As training a MLP is a supervised task, MLP requires a desired response to be trained. They learn how to transform input data in to a desired response, so they are widely used for pattern classification. With one or two hidden layers, they can approximate virtually any input-output map. It has been shown to approximate the performance of optimal statistical classifiers in difficult problems. The MLP is trained with error correction learning, which is appropriate here because the desired response is the arteriography result and as such known.

Using artificial neural networks it is impossible to model the full complexity of the brain of anything other than the most basic living creatures, and generally ANNs will consist of at most a few hundred (or few thousand) neurones, and very limited connections between them, quite small neural networks have been used to solve what have been quite difficult computational problems, ANNs are basic input and output

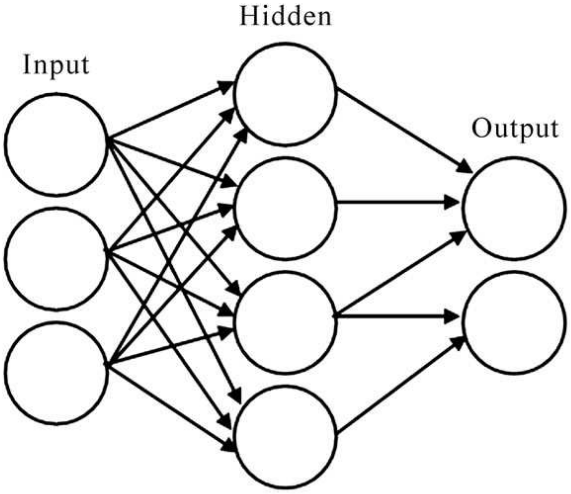


Figure 5-2: Simple Perceptron Architecture.

devices, with the neurones organized into layers. Simple Perceptrons consist of a layer of input neurones, coupled with a layer of output neuron, and a single layer of weights between them, as shown in Figure 5-2.

The learning process consists of finding the correct values for the weights between the input and output layer. The schematic representation given in Figure 5-2 is often how neural nets are depicted in the literature, although mathematically it is useful to think of the input and output layers as vectors of values(I and O respectively), and the weights as a matrix. We define the weight matrix W_{io} as an i x o matrix, where i is the number of input nodes, and o is the number of output nodes. The network output is calculated as follows.

$$O = f(IW_{io}) \tag{5.1}$$

Generally data is presented at the input layer, the network then processes the input by multiplying it by the weight layer. The result of this multiplication is processed by the output layer nodes, using a function that determines whether or not the output node fires.

The process of finding the correct values for the weights is called the learning rule, and the process involves initializing the weight matrix to a set of random numbers between -1 and $+1$. Then as the network learns, these values are changed until it has been decided that the network has solved the problem. Finding the correct values for the weights is effected using a learning paradigm called supervised learning. Supervised learning is sometimes referred to as training. Data is used to train the network, this constitutes input data for which the correct output is known. Starting with random weights, an input pattern is presented to the network, it makes an initial guess as to what the correct output should be [48].

During the training phase, the difference between the guess made by the network and the correct value for the output is assessed, and the weights are changed in order to minimize the error. The error minimization technique is based on traditional gradient descent techniques. While this may sound frighteningly mathematical, the actual functions used in neural networks to make the corrections to the weights are chosen because of their simplicity, and the implementation of the algorithm is invariably uncomplicated.

The perceptron learning rule is comparatively straightforward. Starting with a matrix of random weights, we present a training pattern to the network, and calculate the network output. We determine an error function E:

$$E(O) = (T - O) \tag{5.2}$$

Where in this case T is the target output vector for a training input. In order to determine how the weights should change, this function has to minimized. What this means is find the point at which the function reaches its minimum value. The assumption we make about the error function is that if we were to plot all of its potential values into a graph, it would be shaped like a bowl, with sides sloping down

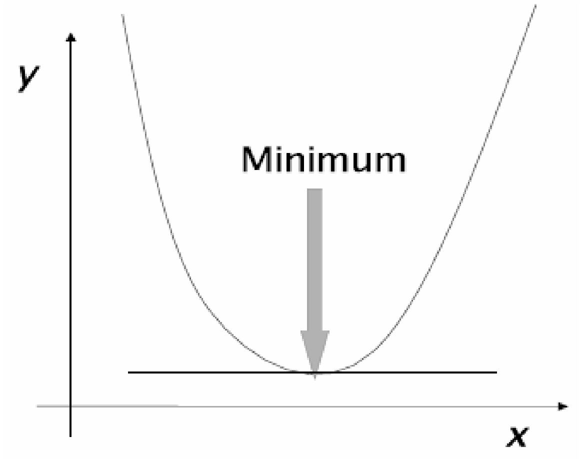


Figure 5-3: Function minimization using differentiation.

to a minimum value at the bottom [17].
In order to find the minimum values of a function differentiation is used. Differentiation is used to give the rate at which functions change, and is often defined as the tangent on a curve at a particular point 1. If our function is perfectly bowl shaped, then there will only be one point at which the minimum value of a function has a tangent of zero (i.e have a perfectly flat tangent), and that is at its minimum point (see Figure 5-3).

In neural network programming the intention is to assess the effect of the weights on the overall error function. We can take Equation 5.1 and combine it with Equation 5.2 to obtain the following.

$$E(O) = (T - O) = T - f(IW_{io}) \tag{5.3}$$

We then differentiate the error function with respect to the weight matrix. The discussion on Multilayer Perceptrons will look at the issues of function minimization in greater detail. Function minimization in the Simple Perceptron Algorithm is very straightforward. We consider the error each individual output node, and add that error to the weights feeding into that node. The perceptron learning algorithm works as follows.

1. Initialise the weights to random values on the interval $[1, -1]$.

2. Present an input pattern to the network.

3. Calculate the network output.

4. For each node n in the output layer...
   (a)calculate the error $E_n = T_n - On$
   (b) add E_n to all of the weights that connect to node n (add E_n to column n of the weight matrix.)

5. Repeat the process from 2. for the next pattern in the training set.

This is the essence of the perceptron algorithm. It can be shown that this technique minimises the error function. In its current form it will work, but the time taken to converge to a solution (i.e the time taken to and the minimum value) may be unpredictable because adding the error to the weight matrix is something of a "blunt instrument" and results in the weights gaining high values if several iterations are required to obtain a solution. This is akin to taking large steps around the bowl in order to and the minimum value, if smaller steps are taken we are more likely to and the bottom.

In order to control the convergence rate, and reduce the size of the steps being taken, a parameter called a learning r ate is used. This parameter is set to a value that is less than unity , and means that the weights are updated in smaller steps (using a fraction of the error). The weight update rule becomes the following.

$$W_{io}(t+1) = W_{io}(t) + \epsilon E_n \tag{5.4}$$

Which means that the weight value at iteration $t+1$ of the algorithm, is equivalent to a fraction of the error ϵE_n added to the weight value at iteration t [56].

5.2 Decision Trees

A decision tree (DT) is a classifier expressed as a recursive partition of the instance space. DT can handle high dimensional data. Their representation of acquired knowledge in tree free from is intuitive and generally easy to assimilate by humans.
A DT is a tree where root and each internal node are labeled with question. The arcs emanating from each node represent each possible answer to the associated question. Each leaf node represents a prediction of a solution to the problem under consideration. The constructions of DT classifier don't require any domain knowledge or parameter setting and therefore is appropriate for exploratory knowledge discovery. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a DT, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values.

In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attributes value. In the case of numeric attributes, the condition refers to a range.
Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Figure 5-4 describes a decision tree that reasons whether or not a potential customer will respond to a direct mailing. Internal nodes are represented as circles, whereas leaves are denoted as triangles. Note that this decision tree incorporates both nominal and numeric attributes. Given this classifier, the analyst can predict the response of a potential customer (by sorting it down the tree), and understand the behavioral characteristics of the entire potential customers population regarding direct mailing. Each node is labeled with the attribute it tests, and its branches are

labeled with its corresponding values.

In case of numeric attributes, decision trees can be geometrically interpreted as a collection of hyperplanes, each orthogonal to one of the axes. Naturally, decision-makers prefer less complex decision trees, since they may be considered more comprehensible. Furthermore, according to Breiman et al. (1984) the tree complexity has a crucial effect on its accuracy. The tree complexity is explicitly controlled by the stopping criteria used and the pruning method employed. Usually the tree complexity is measured by one of the following metrics: the total number of nodes, total number of leaves, tree depth and number of attributes used. DT induction is closely related to rule induction. Each path from the root of a DT to one of its leaves can be transformed into a rule simply by conjoining the tests along the path to form the antecedent part, and taking the leaf's class prediction as the class value [49, 43]. Decision tree inducers are algorithms that automatically construct a decision tree from a given dataset. Typically the goal is to find the optimal decision tree by minimizing the generalization error. However, other target functions can be also defined, for instance, minimizing the number of nodes or minimizing the average depth.

Depending on the outcome of the test, we go to either the left or the right sub-branch of the tree. Eventually we come to a leaf node, where we make a prediction. This prediction aggregates or averages all the training data points which reach that leaf. Figure 5-4 should help to clarify this. Why do this? Predictors like linear or polynomial regression are global models, where a single predictive formula is supposed to hold over the entire data space. When the data has lots of features which interact in complicated, nonlinear ways, assembling a single global model can be very difficult, and hopelessly confusing when you do succeed. Some of the non-parametric smoothers try to fit models locally and then paste them together, but again they can be hard to interpret. (Additive models are at least pretty easy to grasp.) An alternative approach to nonlinear regression is to sub-divide, or partition, the space into smaller regions, where the interactions are more manageable. We then partition the sub-divisions again. This is recursive partitioning, as in hierarchical clustering until finally we get to chunks of the space which are so tame that we can fit simple models

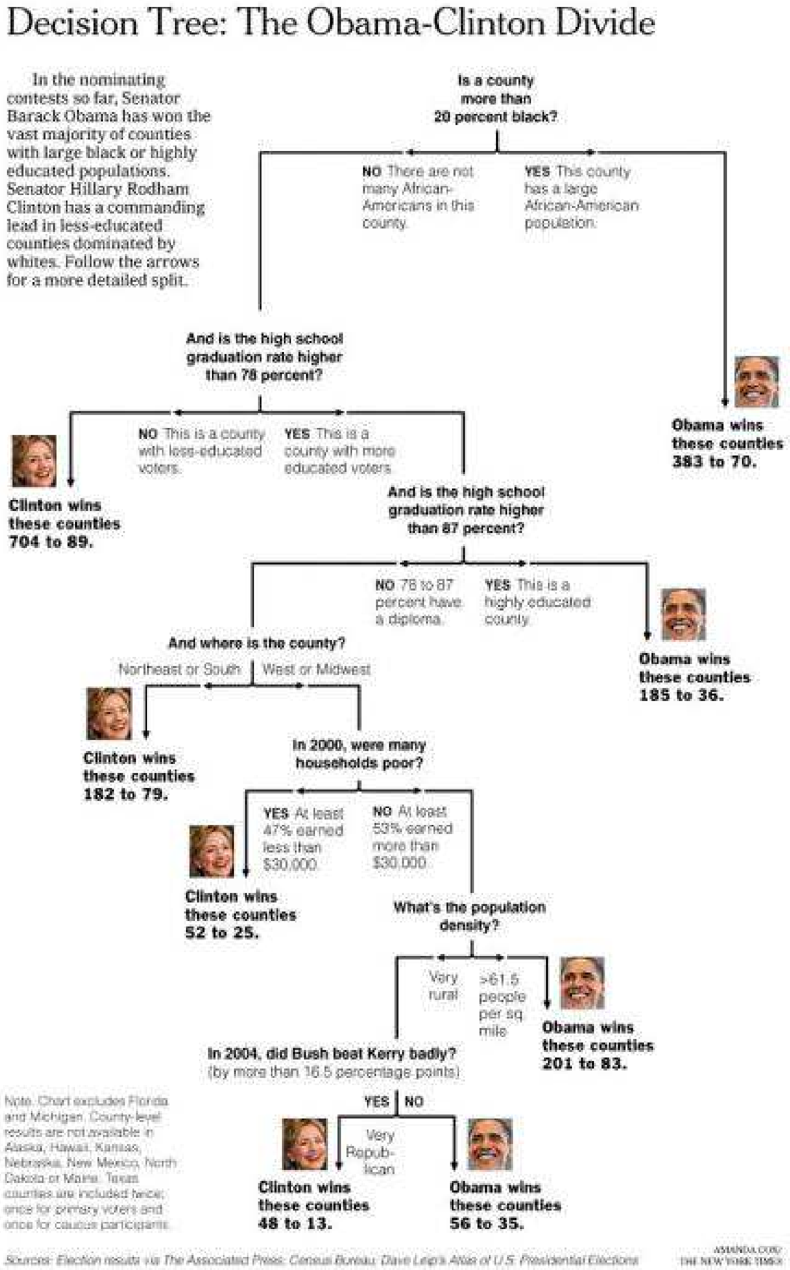

Decision Tree: The Obama-Clinton Divide
In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.
Is a county more than 20 percent black?
NO There are not many African-Americans in this county.
YES This county has a large African-American population.
Obama wins these counties 383 to 70.
And is the high school graduation rate higher than 78 percent?
NO This is a county with less-educated voters.
YES This is a county with more educated voters.
Clinton wins these counties 704 to 89.
And is the high school graduation rate higher than 87 percent?
NO 78 to 87 percent have a diploma.
YES This is a highly educated county.
Obama wins these counties 185 to 36.
And where is the county?
Northeast or South
West or Midwest
Clinton wins these counties 182 to 79.
In 2000, were many households poor?
YES At least 47% earned less than $30,000.
NO At least 53% earned more than $30,000.
Clinton wins these counties 52 to 25.
What's the population density?
Very rural
>61.5 people per sq. mile
Obama wins these counties 201 to 83.
In 2004, did Bush beat Kerry badly? (by more than 16.5 percentage points)
YES
NO
Very Repub-lican
Clinton wins these counties 48 to 13.
Obama wins these counties 56 to 35.
Note: Chart excludes Florida and Michigan. County-level results are not available in Alaska, Hawaii, Kansas, Nebraska, New Mexico, North Dakota or Maine. Texas counties are included twice; once for primary voters and once for caucus participants.
Sources: Election results via The Associated Press; Census Bureau; Dave Leip's Atlas of U.S. Presidential Elections
AMANDA COX/ THE NEW YORK TIMES

Figure 5-4: Classification tree for county-level outcomes in the 2008 Democratic Party primary (as of April 16), by Amanada Cox for the New York Times.

to them. The global model thus has two parts: one is just the recursive partition, the other is a simple model for each cell of the partition.

