NTNU

Norwegian University of
Science and Technology

Machine Learning of Circulatory Oscillations in Cardiac Surgical Patients

Mathias Falk

Master of Science in Electronics
Submission date: September 2016
Supervisor: Ilangko Balasingham, IET

Norwegian University of Science and Technology
Department of Electronics and Telecommunications

Preface

This Master's thesis is submitted to the Norwegian University of Science and Technology (NTNU) in partial fulfillment of the requirements of the degree Masters of Science (MSc). The work has been carried out in the period January 2016 to August 2016 at the Department of Electronics and Telecommunications (IET) under the supervision of Prof. Ilangko Balasingham (IET) and Nils Kristian Skjærvold, Postdoctoral Fellow (ISB).

The assignment was written in collaboration with St. Olav's University Hospital, contributing to ongoing research on the human circulatory system. Frequency analysis of circulatory oscillations and subsequent pattern recognition have been conduction in order to analyze multiple biomedical signals from cardiac patients both before and after heart surgery.

Biomedical Engineering has been a long interest of mine, and a major influence for choosing Signal Processing as my field of study. During this work I have learned much about medical signals and the human body, as well as *machine learning*. This topic I knew very little of, but is now something I would very much like to work with in the future.

This thesis is intended for readers with a basic understanding of calculus, linear algebra, probability theory and signal processin.

Mathias Falk
Trondheim, August 2016

Acknowledgment

I would first like to thank my supervisor Ilangko Balasingham for the opportunity to work in the interesting field of biomedical signal processing, and for guidance throughout the process of the assignment.

I would also like to thank my co-supervisor Nils Kristian Skjærvold for formulating the assignment, and also providing useful medical research and explanations.

A big thank you to Seyyed Hamed Fouladi for using more time than you needed to explain and help me understand the different signal processing methods used in my analysis. And also to Younghak Shin for your participation.

Also thanks to medical students Bjørn Gardsjord Lio og Fredrik Axelsson for producing the medical signals, the detailed description of the procedure and equipment, and help with medical research. And thanks to my fellow Master's student Kristian Stenerud for cooperation in the early process helping each other out.

Finally I would like my parents for always being there, supporting me through all these years of studying, and especially these last months of struggle.

Abstract

The human circulatory system is a complex organ system, with multiple feedback and feedforward mechanism for regulation. When global circulatory variables are assessed, these exhibit distinct oscillatory patterns with several frequency bands with different amplitudes, probably because of the underlying complex regulation. Loss of complexity, or decomplexification, has been reported in several cardiac diseases, as well as after heart surgery. Electrocardiogram (ECG), arterial blood pressure (ABP) and laser Doppler flowmetry (LDF) of peripheral blood flow has been recorded in ten patients subjected to coronary artery bypass grafting (CABG) surgery. As such, cardiovascular oscillations have been obtained in observations pre and post surgery, divided into different classes. Heart rate variability (HRV) signals have been extracted from the ECG signals. Low frequency content in the band 0.004 - 2 Hz has been computed in the HRV, ABP and LDF signals with the continuous wavelet transform (CWT). Principal component analysis (PCA) has been performed of the CWT data, to investigate changes in the circulatory system due to the CABG surgery. This showed a small distinction between pre and post surgery observations. The frequency band for each signal has been further divided into six smaller subbands. Total power in each subband has been computed. This showed a statistical significantly decrease in power from before to after surgery in the HRV signals, for the five lower subbands. Also, the power loss was statistically significant in some of the subbands for the ABP signal. Further, five more features have been computed for each subband. These new features, as well as the principal component (PC) scores, for each observation, have been further analyzed with the ReliefF feature selection method. From this, the dimensionality was reduced to only contain the important features showing distinction between pre and post surgery data. Using these features, supervised classification was performed with the decision tree, discriminant analysis, K-nearest neighbor (KNN) and support vector machine (SVM) classifiers. Performance measures have been assessed with the K-fold cross validation method. Correct classification accuracy of the different classes for before and after surgery was relatively high, however performance was not perfect. This was probably due to the low number of observations, as more is data usually needed to properly train and test classifiers.

Sammendrag

Sirkulasjonssystemet i mennesket er et kompleks system, med flere bakover- og foroverkoblinger for å regulere systemet. Inspeksjon av globale variabler fra sirkulasjonssystemet viser at disse inneholder distinkte mønstere med ulik amplitude i ulike frekvensbånd. Dette skyldes mest sannsynlig underliggende mekanismer som styrer reguleringen av systemet. Tap av kompleksistet har blitt rapportert for flere hjertesykdommer, og også etter hjerteoperasjon. Elektrokardiogram (EKG), blodtrykk og blodgjennomstrømming i huden har blitt målt i ti pasienter som har gjennomført koronar bypassoperasjon (CABG). Fra dette så har oscillasjoner fra sirkulasjonssystemet blitt målt både før og etter operasjon, delt inn i ulike klasser. Et nytt signal for hjertrytmevariasjon (HRV) har blitt funnet fra EKG signaler. Lavfrekvent infromasjon i båndet 0,004 - 2 Hz har blitt regnet ut i signalene for HRV, blodtrykk og blodgjennomstrøming, ved bruk av metoden sammenhengende wavelet transformasjon (CWT). Prinsipalkomponentanalyse (PCA) har blitt utført på CWT dataen, for å undersøke endringer i sirkulasjonssystemet forårsaket av CABG operasjonen. Dette viser en liten endring i observasjonene mellom før og etter operasjon. Det nevnte frekvensbåndet har blitt ytterligere delt inn i seks mindre underbånd. Total amplitude for verdiene i hvert underbånd har blitt regnet ut for signalene HRV og blodtrykk. Dette viser en statistisk signifikant reduksjon mellom observasjoner før og etter CABG operasjon, i de fem laveste båndene for HRV signalet. Det er også en statistisk signifikant reduksjon for noen av båndene i blodtrykkssignalet. Videre så har fem flere særtrekk blitt regnet ut i hvert bånd. Alle disse nye særtrekkene, og prisipalkomponentene fra PCA, har blitt analysert med ReliefF algoritmen. Fra dette, så har dimensjonalitetene blitt redusert, ved å finne og beholde kun de beste særtrekkene som viser en betydelig forskjell mellom klassene med data for før og etter operasjon. Ved bruk av disse sætrekkene, så har klassifisering blitt gjennomførst bed bruk av algoritmene beslutningstre, diskriminativ analyse, K-næreste naboer (KNN) og støttevektormaskiner (SVM). Prestasjonsmål for de ulike algoritmene har blitt målt ved K-fold kryssvalidering. Korrekt klassifiseringstreffsikkerhet av de ulike klassene for observasjoner før og etter operasjon er relativ høy, men ikke perfekt. Dette skyldes mest sannsynlig the lave antallet observasjoner, og mer data trengs for å skikkelig kunne trene og teste ulike klassifiserere.