Now look back at Figure 5-4 and the description which came before it. DT use the tree to represent the recursive partition. Each of the terminal nodes, or leaves, of the tree represents a cell of the partition, and has attached to it a simple model which applies in that cell only. A point x belongs to a leaf if x falls in the corresponding cell of the partition. To finding which cell we are in, we start at the root node of the tree, and ask a sequence of questions about the features. The interior nodes are labeled with questions, and the edges or branches between them labeled by the answers. Which question we ask next depends on the answers to previous questions. In the classic version, each question refers to only a single attribute, and has a yes or no answer, e.g.,$HSGrad < 0.78$ or "Is Region $==$ Midwest?" The variables can be of any combination of types (continuous, discrete but ordered, categorical, etc.). You could do more than binary questions, but that can always be accommodated as a larger binary tree. Asking questions about multiple variables at once is, again, equivalent to asking multiple questions about single variables.

That's the recursive partition part; what about the simple local models? For classic regression trees, the model in each cell is just a constant estimate of Y . That is, suppose the points $(x_i, y_i), (x_2, y_2), ..., (x_c, y_c)$ are all the samples belonging to the leaf-node l. Then our model for l is just $\hat{y} = \frac{1}{c}\sum_{i=1}^{c} y_i$, the sample mean of the response variable in that cell. This is a piecewise-constant model. There are several advantages to this:

1. Making predictions is fast (no complicated calculations, just looking up constants in the tree).

2. It's easy to understand what variables are important in making the prediction (look at the tree).

3. If some data is missing, we might not be able to go all the way down the tree to a leaf, but we can still make a prediction by averaging all the leaves in the sub-tree we do reach.

4. The model gives a jagged response, so it can work when the true regression surface is not smooth. If it is smooth, though, the piecewise-constant surface can approximate it arbitrarily closely (with enough leaves).

5. There are fast, reliable algorithms to learn these trees.

A last analogy before we go into some of the mechanics. One of the most comprehensible non-parametric methods is k-nearest-neighbors: find the points which are most similar to you, and do what, on average, they do. There are two big drawbacks to it: first, you're defining "similar" entirely in terms of the inputs, not the response; second, k is constant everywhere, when some points just might have more very-similar neighbors than others. Trees get around both problems: leaves correspond to regions of the input space (a neighborhood), but one where the responses are similar, as well as the inputs being nearby; and their size can vary arbitrarily. Prediction trees are adaptive nearest-neighbor methods.
Decision tree classifiers are widely used for building classifier ensembles. Three important characteristics of these classifiers are:

1. If all the objects are distinguishable, that is, there are no identical elements of Z with different class labels, then we can build a tree classifier with zero re-substitution error. This fact places tree classifiers in the instable group: capable of memorizing the training data so that small alterations of the data might lead to a differently structured tree classifier. As we shall see later instability can be an advantage rather than a drawback when ensembles of classifiers are considered.

2. Tree classifiers are intuitive because the decision process can be traced as a sequence of simple decisions. Tree structures can capture a knowledge base in

a hierarchical arrangement, most pronounced examples of which are botany, zoology, and medical diagnosis.

3. Both quantitative and qualitative features are suitable for building decision tree classifiers. Binary features and features with a small number of categories are especially useful because the decision can be easily branched out. For quantitative features, a point of split has to be found to transform the feature into a categorical one. Hence, DT do not rely on a concept of distance in the feature space. As discussed earlier, a distance is not easy to formulate when the objects are described by categorical or mixed-type features. This is why decision trees are regarded as non metric methods for classification [25, 37].

Pruning is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. The dual goal of pruning is reduced complexity of the final classifier as well as better predictive accuracy by the reduction of over-fitting and removal of sections of a classifier that may be based on noisy or erroneous data. Sometimes early stopping can be too shortsighted and prevent further beneficial splits. This phenomenon is called the horizon effect [67]. To counter the horizon effect, we can grow the full tree and then prune it to a smaller size. The pruning seeks a balance between the increase of the training error and the decrease of the size of the tree. Downsizing the tree will hopefully reduce over-training. There are different criteria and methods to prune a tree summarized by Esposito et al.[27] as follows.

5.3 Classification and Regression Trees

The Classification and Regression Trees (CART) have been proposed by Breiman et al. in 1991. A CART is sophisticated program for fitting trees to data that chooses the split for each node such that maximum reduction in overall node impurity is achieved, where impurity is measured as the total sum of squared deviations from node centers. A novel method of choosing multi way partitions for classification and

DT was given by Biggs et al. in 1991 which chooses the best partition on the basis of statistical significance. Breiman, again in 1994, developed the bagging predictors which is a method of generating multiple versions of a predictor and using them to get an aggregated predictor. Later on, Loh and Shih (1997) developed the QUEST (Quick Unbiased Efficient Statistical Tree) method to take care of the selection bias towards the variables with more possible splits. To increase the statistical reliability of CART, Mola and Siciliano introduced statistical testing approach in the pruning procedure. Chou (1991) proposed an optimal partitioning method in classification tree for categorical explanatory variables with large number of categories based on a kmeans clustering procedure. Shih (2001) proposed methods of selecting the best categorical split in a tree based on a family of splitting criterion. These methods were shown to be useful to reduce the computational complexity of the exhaustive search methods. Cappeli et al.(2002) suggested the use of statistical significance in the pruning procedure of both classification and regression trees to obtain a statistically reliable tree.

Waheed et al.(2006) investigated the potential of hyper-spectral remote sensing data of experimental corn plots into categories of water stress, and nitrogen application rates for providing better crop management information in precision farming by using the CART algorithm. The results showed that the accuracy for the irrigation factor was 96% while that of the nitrogen application rate was 83%. Olden and Jackson (2002) provided a comparison between logistic regression analysis, linear discriminant analysis, classification trees and ANN to model fish species distributions and they concluded that classification trees and ANN greatly outperformed traditional approaches. Rothwell et al. (2008) applied the CART approach to evaluate the key environmental drivers controlling dissolved inorganic nitrogen (DIN) leaching from European forests which successfully classified the sites into the appropriate leaching category.

Using Classification And Regression Tree (CART) analysis is increasing in various applications. Early diagnosis of heart murmurs in newborns is a novel application of CART for clinical and physiological data. CART analysis is a tree-building technique which is different from traditional data analysis methods. In a number of studies, CART has been found to be quite effective for creating decision rules which perform as well or better than rules developed using more traditional methods. In addition, CART is often able to uncover complex interactions between predictors which may be difficult or impossible using traditional multivariate techniques. It is now possible to perform a CART analysis with a simple understanding of each of the multiple steps involved in its procedure. Classification tree methods such as CART are convenient way to produce a prediction rule from a set of observations described in terms of a vector of features and a response value. The aim is to define a general prediction rule which can be used to assign a response value to the cases solely on the bases of their predictor (explanatory) variables.

Tree-structured classification and regression are nonparametric computationally intensive methods that have greatly increased in popularity during the past dozen years. They can be applied to data sets having both a large number of cases and a large number of variables, and they are extremely resistant to outliers. Tree-structured classifications are not based on assumptions of normality and user-specified model statements, as are some conventional methods such as discriminant analysis and ordinary least square regression. Tree based classification and regression procedure have greatly increased in popularity during the recent years. Tree based decision methods are statistical systems that mine data to predict or classify future observations based on a set of decision rules and are sometimes called rule induction methods because the reasoning process behind them is clearly evident when browsing the trees. The CART methodology have found favor among researchers for application in several areas such as agriculture, medicine, forestry, natural resources management etc. as alternatives to the conventional approaches such as discriminant function method, multiple linear regression, logistic regression etc. In CART, the observations are successively separated into two subsets based on associated variables significantly related to the response variable; this approach has an advantage of providing easily comprehensible decision strategies. CART can be applied either as a classification tree or as a regressive tree depending on whether the response variable is categorical or continuous. Tree based methods are not based on any stringent assumptions. These methods can handle large number of variables, are resistant to outliers, non-parametric, more versatile, can handle categorical variables, though computationally more intensive. CART can be a good choice for the analysts as they give fairly accurate results quickly, than traditional methods. If more conventional methods are called for, trees can still be helpful if there are a lot of variables, as they can be used to identify important variables and interactions. These are also invariant to the monotonic transformations of the explanatory variables and do not require the selection of the variable in advance as in regression analysis.

CART uses so-called learning set which is a set of historical data with pre-assigned classes for all observations. An algorithm known as recursive partitioning is the key to the nonparametric statistical method of CART. It is a step-by-step process by which a decision tree is constructed by either splitting or not splitting each node on the tree into two daughter nodes. An attractive feature of the CART methodology is that because the algorithm asks a sequence of hierarchical questions, it is relatively simple to understand and interpret the results. The unique starting point of a classification tree is called a root node and consists of the entire learning set L at the top of the tree. A node is a subset of the set of variables, and it can be terminal or nonterminal node. A non terminal (or parent) node is a node that splits into two daughter nodes (binary split). Such a binary split is determined by a condition on the value of a single variable, where the condition is either satisfied or not satisfied by the observed value of that variable. All observations in L that have reached a particular (parent) node and satisfy the condition for that variable drop down to one of the two daughter nodes; the remaining observations at that (parent) node that do not satisfy the condition drop down to the other daughter node. A node that does not split is called a terminal node and is assigned a class label. Each observation in L falls into one of the terminal nodes. When an observation of unknown class is "dropped down" the tree and ends up at a terminal node, it is assigned the class corresponding to the class label attached to that node. There may be more than one terminal node with the same class label.

Two major problems addressed in this point are: i) number of splits ii) Query Selection and Node Impurity.

To produce a tree-structured model using recursive binary partitioning, CART determines the best split of the learning set L to start with and thereafter the best splits of its subsets on the basis of various issues such as identifying which variable should be used to create the split, and determining the precise rule for the split, determining when a node of the tree is a terminal one, and assigning a predicted class to each terminal node. The assignment of predicted classes to the terminal nodes is relatively simple, as is determining how to make the splits, whereas determining the right-sized tree is not so straightforward. In order to explain these in details, procedure of growing a fully expanded tree and obtaining a tree of optimum size is explained subsequently.

In general, the number of splits is set by the designer and could vary throughout the tree. The number of links descending from a node is sometimes BRANCHING called the node's branching factor or branching ratio, denoted B. However, every FACTOR decision (and hence every tree) can be represented using just binary decisions. Thus, the root node querying fruit color (B = 3) in our example could be replaced by two nodes: The first would ask fruit = green?, and at the end of its "no" branch, another node would ask fruit = yellow?. Because of the universal expressive power of binary trees and the comparative simplicity in training, we shall concentrate on such trees Figure 5-5.

ii) Query Selection and Node Impurity: Recently a number of the work in designing trees focuses on deciding which property test or query should be performed at each node. With nonnumeric data, there is no geometrical interpretation of how the query at a node splits the data. However, for numerical data, there is a simple way to visualize the decision boundaries that are produced by decision trees. For example, Assumed that the query at each node has the form $x_i \leq x_{is}$ This leads to hyperplane decision boundaries that are perpendicular to the coordinate axes, and to decision

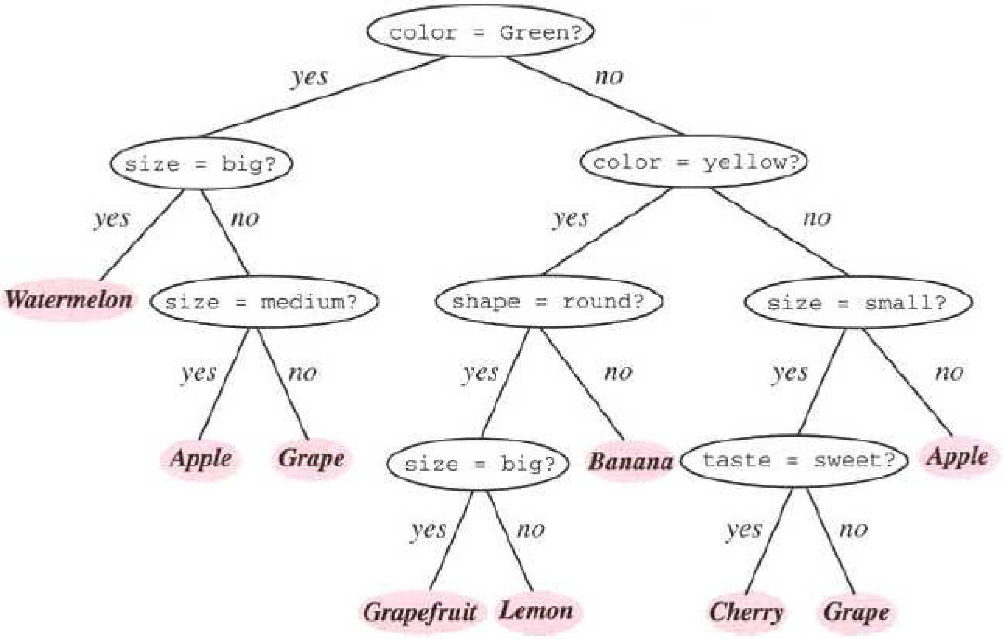


Figure 5-5: A tree with arbitrary branching factor at different nodes can always be represent by a functionally equivalent binary tree which is one having branching factor B=2 throughout, as shown here.

regions of the form shown in Figure 5-6.

The fundamental principle underlying tree creation is that of simplicity: The decisions that lead to a simple, compact tree with few nodes should be preferred. The important problem is why there is no reason the query at a node has to involve only one property. One might well consider logical combinations of properties, such as using (size=medium) AND (NOT(color=yellow))? as a query. Trees in which each query is based on a single property are called monothetic; if the query at any of the nodes involves two or more properties, the tree is called polythetic. For simplicity, we generally restrict our treatment to monothetic trees. In all cases, the key requirement is that the decision at a node be well-defined and unambiguous so that the response leads down one and only one branch. To this end, we seek a property query T at each node N that makes the data reaching the immediate descendant nodes as "pure" as possible. In formalizing this notion, it turns out to be more convenient to define the impurity, rather than the purity of a node.

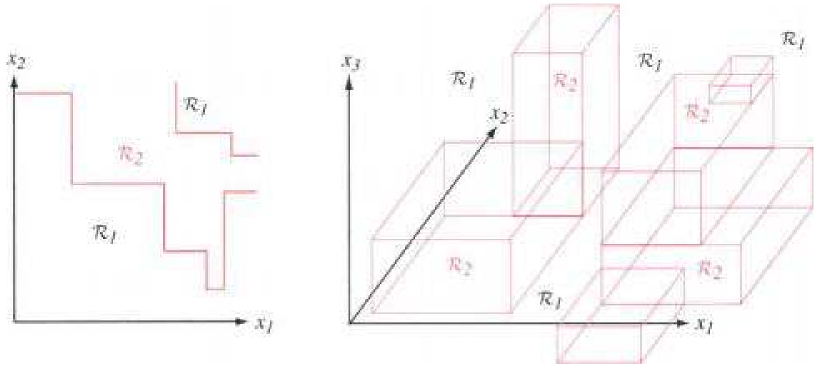


Figure 5-6: Monothetic decision trees create decision boundaries with portions perpendicular to the feature axes.

Several different mathematical measures of impurity have been proposed, all of which have basically the same behavior. Let $i(N)$ denote the impurity of a node N. In all cases, we want $i(N)$ to be 0 if all of the patterns that reach the node bear the same category label, and to be large if the categories are equally represented.
The most popular measure is the entropy impurity (or occasionally information impurity):

$$i(N) = -\sum_j P(\omega_j) log_2 P(\omega_j) \tag{5.5}$$

where $P(\omega_j)$ is the fraction of patterns at node N that are in category ω_j. By the well-known properties of entropy, if all the patterns are of the same category, the impurity is 0; otherwise it is positive, with the greatest value occurring when the different classes are equally likely. Another definition of impurity is particularly useful in the two-category case. Given the desire to have zero impurity when the node represents only patterns of a single category, the simplest polynomial form is

$$i(N) = P(\omega_1)P(\omega_2). \tag{5.6}$$

This can be interpreted as a variance impurity because under reasonable assumptions it is related to the variance of a distribution associated with the two categories. A generalization of the variance impurity, applicable to two or more categories, is the

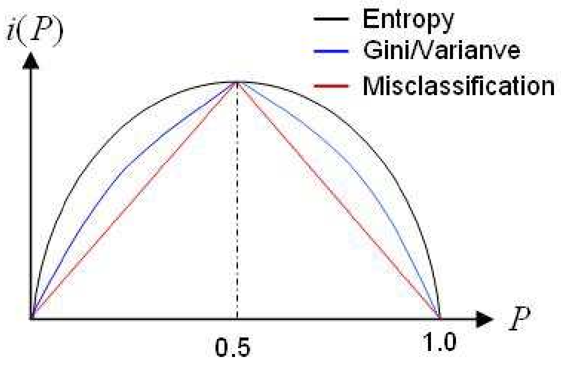


Figure 5-7: For the two-category case, the impurity functions peak at equal class frequencies and the variance and the Gini impurity functions are identical. The entropy, variance, Gini, and misclassification impurities have been adjusted in scale and offset to facilitate comparison here; such scale and offset do not directly affect learning or classification.

Gini impurity:

$$i(N) = -\sum_{i \neq j} P(\omega_j) log_2 P(\omega_j) = \frac{1}{2}\left[1 - \sum_j P^2(\omega_j)\right]. \qquad (5.7)$$

This is just the expected error rate at node N if the category label is selected randomly from the class distribution present at N. This criterion is more strongly peaked at equal probabilities than is the entropy impurity Figure 5-7.

The misclassification impurity can be written as

$$i(N) = 1 - maxP(\omega_j) \qquad (5.8)$$

and it measures the minimum probability that a training pattern would be misclassified at N. Of the impurity measures typically considered, this measure is the most strongly peaked at equal probabilities. It has a discontinuous derivative, though, and this can present problems when searching for an optimal decision over a continuous parameter space. Figure 5-7 shows these impurity functions for a two-category case, as a function of the probability of one of the categories.

There is still a key question: Given a partial tree down to node N, what value s should we choose for the property test T. An obvious heuristic is to choose the query that decreases the impurity as much as possible. The drop in impurity is defined by

$$\Delta i(N) = i(N) - P_L i(N_L) - (1 - P_L) i(N_R) \tag{5.9}$$

where N_L and N_R are the left and right descendant nodes, $i(N_L)$ and $i(N_R)$ are their impurities, and P_L is the fraction of patterns at node N that will go to N_L when property query T is used. Then the "best" query value s is the choice for T that maximizes $\delta i(T)$. If the entropy impurity is used, then the impurity reduction corresponds to an information gain provided by the query. Because each query in a binary tree is a single "yes/no" one, the reduction in entropy impurity due to a split at a node cannot be greater man one bit.

The way to find an optimal decision for a node depends upon the general form of decision. Because the decision criteria are based on the extrema of the impurity functions, we are free to change such a function by an additive constant or overall scale factor and this will not affect which split is found. Designers typically choose functions that are easy to compute, such as those based on a single feature or attribute, giving a monothetic tree. If the form of the decisions is based on the nominal attributes, we may have to perform extensive or exhaustive search over all possible subsets of the training set to find the rule maximizing δi. If the attributes are real-valued, one could use gradient descent algorithms to find a splitting hyperplane, giving a polythetic tree. An important reason for favoring binary trees is that the decision at any node can generally be cast as a one-dimensional optimization problem. If the branching factor B were instead greater than 2, a two or higher-dimensional optimization would be required; this is generally much more difficult.

5.4 Ensembles of Decision Trees

Machine learning approaches have wide applications in medical and decision tree is one of the oldest machine learning model and usually is used to illustrate the very basic idea of machine learning which is applied in this field.

The aim of adopting a decision tree ensemble is to obtain highly accurate performance in clinical decisions. It is well known that the classification performance can be improved by using an ensemble of individual classifiers, which concur to reach the final decision . Several ensemble learning methods exist and different strategies have been proposed to blend the predictions of individual classifiers, which are expected to be diverse and yet accurate. In particular, in classifier fusion, each ensemble member is supposed to have knowledge of the whole feature space. The features set can be as inputs in two categories, i) same feature set for all classifiers or ii) different feature set for each classifiers. In Figure 5-8 a schematic representation of these models are shown.

Two of the most popular techniques for ensemble fusion are bagging i.e., boostrap aggregation, Breiman (1996), and boosting, i.e., Adaboost Freund and Schapire (1996). Let us spend a few words on the former technique, which has been used in this work.

Bagging

The idea of bagging is simple and appealing: the ensemble is made of classifiers built on bootstrap replicates of the training set. The classifier outputs are combined by the plurality vote[15].

The diversity necessary to make the ensemble work is created by using versions of a training set. Ideally, the training sets should be generated randomly from the distribution of the problem. Each of these bootstrap data sets is used to train a different component classifier and the final classification decision is based on the vote of each component classifier. Traditionally the component classifiers are of the same general form-for example, all Hidden Markov Models, or all neural networks, or all decision trees merely the final parameter values differ among them due to their different sets of training patterns.

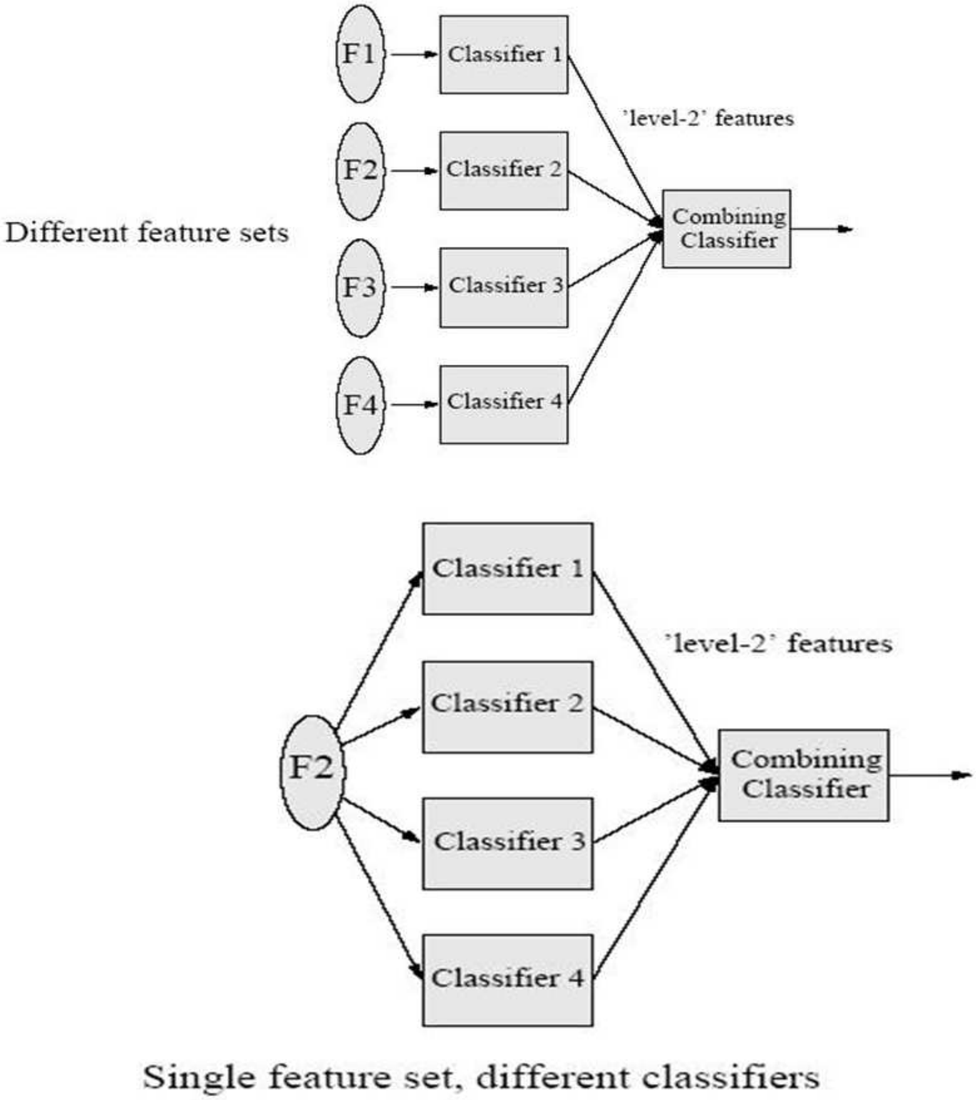


Figure 5-8: A schematic representation of the inputs system.

In practice, we can only afford one labeled training set, $Z = \{z_1, ..., z_N\}$, and have to imitate the process or random generation of L training sets. We sample with replacement from the original training set (bootstrap sampling [26]) to create a new training set of length N. To make use of the variations of the training set, the base classifier should be unstable, that is, if "small" changes in the training data lead to significantly different classifiers and relatively "large" changes in accuracy. Otherwise, the resultant ensemble will be a collection of almost identical classifiers, therefore unlikely to improve on a single classifier's performance. Examples of unstable classifiers are neural networks and decision trees while k-nearest neighbor is an example of a stable classifier. The training and operation of bagging is (The bagging algorithm):

Training phase

1. Initialize the parameters

   $D = 0$, the ensemble.

   L, the number of classifiers to train.

2. For $k = 1, ..., L$

   Take a bootstrap sample S_k from Z. Build a classifier D_k using S_k as the training set. Add the classifier to the current ensemble, $D = D \cup D_k$.

3. Return D.

Classification phase

4. Run $D_1, ..., D_L$ on the input x.

5. The class with the maximum number of votes is chosen as the label for x.

If the outputs of classifier were independent and classifiers had the same individual accuracy p, then the majority vote (we will talk in this chapter) is guaranteed to improve on the individual performance [42, 13]. Bagging aims at developing independent classifiers by taking bootstrap replicates as the training sets. Bagging is our first encounter with multiclassifier systems, where a final overall classifier is based on the outputs of a number of component classifiers. The global decision rule in bagging- a simple vote among the component classifiers is the most elementary method of pooling or integrating the outputs of the component classifiers. The samples are pseudo-independent because they are taken from the same Z. However, even if they were drawn independently from the distribution of the problem, the classifiers built on these training sets might not give independent outputs.

Boosting

We now look at the ensemble method of boosting. Boosting was inspired by an on-line learning algorithm called Hedge(β) [30]. The goal of boosting is to improve the accuracy of any given learning algorithm. In boosting we first create a classifier with accuracy on the training set greater than average, and then add new component classifiers to form an ensemble whose joint decision rule has arbitrarily high accuracy on the training set. In such a case we say that the classification performance has been "boosted". In overview, the technique trains successive component classifiers with a subset of the training data that is "most informative" given the current set of component classifiers. For example, we suppose that as a patient, you have certain symptoms. Instead of consulting one doctor, you choose to consult several. Suppose you assign weights to the value or worth of each doctor's diagnosis, based on the accuracies of previous diagnoses they have made. The final diagnosis is then a combination of the weighted diagnoses. This is the essence behind boosting.

For definiteness, consider creating three component classifiers for a two-category problem through boosting. First we randomly select a set of $n_1 < n$ patterns from the full training set D (without replacement); call this set D_1. Then we train the first classifier, C_1, with D_1. Classifier C_1 need only be a weak learner-that is, have accuracy only slightly better than chance. (Of course, this is the minimum requirement; a weak learner could have high accuracy on the training set. In that case the benefit of boosting will be small).

Now we seek a second training set, D_2, that is the "most informative" given component classifier C_1. Specifically, half of the patterns in D_2 should be correctly classified by C_1, half incorrectly classified by C_1. Such an informative set D_2 is created as follows: We flip a fair coin. If the coin is heads, we select remaining samples from D and present them, one by one to C_1 until C_1 misclassifies a pattern. We add this misclassified pattern to D_2. Next we flip the coin again. If heads, we continue through D to find another pattern misclassified by C_1 and add it to D_2 as just described; if tails, we find a pattern that C_1 classifies correctly. We continue until no more patterns can be added in this manner. Thus half of the patterns in D_2 are correctly classified by C_1, half are not. As such, D_2 provides information complementary to that represented in C_1. Now we train a second component classifier C_2 with D_2.

Next we seek a third data set, D_3, which is not well classified by voting by C_1 and C_2. We randomly select a training pattern from those remaining in D and then classify that pattern with C_1 and with C_2. If C_1 and C_2 disagree, we add this pattern to the third training set D_3, otherwise we ignore the pattern. We continue adding informative patterns to D_3, in this way; thus D_3 contains those not well represented by the combined decisions of C_1 and C_2. Finally, we train the last component classifier, C_3, with the patterns in D_3.

Now consider the use of the ensemble of three trained component classifiers for classifying a test pattern x. Classification is based on the votes of the component classifiers. Specifically, if C_1 and C_2 agree on the category label of x, we use that label; if they disagree, then we use the label given by C_3 (See Figure 5-9).

We skipped over a practical detail in the boosting algorithm: how to choose the number of patterns n_1 to train the first component classifier. We would like the final system to be trained with all patterns in D of course; moreover, because the final decision is a simple vote among the component classifiers, we would like to have a roughly equal number of patterns in each (i.e., $n_1 \cong n_2 \cong n_3 \cong n/3$). A reasonable first guess is to set $n_1 \cong n/3$ and create the three component classifiers. If the classification problem is very simple, however, component classifier C_1 will explain most of

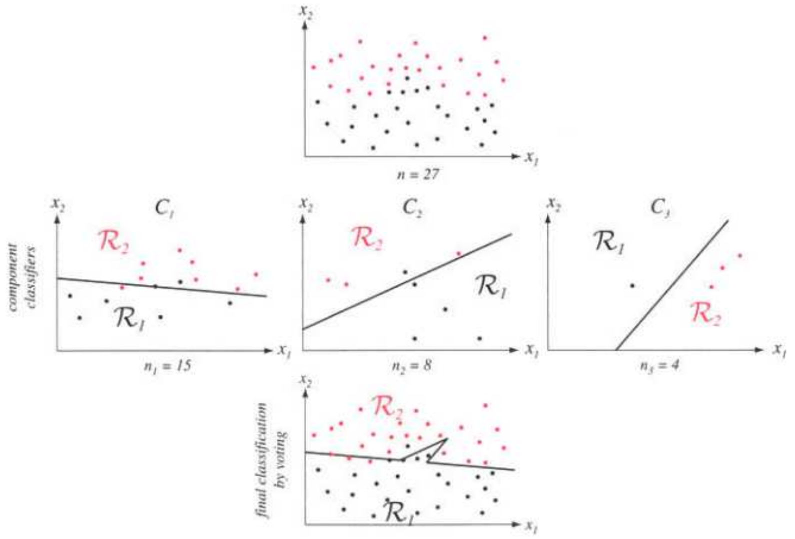


Figure 5-9: A two-category classification task is shown at the top in two-dimensions. The middle row shows three component (linear) classifiers C_k trained. where their training patterns were chosen through the basic boosting procedure. The final classification is given by the voting of the three component classifiers and yields a nonlinear decision boundary, as shown at the bottom. Given that the embedded classifiers are weak learners (i.e., each can learn a training set at least slightly better than chance), the ensemble classifier will have a lower training error on the full training set D than does any single component classifier. Of course, the ensemble classifier has lower error than a single linear classifier trained on the entire data set.

the data and thus n_2 (and n_3) will be much less than n_1, and not all of the patterns in the training set D will be used. Conversely, if the problem is extremely difficult, then C_1 will explain only a small amount of the data, and nearly all the patterns will be informative with respect to C_1; thus n_2 will be unacceptably large. Thus in practice we may need to run the overall boosting procedure a few times, adjusting n_1 in order to use the full training set and, if possible, get roughly equal partitions of the training set.

There are a number of variations on basic boosting. The most popular, AdaBoost from "adaptive boosting-allows" the designer to continue adding weak learners until some desired low training error has been achieved. In AdaBoost each training pattern receives a weight that determines its probability of being selected for a training set for an individual component classifier. The AdaBoost algorithm is as follows:

AdaBoost

1. begin initialize $D = x^1, y_1, ..., x^n, y_n, k_{max}, W_1(i) = 1/n, i = 1, ..., n$
   $k \rightarrow 0$

2. do $k + 1 \rightarrow k$

3. train week learner C_k using D sampled according to $W_k(i)$

4. training error of C_k measured on D using $W_k(i)$

5. $\frac{1}{2} ln[(1 - E_k)/E_k] \rightarrow \alpha_k$

6. $\frac{w_k(i)}{Z_k} \times e^{\alpha}$

```
7. $k = k_{max}$
8. return $C_k$ and $\alpha_k$ for $k = 1$ to $k_{max}$ (ensemble of classifiers with weights)
9. end
```

Note that in line 5 the error for classifier C_k is determined with respect to the distribution $W_k(i)$ over D on which it was trained. In line 7, Z_k is simply a normalizing constant computed to ensure that $W_k(i)$ represents a true distribution, and $h_k(x^i)$ is the category label ($+1 or -1$) given to pattern x^i by component classifier C_k. Naturally, the loop termination of line 8 could instead use the criterion of sufficiently low training error of the ensemble classifier.

"How does boosting compare with bagging?" Because of the way boosting focuses on the misclassified tuples, it risks overfitting the resulting composite model to such data. Therefore, sometimes the resulting "boosted" model may be less accurate than a single model derived from the same data. Bagging is less susceptible to model overfitting. While both can significantly improve accuracy in comparison to a single model, boosting tends to achieve greater accuracy.

6

Results and Discussion

This chapter illustrates experimental results concerning signal processing and classification of heart sound signals. The most important the goal of signal processing techniques is to extract efficient features, to be used as inputs for the classification process. The problem faced in feature extraction from row data is to determine what features are to be used. If too many features are extracted and used, the training process of the classifier at hand might be complex. On the other hand, if few features are selected the the information given to the training algorithm might be poor. Moreover, training of the classifier will be also difficult and testing results will be poor. In this research, we applied in pipeline feature extraction.

Classification involves assigning a class to an unknown object. Both supervised and unsupervised classification methods have been used for obtaining the final results of the analysis. In this work, several classification techniques have been experimented on available data.

6.1 Data Analysis and Feature Extraction

Data Analysis Results

Some relevant signal processing tools have been applied to newborn heart sound signals. This section reports experimental results and discusses the applications of each tool that better highlights the properties of the PCG signal, with the goal of identifying those that are more suitable for classification purposes. In particular, the following tools have been selected: Shannon Energy, Spectrum, Bispectrum, Wigner Distribution and Wigner Bispectrum.

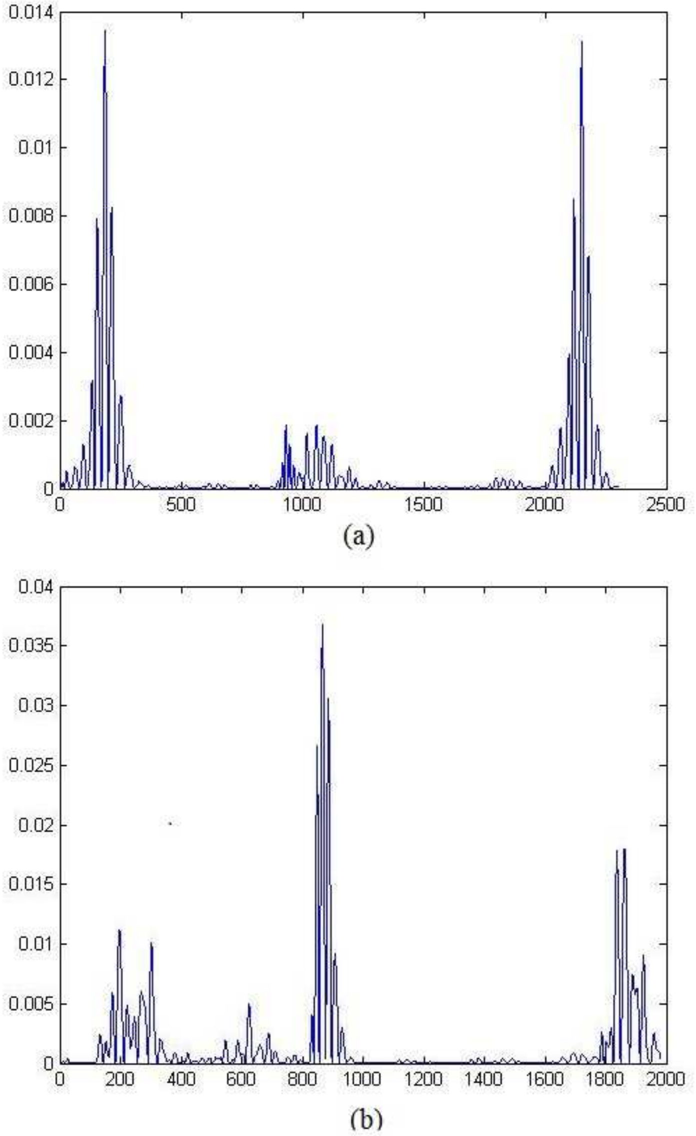
Shannon Energy The Shannon Energy of a PCG signal is shown in figure 6-1. In our case signal segments have been obtained with a granularity of 0.02 seconds and with signal segment overlapping of a 0.01 seconds. According to these figures (as shown in chapter 4), we can see that Shannon entropy and Shannon Energy can absorb the magnitude of oscillations of high intensity as well as those in low amplitudes.

Figure 6-1 a) shows a cardiac cycle magnified, indicating the normal heart sound murmurs. Figure 6-1 b) shows pathological heart sound magnified, indicating early systolic murmurs. The figures clearly highlights the existence of significant extra peaks in the systolic area of pathological sample. This technique proves very useful in enhancing signal details and can generally be applied on any pathological case.

Spectrum Time-frequency visualization is a common preliminary step in the analysis of nonstationary signals. The most popular technique is the spectrogram, which estimates the power spectral density (PSD) by applying the periodogram to windowed segments separated by a fixed interval. This is computationally efficient because it incorporates the Fast Fourier transform (FFT). The user specifies the window shape and length that controls the trade-off between time and frequency resolution of the image.

The time-frequency representation of the PCG signal is evaluated with the help of the spectrogram. This is a very important parameter for window analysis. We estimated spectrograms using an FFT applied to a series of signal segment multiplied by a Hann window. The window was based on first and second heart sound. Figure 6-2 shows the spectrogram of first and second heart sound in newborn.

The decreasing frequency and growing amplitude are clearly visible in the spectrogram. Two beats of normal PCG signal and appropriate spectrogram are shown in figure 6-2. Heartbeat consists of time intervals that are determined with S1 and S2,



(a)

(b)

Figure 6-1: Shannon energy of heart sound with a) normal and b) pathological murmurs.

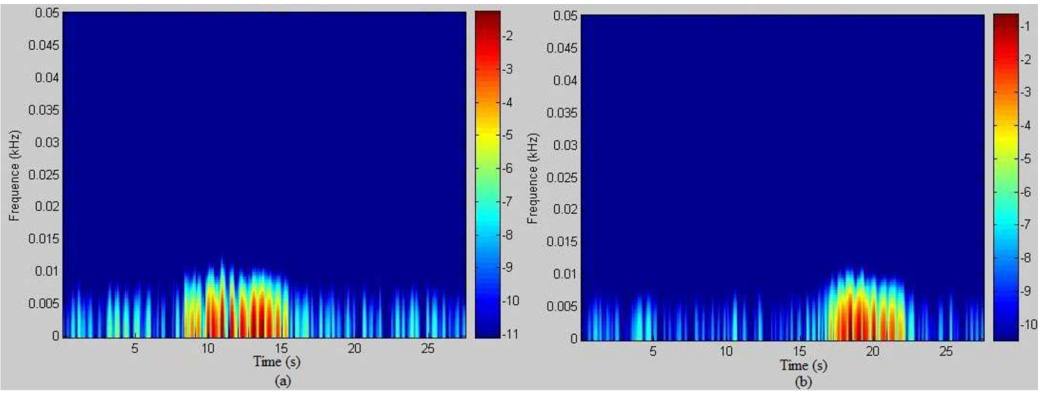


Figure 6-2: Spectrogram of a) first (S1) and b) second (S2) heart sound.

i.e., systole and diastole peaks, where diastole has a longer interval than systole. In this figure we can find out that the first heart sound S1 contains more information than the second one. Indeed, this systolic sound occupies a relatively large bandwidth, from 20 Hz up to 100 Hz, when acquired from mitral focus upon the chest. The diastolic heart sound S2 appears in very short duration with approximately the same spectral range of S1, but with different shape. Periods of heart murmurs are characterized by horizontal bands of elevated power spectral density (PSD) in the range of 0.01 kHz. These bands indicate low-frequency oscillations in the heart sound. By applying the spectral analysis to different PCG signals, we can identify which sounds (either S1 or S2) is directly concerned by the pathology, and more precisely which component of these sounds is affected.

Bispectrum Bispectral analysis makes use of phase information by detecting whether the phase of signal components at frequencies f_1, f_2, f_3 and are interdependent. The Bispectrum has applied to the normal first and second sounds to analyze the frequency content, as shown in Figure 6-3. The two internal components for the sound S1 (M1 and T1) and the two components A2 and P2 of the sound S2 are obvious in figure 6-3. This Biscpectrum analysis cannot give the time delay between these internals components. Therefore, the usual Bispectrum is unable to accurately diagnose heart diseases. It is thus essential to look for a transform which will describe a kind of "time-varying" spectrum.

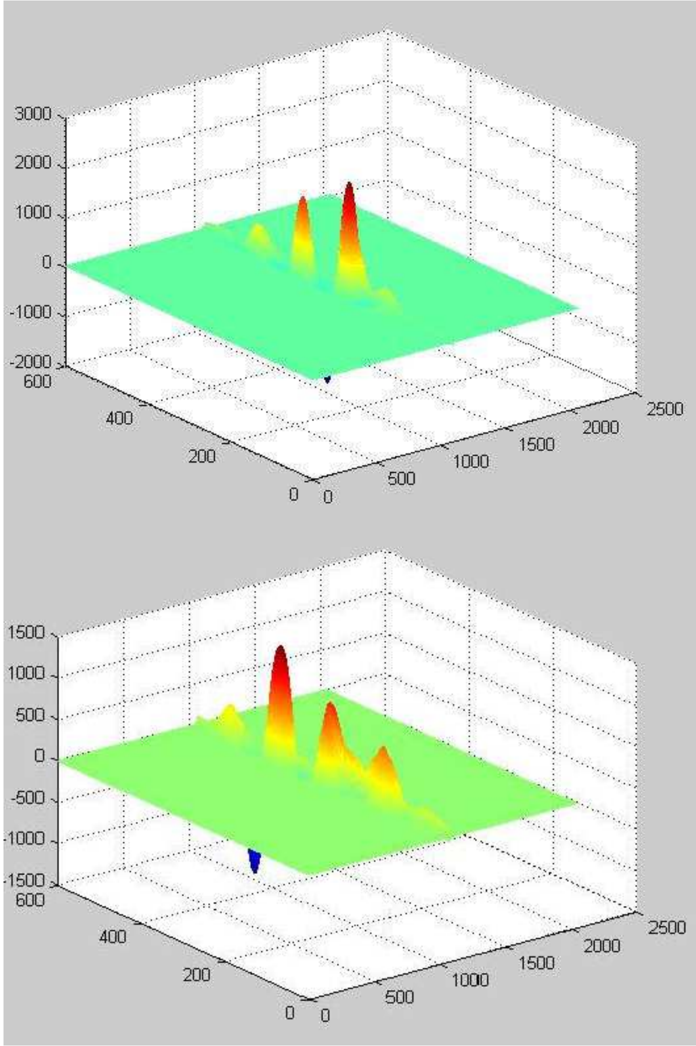


Figure 6-3: Bispectrum of first and second heart sounds with components.

The implementation of Bispectrum on normal and pathological heart murmurs data is shown in figure 6-4, which clearly highlights the existence of significant peaks in the bispectra. Thorough experimental results shows that the same kind of heart murmurs have significant similarities in their Bispectra shapes and in the locations of peaks. The location of significant bispectral peaks in bifrequency are quite different depending on the PCG signal in different kind of diseases.

Wigner Distribution

As we discussed in chapter 4, Wigner Distribution provide a time-frequency shifted

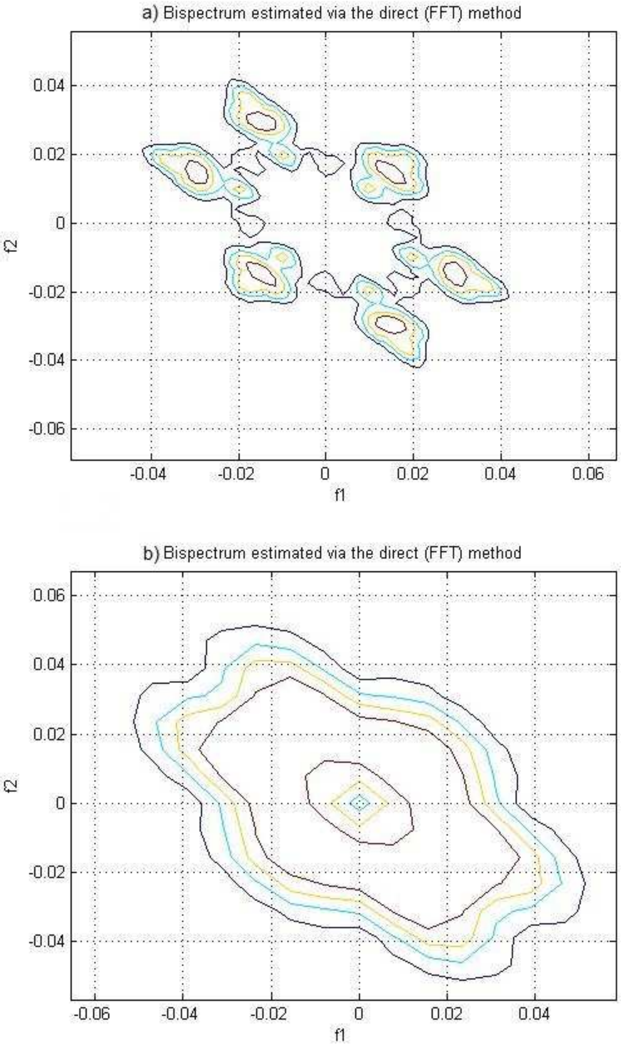


Figure 6-4: A contour plot of the magnitude of the indirect estimated bispectrum on the bifrequency plane, for (a) normal and (b) pathological heart murmurs.

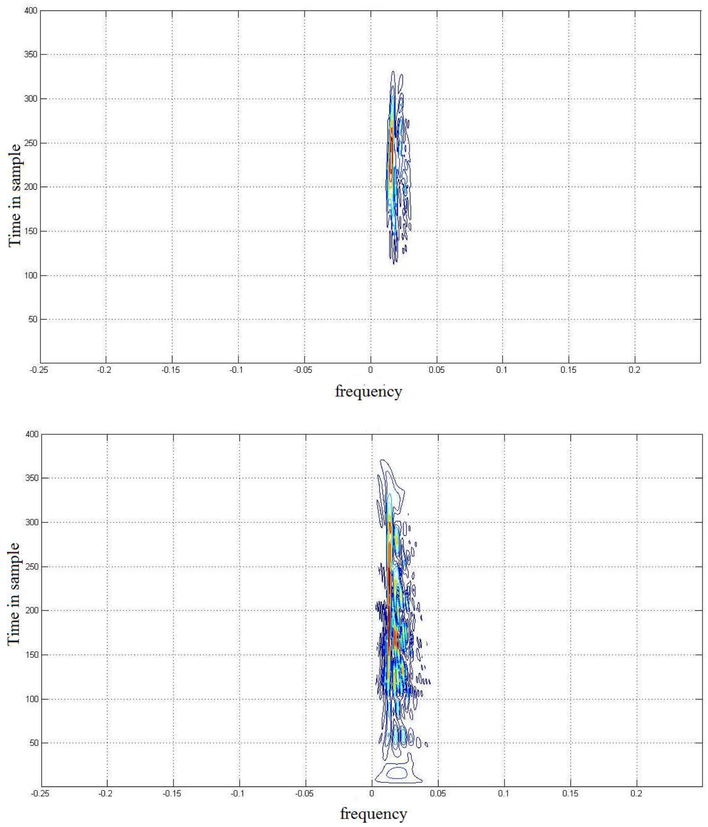


Figure 6-5: Wigner Distribution of normal (upper) and pathological (lower) first cardiac sound.

versions of PCG signal. Figure 6-5 shows the WD applied to a normal (upper) and pathological lower (S1). The spectrogram is calculated by a linear then a bilinear operations. Firstly, the linear operator consists of a Fourier transform, and secondly the squared modulus as a bilinear operator is applied to the signal to be analysed. In contrast, the WVD begins with a quadratic estimation of the energy and then a Fourier transform is applied to the signal according to equation 4.21. The WVD combines the time and the frequency representations with some required properties to adequately represent a given signal $x(t)$ in the time-frequency domain [14].

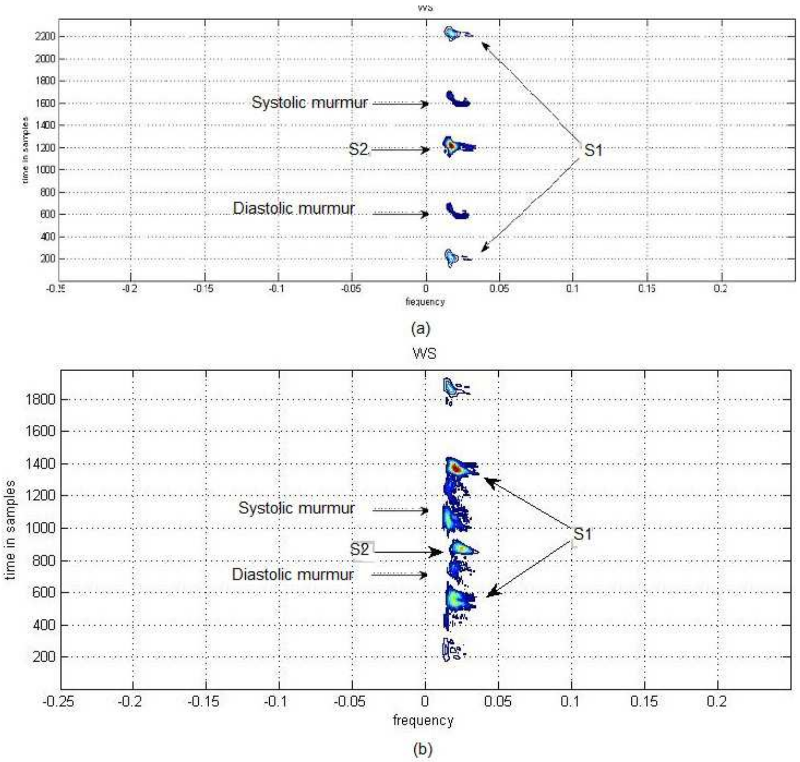
In contrast, the WVD provides an extraordinary time-frequency representation which retraces perfectly the differentiation of the S1 heart sound. One can notice here that the two main components (A2 and P2) start to appear in the presence cross-terms. The information contained in a WD may be improved by increasing the sampling rate of the original signal, but it would still suffer from the cross-terms problem due to the nonlinearity of the WD analysis. However the WD have shown good performances in the analysis of non-stationary signals. This comes from ability to separate signals along both time and frequency directions. One advantage of the WD over the STFT is that it does not suffer from the time-frequency trade-off problem. On the other hand, the WD has a disadvantage since it shows cross-terms in its response. These cross-terms are due to the nonlinear behavior of the WD, and bear no physical meaning. One way to remove these cross-terms is by smoothing the time-frequency plane, but this will be at the expense of decreased resolution in both time and frequency. We believe that the time-frequency scaling of the PCG may find important applications in the improvement of the diagnosis of heart and heart valve disease.

Wigner Bispectrum The Wigner distribution is defined somewhat differently from the spectrogram, but also provides information regarding the frequency content of the signal versus time. Like the spectrogram, it reduces to the power spectrum when the signal is stationary. Using the Bispectrum as a model, we have extended the Wigner distribution to third order in away that preserves many of its essential features and appealing properties. A third-order Wigner distribution or Wigner Bispectrum of heart sound with normal (innocent) and pathological murmurs are shown in figure 6-6.

Figure 6-6 clearly highlights S1 and S2, together with the feature of heart murmur which are systolic and diastolic murmurs. For the sake of readability, they are put into deviance for separating innocent and pathologic murmurs.

Feature Extraction and Selection

This phase is focused on extracting signal features that better highlight the properties of the PCG signal, with the goal of identifying those that are more suitable for classification purposes. This part consists of two major steps: feature extraction from row data and feature selection.



(a)

(b)

Figure 6-6: Contour map of the Wigner Bispectrum from heart sound: (a) Heart sound with innocent murmur and (b) pathological murmur.

Feature Extraction

Feature extraction is an essential perprocessing step in pattern recognition and machine learning problems. It is often followed by feature selection.
In our case, each signal is represented by the features summarized in Table 6.1.

No	**Variables**	**Feature Set**
1	Max	Maximum Value Amplitude
2	Min	Minimum Value Amplitude
3	Positive Area	Sum of Positive Area
4	Absolute Negative Area	Absolute Sum of Negative Area
5	Total Absolute Area	Sum of Absolute Area
6	Variance	Variance
7	Peak to Peak	Peak to Peak Time window
8	SE	Shannon Energy
9	C_1, C_2, C_3	Bispectrum
10	WD	Wigner Distribution
11	WB	Wigner Bispectrum

Table 6.1: List of Features Extracted for Classification

Overall, a total of 13 features from time domain, frequency domain, higher order spectral and statistical features were extracted that could have potential to discriminate among the normal and murmur signals. This study uses the common assumption that systole is shorter than diastole. Unlike other studies, the features in the whole signal as well as separately in systolic and diastolic regions have been extracted in order to deal with the situations of systolic and diastolic murmur. All the features are calculated from the available PCG signals which are 116 samples.

Feature Selection

Feature selection has then been used to reduce the size of the feature vector. To measure the score of each variable we made use of the gain and variable importance

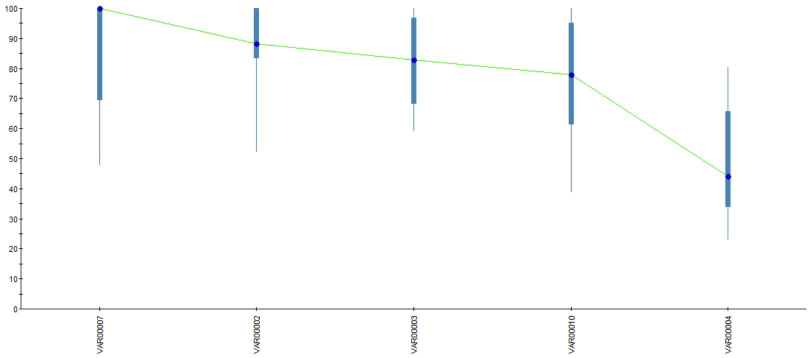


Figure 6-7: Variable importance averaging.

metrics which are provided as secondary outputs by the algorithm used for training an ensemble of decision trees.

To calculate the importance score of a variable, the training algorithm for decision trees looks at the improvement measure of each variable, in its role as a surrogate to the primary split. The values of these improvements are summed over each node and are scaled according to the best performing variable [12].
A variable can obtain an importance score of zero in a decision tree only if it never appears as primary or surrogate splitter. As such kind of variables play no role anywhere in the tree, eliminating them from the data set does not affect the training process. Importance variable scores for the process in hand are reported in Figure 6-7.

Figure 6-7 shows all variables used (or not used) in the tree building process. A score is associated to each variable, based on the improvement each variable makes as a surrogate to the primary splitting variable. Variable importance allows to highlight variables whose significance is masked or hidden by other variables in the tree building process.

In this study some significant features have been introduced. By using these features the classification accuracy improved in case of various classifiers as shown in Table 6.2. So, these new features proved to be very efficient for classification of normal and murmur signals. Here, also proposes the approach of evaluating some

Table 6.2: Score of importance variable

No	Feature	Improvement	IVS
7	Peak to Peak	0.26727	100.000
1	Maximum	0.24242	98.3110
8	Shannon Energy	0.22668	82.6663
9	Bispectrum, C1	0.20649	72.6115
12	Wigner Bispectrum	0.16474	54.7318
4	Absolute Negative Area	0.14060	48.2417
11	Bispectrum, C2	0.03019	43.8223

already existing features in both the systolic and diastolic regions in order to deal with the situations of systolic and diastolic murmurs. In particular, the variable with the highest sum of improvements is scored 100, while other variables have lower score. Importance variable scores (IVS) are summarized in Table 6.2.

6.2 Classification of Heart Diseases in Newborn

This section reports experimental results and discusses about machine learning and data mining techniques that applied in this research. Various classifiers were used in this study in order to find out the best classifier that suits the problem. The goal of implementation of classifiers is reducing of two types of errors. As we discussed about heart murmurs in newborns that an innocent heart murmur does not entail a disease condition, a physician assuming that a newborn is healthy typically orders an echocardiogram for reassurance, although its cost may be not negligible. Overall, the result of this practice is a misallocation of health care funds. Indeed, while it is clearly important to avoid type-I errors, i.e. healthy newborn sent for echocardiogram, it is also important to avoid type-II errors, i.e. newborns having a pathological heart murmur sent home without proper treatment

Artificial Neural Networks

The first classifier that has been implemented for diagnosis of systolic defect consists of a multilayer perceptron with one hidden layer. The final ANN architecture was determined by trial and error. In particular, to identify the number of hidden layers and the amount of neurons to be used in the input and hidden layers, system complexity was reduced until performance began to degrade. The result of this procedure was MLP architecture with the hidden layer equipped with 5 neurons.

We have trained the classifiers on medical data labeled healthy or unhealthy. Sensitivity, see Equation 6.1, is a very important measure for this particular research, as in our case it measures the percentage of patients with unhealthy hearts that are recognized as such. High sensitivity means that the system has fewer Type II errors, i.e. few unhealthy hearts classified as healthy [65].

$$Sensitivity = \frac{(true \;\; positives)}{(true \;\; positives + false \;\; negatives)} \tag{6.1}$$

Specificity, see Equation 6.2, in our case gives the percentage of healthy cases that are classified as healthy. With high specificity, the system has fewer Type I errors, i.e. a healthy newborn classified as unhealthy.

$$Specificity = \frac{(true \;\; negatives)}{(true \;\; negatives + false \;\; positives)} \tag{6.2}$$

The classification accuracy was calculated using the leave-one-out cross-validation[1] method, which repeatedly trains the classifier with all samples but one and tests it on the sample excluded from training. The method iterates over all available samples and the final performance metrics are obtained by considering the results of all steps. Results are shown in Table 6.3 in the form of a confusion matrix, together with percentage classification accuracy.

[1] One of the most common forms of cross validation is "leave-one-out" (LOO) in which the model is repeatedly refit leaving out a single observation and then used to derive a prediction for the left-out observation.

	Normal	Pathological
Normal	96.4%	3.6%
Pathological	3.6%	96.4%

Table 6.3: Classification results of systolic murmurs in newborns. Cross-validation: obtained with the leave-one-out method, %: Percentage of classification.

It can be seen that out of 28 normal signals, 96.4% were correctly classified as normal, and 3.6% were misclassified as pathological. Similarly, out of 28 pathological signals, 96.4% were correctly classified as pathological and 3.6% were misclassified as normal. A detailed analysis of the misclassified signals showed that they were in fact very difficult to classify, even by human experts. Summarizing, 96.4% accuracy, 97% sensitivity and 97% specificity were obtained by the MLP, distinguishing between the 58 innocent and pathological heart murmurs in newborns.

Let us point out that for this system, both high sensitivity and specificity are important. In particular, higher specificity reduces the number of newborns with innocent murmurs who are identified as pathological murmur and sent to echocardiogram for further testing. More importantly, higher sensitivity reduces the number of newborns with pathological murmurs that are identified as innocent murmurs and have been released with a potentially deadly heart condition.

Classification and Regression Trees (CART)

This section reports experimental results and discusses the application of Classification and Regression Trees (CART) to early diagnosis of heart disease in newborns. Early diagnosis of heart murmurs in newborns is a novel application of CART for clinical and physiological data. K-fold cross validation (K=10) has been used as training and test strategy. Experiments have been run using an implementation of CART provided by Salford System Inc, USA. It is a step-by-step process in which a decision tree is constructed by either splitting each node on the tree in two daughter nodes.

The realistic objective of partitioning is to find partitions of the data such that terminal nodes are as such homogeneous as possible. The quantitative measure of node homogeneity is called the impurity function. The simplest idealization of the impurity function is the number of patients who meet an objective criteria divided by the total number of patients in the node. Ratios close to 0 or 1 are considered more pure.
To partition a node, CART examines all possible splits of the explanatory variables. In general, the number of possible splits for ordinal or continuous variables is 1 less the number of distinctly observed values. A potential split is judged by its reduction of the impurity function for both daughter nodes it creates. The partitioning iteratively continues by splitting each node in two daughter nodes and continues until the tree is saturated that is, until no further partitions can be found [70].
The DT start at the top of the tree and follow different branches, depending on conditions involving the predictor variables. Trees with multiple layers of splits may be conceptualized as describing interactions between predictor variables. Once we arrive at an end-point of the tree, we used 12 nodes and variables classified in two classes (classes 0 and 1 were Innocent and pathological murmurs respectively [47]).

We calculated the likelihood ratio (LR) to obtain sensitivity and specificity on a tree, defined as follows:

$$LR+ = \frac{sensitivity}{1 - specificity} \tag{6.3}$$

$$\tag{6.4}$$

$$LR- = \frac{1 - sensitivity}{specificity}$$

The interpretation of likelihood ratios is intuitive: the larger the positive likelihood ratio, the greater the likelihood of heart disease; the smaller the negative likelihood ratio, the lesser the likelihood of heart disease.

Results are shown in Table 6.2 in the form of a confusion matrix, together with classification accuracy. It can be seen that out of 58 normal signals, 57 were correctly

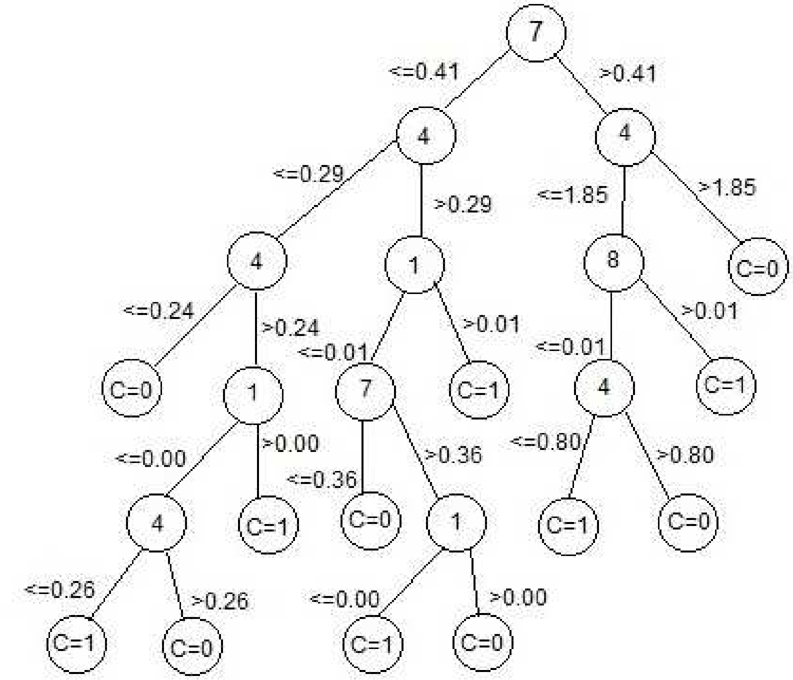


Figure 6-8: Illustration of decision tree structure.

Actual Group	Normal	Pathological	Percent Correct
Normal	57	1	98.28%
Pathological	0	58	100%
Average/Overall	116		99.14%

Table 6.4: Classification result of heart disease in newborns using CART.

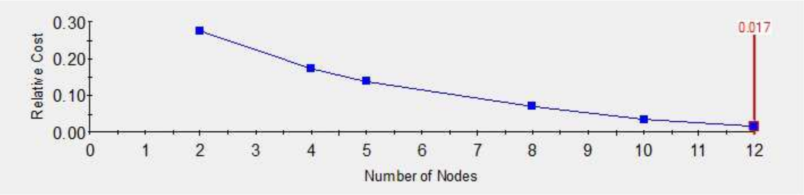


Figure 6-9: CART decision tree Error curve.

classified as normal, and 1 was misclassified as pathological. As for 58 pathological signals, they were correctly classified as pathological without misclassification. A detailed analysis of the misclassified example showed that it was in fact very difficult to classify, even by human experts.
Summarizing, 99.14% accuracy, 100% sensitivity and 98.28% specificity were obtained by CART, when used to distinguish between the 116 innocent and pathological heart murmurs in newborns.

Let us point out that, for this system, both high sensitivity and specificity are important. In particular, high sensitivity reduces the number of newborns with innocent murmurs who are identified as pathological murmur and sent to echocardiogram for further testing. More importantly, high specificity reduces the number of newborns with pathological murmurs that are identified as innocent murmurs and have been released with a potentially deadly heart condition.

For each fold, learning has been performed in two steps: growing and pruning. It is worth noting that pruning has been performed provided that decision tree error curve did not trespass the threshold of 1%.

Figure 6-9 shows a curve which outlines the relationship between classification errors and tree size. The scale is always between 0 and 1, so it is called a relative error curve. A tree with a relative error of 0 or nearly 0 is usually too good to be true. The proposed model shows excellent performance for application of diagnosis of heart disease.

The CART decision tree error curve archived automated growing of a too large tree, followed by automated pruning to find the right-sized tree [52]. The rationale for the growing/pruning process is illustrated in the error curve (figure 6-9). In a

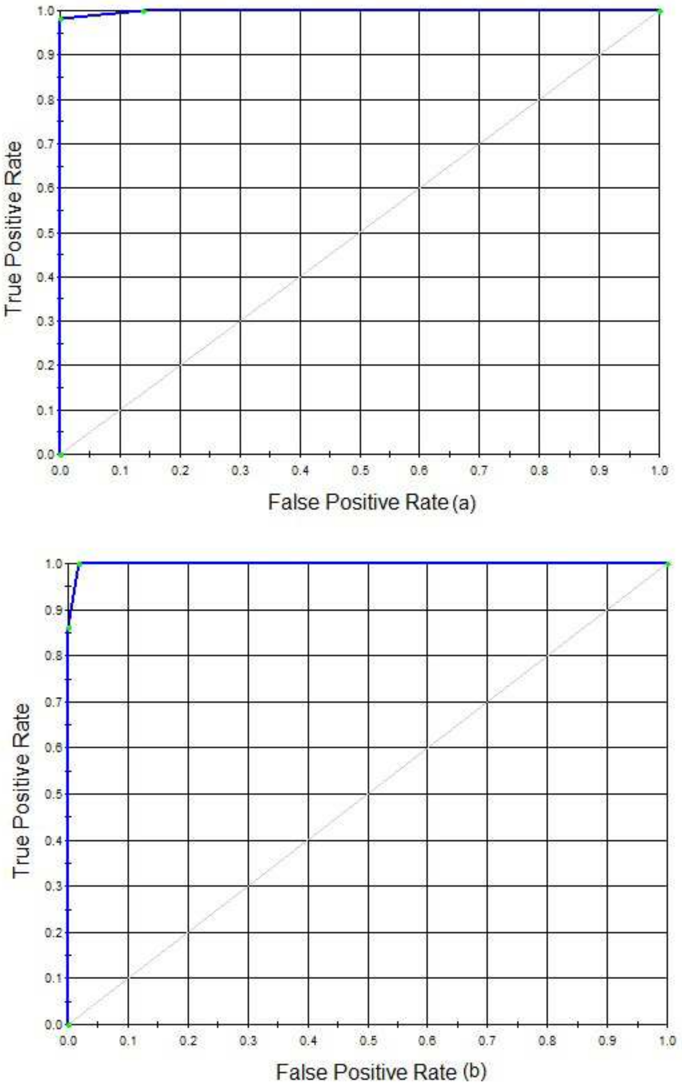


Figure 6-10: ROC curve of innocent (a) and pathological murmurs (b) classified.

Receiver Operating Characteristic (ROC) [2] curve for a binary classification problem, the true positive rate (Sensitivity) is reported as function of the false positive rate (100-Specificity) for different cut-off points. ROC curve are reported in figure 6-10 a and figure 6-10 b are targeted for innocent and pathological murmurs, respectively.

A successful classifier will result in an ROC curve tending towards the upper-left corner. The area under the ROC curve (AUC) is often used as a summary statistic since

[2] In statistics, a receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings.

it relates to the Mann-Whitney U-test. A predictive model with perfect performance has an area under ROC curve equal to 1. We obtained, on average, an accuracy of 0.99 the ROC curve highlights the excellent performance of CART to discriminate of heart murmurs.

Ensembles of Decision Trees

As discussed in a previous chapter a decision tree is typically trained using a greedy procedure which, at each node of the tree, decides whether to assign a class label to the node or to recursively split the node in two or more daughter nodes. The whole process ends when no more splitting is required (or feasible). The decision of assigning a class label to a node or continuing with splitting depends on the training samples associated to the node.

The goal of the underlying partitioning procedure is to split data such that terminal nodes are as much homogeneous as possible. The quantitative measure of node homogeneity is called impurity function. The simplest idealization of the impurity function is the number of patients who meet an objective criterion divided by the total number of patients in the node. Ratios close to 0 or 1 are considered more pure. To partition a node, the training algorithm of a decision tree examines all possible splits of the explanatory variables. In general, the number of splits for ordinal variables is the number of distinctly observed values minus 1. Several proposals have also been made to deal with continuous variables (see for instance Usama M. Fayyad, Keki B. Irani (1992)). A potential split is judged by its reduction of the impurity function for all daughter nodes it creates. The partition process keeps splitting nodes until no further partitions can be found [57].

A DT starts at the top of the tree and follow different branches, depending on the conditions involving the predictor variables. Trees with multiple layers of splits may be conceptualized as describing interactions between predictor variables [70].

The aim of adopting a decision tree ensemble is to obtain highly accurate performance in clinical decisions. It is well known that the classification performance can be improved by using an ensemble of individual classifiers, which concur to reach the final decision. Several ensemble learning methods exist and different strategies have been proposed to blend the predictions of individual classifiers, which are expected to be diverse and yet accurate. In particular, in classifier fusion, each ensemble member is supposed to have knowledge of the whole feature space. As for the corresponding voting policy, let us use a metaphor involving doctors and newborns: let us suppose that there is a newborn patient and that we would like to have a diagnosis made based on her/his symptoms. Instead of asking one doctor, one may decide to ask several doctors. The final diagnosis can be obtained using majority voting, in which the underlying assumption is that each doctor has the same power to influence the outcome of voting. Replacing each doctor with a single decision tree, we render the main idea that lies behind bagging with majority voting. It is worth pointing out that bagging works better when base classifiers are unstable. Indeed, decision trees are known to be unstable (also due to the greedy policy adopted by the partitioning procedure), as small changes in the training set can result in significantly different trees.

For testing the performance of individual classifiers, the hold-out method (sometimes called test sample estimation) has been used. This method requires the given data be randomly partitioned in two independent sets (training and testing) [40]. A common solution consists of designating two thirds of the data as training set and one third as test set. The hold-out method performs a fixed number of experiments on the given data. At each experiment, data is split in training and test set using random sub-sampling. Experimental results are then averaged over the splits.
This section reports experimental results and discusses the application of the adopted ensemble of decision trees to perform clinical decisions for newborns with heart murmurs. The final ensemble consists of 12 trees, used as classifier ensemble –as shown in Figure 6-11.

Results are shown in Table 6.5 in form of confusion matrix. The blending policy for decision trees was majority voting.

Experiments have been performed on a balanced set of 110 samples (meaning that

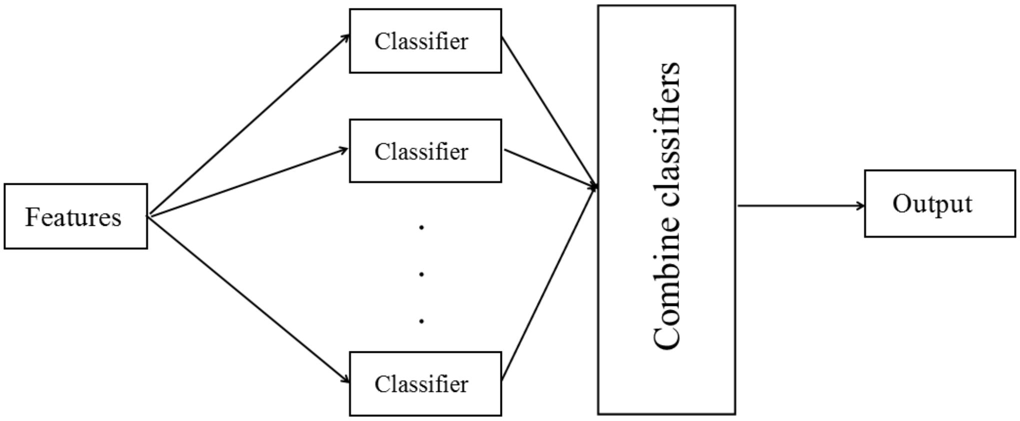


Figure 6-11: A schematic representation of the system.

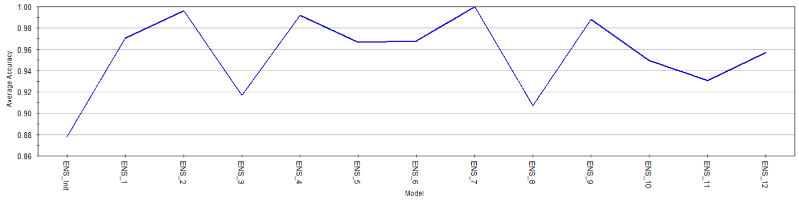


Figure 6-12: Average accuracy of modules.

the number of samples was the same for physiological and pathological murmurs). The overall accuracy has been calculated by averaging the results obtained over 20 runs with decision tree ensembles. Note that the minimum error on accuracy is obtained by tree 7 (see Figure 6-12).

It can be seen that, out of 36 normal signals (72 random samples as training set and 38 random samples as testing set), on average 91.82 % were correctly classified as normal, and 8.18% were misclassified as pathological. As for the 36 pathological signals, on average 96.28% were correctly classified as pathological and 3.72% were misclassified as physiological. It is worth pointing out that a detailed analysis performed on the misclassified examples highlighted that it was indeed very difficult to classify them even by human experts.

As already pointed out, the final decision on each sample submitted to the system is taken by majority voting. Majority voting considers all outputs of classifiers and makes decision based on 50 percent of votes+1. With N_c number of classifiers and p probability for each classifier to give the correct answer, the following equation holds

Table 6.5: Ensemble of decision trees results.

Model	Average ROC	Average Accuracy	Overall Accuracy	P_{maj}	$P_{maj} - p$
Tree 0	0.8784	0.8784	0.8750	0.9	0.0
Tree 1	0.9704	0.9667	0.9583	0.9	0.0
Tree 2	0.9965	0.9762	0.9792	0.9	0.0
Tree 3	0.9167	0.9167	0.8958	0.9	0.0
Tree 4	0.9921	0.9444	0.9375	0.9	0.0
Tree 5	0.9667	0.9667	0.9583	0.9	0.0
Tree 6	0.9677	0.9677	0.9583	0.9	0.0
Tree 7	1.0000	1.0000	1.0000	1	0.1
Tree 8	0.9071	0.9071	0.9167	0.9	0.0
Tree 9	0.9883	0.9844	0.9792	0.9	0.0
Tree 10	0.9496	0.9165	0.9167	0.9	0.0
Tree 11	0.9310	0.9310	0.9167	0.9	0.0
Tree 12	0.9570	0.9375	0.9583	0.9	0.0

for the accuracy of the ensemble [41], provided that (i) N_c is odd, (ii) classifiers are homogeneous, and (iii) the outputs of classifiers are (largely) independent:

$$P_{maj} = \sum_{M=(N/2)+1}^{N} \begin{pmatrix} L \\ M \end{pmatrix} p^M 1 - p^{N-M} \tag{6.5}$$

It can also be shown that, with $p > 0.5$ and P_{maj} monotonically increasing, $P_{maj} \to 1$ for $N \to \infty$.

The probabilities of correct classification of the ensemble for p accuracies of 12 trees are displayed in Table 6.5, which shows the individual accuracy required by a pool of decision trees so that highest possible $P_{maj} = 1$ is obtained. Tree 7 can be identified as the "pattern of success". According to Kuncheva's definition [41] the pattern of success is a distribution of the Y classifier outputs for a pool D such that the probability of any combination of correct and incorrect votes and the probability of all Y votes being incorrect.

For values of individual accuracy $p > 0.5$, the pattern of failure is always possible. The pattern of failure is symmetrical with the pattern of success in ensemble classifiers.

The upper and lower bounds of the majority vote accuracy in various individual accuracies are defined by Matan [50] as function of the pattern of success and the pattern of failure respectively. Given an ensemble of N classifiers $\{D_1, D_2, ..., D_i\}$ and $k = N + 1/2$. The upper and lower bound for majority voting are:

$$max\ P_{maj} = min\left\{1, \sum(k), \sum(k-1), ..., \sum(1)\right\} \tag{6.6}$$

$$min\ P_{maj} = max\left\{0,\ \xi(k),\ \xi(k-1),\ ...,\ \xi(1)\right\} \tag{6.7}$$

where

$$\sum(m) \equiv \frac{1}{m}\sum_{i=1}^{N-k+m} p_i, \qquad m = 1, ..., k \tag{6.8}$$

$$\xi(m) \equiv \frac{1}{m}\sum_{i=k-m+1}^{N} p_i - \frac{N-k}{m} \qquad m = 1, ..., k \tag{6.9}$$

We have archived 1 and 0.79 as upper and lower bounds on the classifiers, which are indicators for the performance of individual members. Results are best- and worst-case scenarios and not necessarily typical. All possible combinations of correct/incorrect votes of the N classifier outputs. The pattern of success is when the correct votes are used in the most efficient way, whereas the pattern of failure is when most correct votes are "wasted"

In a Receiver Operating Characteristic (ROC) curve for a binary classification problem, the true positive rate (i.e., sensitivity) is reported as function of the false positive rate (1-specificity) for different cut-off points. The ROC for normal and pathological murmurs is reported in Figure 6-13. The area under the ROC curve (AUC) is used as a measure of performance. A predictive model with perfect performance has an AUC equal to 1. On average, we obtained a value of 0.9587 for AUC, highlighting the excellent performance of the decision tree ensemble in the task of discriminating heart murmurs.

Summarizing, 93.91% accuracy, 96.15% sensitivity and 91.67% specificity were obtained by the decision tree ensemble, when used to distinguish between the 116 innocent and pathological heart murmurs in newborns.

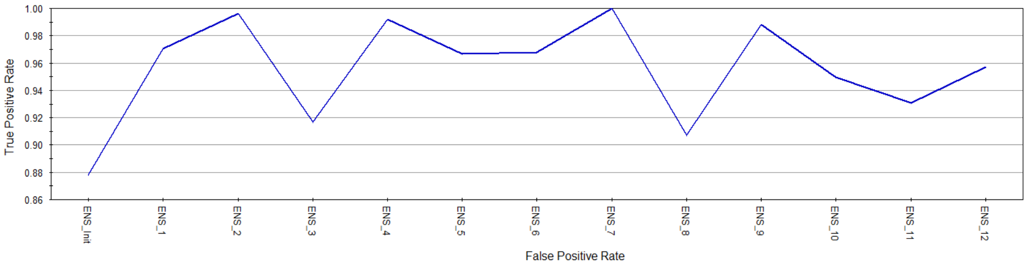


Figure 6-13: ROC curve of normal and pathological murmurs.

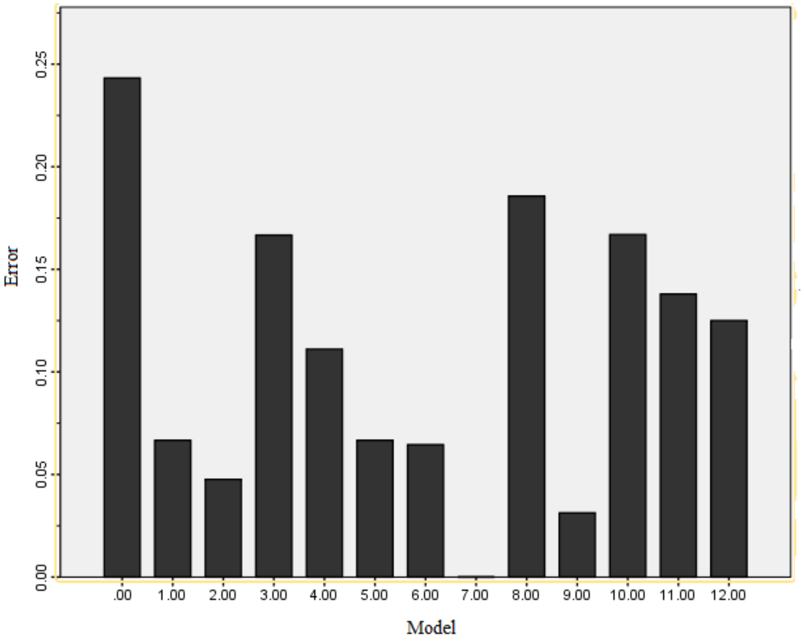


Figure 6-14: Ensemble decision trees error curve.

A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate and diverse. An accurate classifier is one that has an error rate of better than random guessing on new data. The classifiers are diverse if they make different errors on new data points. Figure 6-14 shows uncorrelated errors for the ensemble of decision trees.

The proposed model shows excellent performance for application of diagnosis of heart disease that caused by varieties of error.

7

Conclusions

In this thesis, we have studied PCG signal processing and classification techniques in order to design an intelligent diagnostic system for screening newborns application, with three main objectives: 1) Preprocessing, for removing unwanted noises and preparing signal for signal processing, 2) Feature extraction and selection, based on signal processing and data mining tools that used as input and training process for the neural network, 3) Designing an intelligent system using machine learning and data mining tools to distinguish heart sound with normal and pathological murmurs.

Chapter 1 and 2 contain background material deemed relevant to better understand the content of this thesis. In chapter 1, firstly, the basic concept heart physiology, anatomy and types of heart diseases were presented and heart diseases in newborn or congenital heart defects (CHD) modalities which were considered throughout the thesis were introduced. Then, a theoretical overview of heart valves disease, most common heart defect in newborns such as Patent Ductus Arteriosus (PDA) was provided. The research presented in this thesis finds its application in diagnosing heart sound normal (innocent) and pathological murmurs. After that, a short review of these two medical case studies, underlying the potential of signal processing and classification techniques in their predication and diagnosis was considered.

In chapter 3, several preprocessing steps are described, aimed at improving the quality of data and to facilitate accurate classification. A filter is designed to remove unwanted noises and a down sampling technique is applied to reducing the number of sampling frequency. The detection of cardiac sounds and the definition of systole and diastole are the first steps towards automating the analysis of cardiac acoustic signals. We introduced various methods for manually and automatically detecting and segmentation of cardiac cycles in heart sound signal. A novel automatic segmentation algorithm is proposed to detecting first (S1) and second (S2) heart sound. Also, selecting the best cardiac cycles to extracting feature is presented in this chapter.

Chapter 4, investigates different techniques for feature extraction and selection aimed at improving classification performance. The features, which represent the classification information contained in the signals, are used as inputs to the classifiers. Time-frequency analysis has been applied on PCG signals. This part consists of two major steps: feature extraction and feature selection. In the former step we extracted several features including Maximum value amplitude, Peak to Peak, Energy Shannon, Bispectrum and Wigner bispectrum. The latter step (i.e.feature selection) was aimed at reducing the size of the feature vector. In particular, we used gains and importance variable in CART to measure the score each of variable.

In chapter 5, different classification algorithms considered throughout the studies presented in this thesis were over viewed. multi layer perceptrons, decision trees, classification and regression trees and ensemble of decision trees were introduced. The high performance is archived by ensemble of decision trees which is use bagging technique to combine classifiers.

This thesis introduces novelties in both segmentation of heart sound and application of classification. We demonstrated that CART and a suitable data encoding have significant potential for classifying heart sound data as innocent or pathological murmurs in newborns. Given an unknown heart sound, the system outputs its classification. The corresponding support system has shown high discriminant capability on both type-I and type-II errors, thus becoming a good candidate for giving help to doctors in the activity of monitoring newborns at health care centers.

We expect this system to be very useful for a doctor to decide whether a newborn should be sent for proper treatment or not. The proposed technology is intended for high-volume screening of newborns suspected of having a heart disease. The software system proposed in this work can be considered as a diagnostic tool for the first release able to support physicians in their diagnostic task or health care support system in telehealth care or mobile health as a diagnostic system.

Bibliography

[1] Christer Ahlstrom, Peter Hult, Peter Rask, Jan-Erik Karlsson, Eva Nylander, Ulf Dahlström, and Per Ask. Feature extraction for systolic heart murmur classification. *Annals of biomedical engineering*, 34(11):1666–1677, 2006.

[2] Ömer Akgün and H Selçuk Varol. Determining the degree of aortic stenosis caused by the bicuspid valve with bispectral analysis of heart sound signals. In *5th International Advanced Technologies Symposium, Karabuk, Turkey*, 2009.

[3] Hussnain Ali, TJ Ahmed, and K Shoab. Heart sound signal modeling and segmentation based on improved shannon energy envelogram using adaptive windows. In *Asialink International Conference on Biomedical Engineering & Technology*.

[4] Amir Mohammad Amiri and Giuliano Armano. Detection and diagnosis of heart defects in newborns using cart. *Journal of Life Sciences and Technologies Vol*, 1(2):103–107, 2013.

[5] Amir Mohammad Amiri and Giuliano Armano. Diagnosis and classification of systolic murmur in newborns. In *The 10th IASTED International Conference on Biomedical Engineering*, pages 480–484. IASTED, 2013.

[6] Amir Mohammad Amiri and Giuliano Armano. Early diagnosis of heart disease using classification and regression trees. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–4. IEEE, 2013.

[7] Amir Mohammad Amiri and Giuliano Armano. Heart sound analysis for diagnosis of heart diseases in newborns. *APCBEE Procedia, journal under Elsevier*, 7(4):109–116, 2013.

[8] Amir Mohammad Amiri and Giuliano Armano. An intelligent diagnostic system for congenital heart defects. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 4(7):16–22, 2013.

[9] Amir Mohammad Amiri and Giuliano Armano. Segmentation and feature extraction of heart murmurs in newborns. *Journal of Life Sciences and Technologies Vol*, 1(2), 2013.

[10] Amir Mohammad Amiri and Giuliano Armano. A decision support system to diagnose heart diseases in newborns. In *The 3rd International Conference on HEALTH SCIENCE and BIOMEDICAL SYSTEMS (HSBS '14)*, pages 16–21. NANU, 2014.

[11] Prabhu Babu, Erik Gudmundson, and Petre Stoica. Automatic cepstrum-based smoothing of the periodogram via cross-validation. In *16th European Signal Processing Conference (EUSIPCO 2008), Switzerland*, volume 70, 2008.

[12] A Kumar Banerjee, Neelima Arora, and USN Murty. Classification and regression tree (cart) analysis for deriving variable importance of parameters influencing average flexibility of camk kinase family. *Electronic Journal of Biology*, 4(1):27–33, 2008.

[13] Robert E Banfield, Lawrence O Hall, Kevin W Bowyer, and W Philip Kegelmeyer. A new ensemble diversity measure applied to thinning ensembles. In *Multiple Classifier Systems*, pages 306–316. Springer, 2003.

[14] Boualem Boashash. Time-frequency signal analysis. 1991.

[15] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[16] C Sidney Burrus, Ramesh A Gopinath, Haitao Guo, Jan E Odegard, and Ivan W Selesnick. *Introduction to wavelets and wavelet transforms: a primer*, volume 23. Prentice hall New Jersey, 1998.

[17] Robert Callan. *Essence of neural networks*. Prentice Hall PTR, 1998.

[18] Leon Cohen. Generalized phase-space distribution functions. *Journal of Mathematical Physics*, 7(5):781–786, 1966.

[19] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 41(7):909–996, 1988.

[20] Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *Information Theory, IEEE Transactions on*, 36(5):961–1005, 1990.

[21] SM Debbal and F Bereksi-Reguig. Time-frequency analysis of the first and the second heartbeat sounds. *Applied Mathematics and Computation*, 184(2):1041–1052, 2007.

[22] SM Debbal and Fethi Bereksi-Reguig. Computerized heart sounds analysis. *Computers in biology and medicine*, 38(2):263–280, 2008.

[23] Curt G DeGroff, Sanjay Bhatikar, Jean Hertzberg, Robin Shandas, Lilliam Valdes-Cruz, and Roop L Mahajan. Artificial neural network–based method of screening heart murmurs in children. *Circulation*, 103(22):2711–2716, 2001.

[24] Richard L Donnerstein. Continuous spectral analysis of heart murmurs for evaluating stenotic cardiac lesions. *The American journal of cardiology*, 64(10):625–630, 1989.

[25] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons,, 1999.

[26] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*, volume 57. CRC press, 1994.

[27] Floriana Esposito, Donato Malerba, Giovanni Semeraro, and J Kay. A comparative analysis of methods for pruning decision trees. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(5):476–491, 1997.

[28] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, 1996.

[29] Javier R Fonoliosa and Chrysostomos L Nikias. Wigner higher order moment spectra: definition, properties, computation and application to transient signal analysis. *Signal Processing, IEEE Transactions on*, 41(1):245, 1993.

[30] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[31] Neil L Gerr. Introducing a third-order wigner distribution. *Proceedings of the IEEE*, 76(3):290–292, 1988.

[32] Michael Goebel and Le Gruenwald. A survey of data mining and knowledge discovery software tools. *ACM SIGKDD Explorations Newsletter*, 1(1):20–33, 1999.

[33] Erik Gudmundson, Niclas Sandgren, and Petre Stoica. Automatic smoothing of periodograms. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 3, pages III–III. IEEE, 2006.

[34] Cota Navin Gupta, Ramaswamy Palaniappan, Sundaram Swaminathan, and Shankar M Krishnan. Neural network classification of homomorphic segmented heart sounds. *Applied Soft Computing*, 7(1):286–297, 2007.

[35] Sandro AP Haddad and Wouter Serdijn. *Ultra low-power biomedical signal processing: an analog wavelet filter approach for pacemakers.* Springer Science & Business Media, 2009.

[36] Ibrahim R Hanna and Mark E Silverman. A history of cardiac auscultation and some of its contributors. *The American journal of cardiology*, 90(3):259–267, 2002.

[37] Anil K Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4–37, 2000.

[38] Nick Kingsbury. Complex wavelets for shift invariant analysis and filtering of signals. *Applied and computational harmonic analysis*, 10(3):234–253, 2001.

[39] David Kriesel. A brief introduction to neural networks. *Retrieved August*, 15:2011, 2007.

[40] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms.* John Wiley & Sons, 2004.

[41] Ludmila I Kuncheva, Christopher J Whitaker, Catherine A Shipp, and Robert PW Duin. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications*, 6(1):22–31, 2003.

[42] Louisa Lam and Ching Y Suen. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 27(5):553–568, 1997.

[43] Mark Last, Oded Maimon, and Einat Minkov. Improving stability of decision trees. *International Journal of Pattern Recognition and Artificial Intelligence*, 16(02):145–159, 2002.

[44] Aimé Lay-Ekuakille, Patrizia Vergallo, Diego Caratelli, Francesco Conversano, Sergio Casciaro, and Antonio Trabacca. Multispectrum approach in quantitative eeg: accuracy and physical effort. *Sensors Journal, IEEE*, 13(9):3331–3340, 2013.

[45] Zhixiong Li, Xinping Yan, Chengqing Yuan, and Zhongxiao Peng. Intelligent fault diagnosis method for marine diesel engines using instantaneous angular speed. *Journal of Mechanical Science and Technology*, 26(8):2413–2423, 2012.

[46] DR LIANG. Heart murmur in a child. *Hospital Physician*, page 27, 2004.

[47] Wei-Yin Loh and Nunta Vanichsetakul. Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association*, 83(403):715–725, 1988.

[48] George F Luger. *Artificial intelligence: structures and strategies for complex problem solving*. Pearson education, 2005.

[49] Oded Z Maimon and Lior Rokach. *Data mining and knowledge discovery handbook*, volume 1. Springer, 2005.

[50] Ofer Matan. On voting ensembles of classifiers. In *Proceedings of AAAI-96 workshop on integrating multiple learned models*, pages 84–88. Citeseer, 1996.

[51] Michael E McConnell and Alan Branigan. *Pediatric Heart Sounds*. Springer, 2008.

[52] John Mingers. An empirical comparison of pruning methods for decision tree induction. *Machine learning*, 4(2):227–243, 1989.

[53] Ali Moukadem, Alain Dieterlen, and Christian Brandt. Phonocardiogram signal processing module for auto-diagnosis and telemedicine applications. 2012.

[54] Zoran Nenadic and Joel W Burdick. Spike detection using the continuous wavelet transform. *Biomedical Engineering, IEEE Transactions on*, 52(1):74–87, 2005.

[55] Chrysostomos L Nikias. Higher-order spectral analysis. In *Engineering in Medicine and Biology Society, 1993. Proceedings of the 15th Annual International Conference of the IEEE*, pages 319–319. IEEE, 1993.

[56] Leonardo Noriega. Multilayer perceptron tutorial. *School of Computing. Staffordshire University*, 2005.

[57] M Rabinoff, CMR Kitchen, IA Cook, and AF Leuchter. Evaluation of quantitative eeg by classification and regression trees to characterize responders to antidepressant and placebo treatment. *The open medical informatics journal*, 5:1, 2011.

[58] Norhashimah Mohd Saad, Abdul Rahim Abdullah, and Yin Fen Low. Detection of heart blocks in ecg signals by spectrum and time-frequency analysis. In *Research and Development, 2006. SCOReD 2006. 4th Student Conference on*, pages 61–65. IEEE, 2006.

[59] Milan Šamánek, Zdeněk Slavík, Božena Zbořilová, Věra Hroboňová, Marie Voříšková, and Jan Škovránek. Prevalence, treatment, and outcome of heart disease in live-born children: a prospective analysis of 91,823 live-born children. *Pediatric cardiology*, 10(4):205–211, 1989.

[60] Daniele Paolo Scarpazza. A brief introduction to the wigner distribution. *Dipartimento di Elettronica e Informazione, Politecnico di Milano*, 2003.

[61] Ljubisa Stankovic and Srdjan Stankovic. An analysis of instantaneous frequency representation using time-frequency distributions-generalized wigner distribution. *Signal Processing, IEEE Transactions on*, 43(2):549–552, 1995.

[62] Christos Stergiou and Dimitrios Siganos. Neural networks. 1996, 2010.

[63] P Sandgren Stoica et al. Smoothed nonparametric spectral estimation via cepsturm thresholding-introduction of a method for smoothed nonparametric spectral estimation. *Signal Processing*, 2006.

[64] Petre Stoica and Niclas Sandgren. Total-variance reduction via thresholding: Application to cepstral analysis. *Signal Processing, IEEE Transactions on*, 55(1):66–72, 2007.

[65] Spencer L Strunic, Fernando Rios-Gutiérrez, Rocío Alba-Flores, Glenn Nordehn, and Stanley Burns. Detection and classification of cardiac murmurs using seg-

mentation techniques and artificial neural networks. In *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on*, pages 397–404. IEEE, 2007.

[66] Morton E Tavel and Hart Katz. Usefulness of a new sound spectral averaging technique to distinguish an innocent systolic murmur from that of aortic stenosis. *The American journal of cardiology*, 95(7):902–904, 2005.

[67] Julius T Tou. Engineering principles of pattern recognition. In *Advances in Information Systems Science*, pages 173–249. Springer, 1969.

[68] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[69] John C Wood, Andrew J Buda, and Daniel T Barry. Time-frequency transforms: a new approach to first heart sound frequency dynamics. *Biomedical Engineering, IEEE Transactions on*, 39(7):730–740, 1992.

[70] Heping Zhang and Burton Singer. *Recursive partitioning and applications*. Springer Science & Business Media, 2010.

List of Publications Related to the Thesis

Published papers

- Amir Mohammad Amiri and Giuliano Armano. Heart sound analysis for diagnosis of heart diseases in newborns. *APCBEE Procedia, journal under Elsevier*, 7(4):109–116, 2013

- Amir Mohammad Amiri and Giuliano Armano. Early diagnosis of heart disease using classification and regression trees. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–4. IEEE, 2013

- Amir Mohammad Amiri and Giuliano Armano. Segmentation and feature extraction of heart murmurs in newborns. *Journal of Life Sciences and Technologies Vol*, 1(2), 2013

- Amir Mohammad Amiri and Giuliano Armano. Diagnosis and classification of systolic murmur in newborns. In *The 10th IASTED International Conference on Biomedical Engineering*, pages 480–484. IASTED, 2013

- Amir Mohammad Amiri and Giuliano Armano. A decision support system to diagnose heart diseases in newborns. In *The 3rd International Conference on HEALTH SCIENCE and BIOMEDICAL SYSTEMS (HSBS '14)*, pages 16–21. NANU, 2014

- Amir Mohammad Amiri and Giuliano Armano. An intelligent diagnostic system for congenital heart defects. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 4(7):16–22, 2013

- Amir Mohammad Amiri and Giuliano Armano. Detection and diagnosis of heart defects in newborns using cart. *Journal of Life Sciences and Technologies Vol*, 1(2):103–107, 2013