Contents

List of Tables

List of Figures

Abbreviations

ABP Arterial blood pressure

ANS Autonomic nervous system

BP Blood pressure

CABG Coronary artery bypass grafting

CWT Continuous wavelet transform

ECG Electrocardiogram

HR Heart rate

HRV Heart rate variability

KNN K-nearest neighbor

LDA Linear discriminant analysis

LDF Laser Doppler flow(metry)

PC Principal component

PCA Principal component analysis

QDA Quadratic discriminant analysis

SVM Support vector machine

1 Introduction

Machine learning, often used synonymously with *pattern recognition*,*statistical learning* and *data mining*, all generally refers to a set of tools to *model* and understand some data [1–4]. It consists of algorithms with the ability to extract useful information from complex measurements, sometimes seemingly with no apparent structure. These methods can be utilized in several ways, some being constantly improve systems, visually representation, or making future predictions. The last decade machine learning has spread rapidly through a wide variety of industries, solving previously unsolved problems [5]. Today, machine learning is an integral part of many systems, and used in Web search, spam filters, ad placement, credit scoring, fraud detection, stock trading and numerous other applications [6]. This is only expected to increase in the years to come, with novel solutions leading innovations [7]. There is an immense potential in *big data* situations, like *Internet of Things*[1] (IoT). In medicine, biomedical signals can be analyzed to build better health profiles and predictive models to better diagnose and treat diseases [10, 11].

1.1 Motivation

Cardiovascular diseases (CVDs) are the leading mortality cause globally, and annual deaths are only expected to increase in the future [12, 13]. Naturally, a deeper understanding of the *human circulatory system* would be beneficial for treatment of such diseases. This is a physiologically complex organ system, with multiple feedback and forward loops, and the main purpose is to provide the *cells* and its *organelles* with oxygen [14]. When global circulatory variables are assessed, these exhibit distinct oscillatory patterns with several frequency bands with different amplitudes, probably as a consequence of the underlying complex regulation [15, 16]. *Homeostasis* is considered as the active regulation of variables in healthy systems to sustain stability despite external and internal perturbations, and thus keep the complexity of circulatory oscillations [17, 18].

[1] Network of items (buildings, vehicles, medical devices, etc.) embedded with sensors, actuators and Internet connection that can communicate and exchange data [8], also in medical settings. Machine learning can be the brain perform analysis of this vast amount of data leading to embedded intelligence [9].

Loss of irragularity or complexity, i.e. *decomplexification*, has further been found in pathological patients with different cardiac diseases, as well as non-cardiac diseases and fatigue from aging [19–22]. *Hemodynamic* and *physiological* parameters such as *blood pressure, electrocardiogram* (ECG), *blood flow, muscle sympathetic nerve activity*, etc. can provide information about the health status of a patient. If the patient is put through a necessary cardiac surgery, this is a major cardiovascular challenge, and the hemodynamic and physiological parameters can change. The changes can provide important clinical information regarding the individual patient's underlying physiologic regulation and the circulatory system. As for now, there are no existing mathematical tools available in order to access and understand these circulatory dynamics as a function of health condition, age and gender or predict the outcome in a robust manner. A future goal would therefore be to know the *features* reflecting the human circulatory system, and be able to perfectly model all major and latent functions for complete comprehension. Consequently, reliable algorithm can be implemented in applications designed for early and precise detection of cardiovascular anomalies.

1.2 Problem Description

Doctors at St. Olav's University Hospital in Trondheim have performed a systematic clinical study on several cardiac surgical patients, where high resolution digital data have been obtained. The data are transcribed with timestamps for predefined periods; before surgery, right after surgery, after *extubation* and more than 12 hours after surgery.

The signal processing methods so far used in the literature have relied on frequency methods like *wavelet transforms*. The objective of this master thesis is to extend the frequency based methods with *feature extraction* and *classification* to study the signals and their behavior/dynamics from extremely low frequencies to just above the heart rate. Based on the study we will use statistical methods to analyze/predict certain patterns and compare the aforementioned classes using methods like SVM, PCA, etc.

1.3 Background

As mentioned, low frequency circulatory oscillations show steady power in some frequency bands accompanied with peaks in other bands. These are the rhythmic and nonrhythmic oscillations regulating the complex circulatory system [14]. This has been reported in ECG, *heart rate variability* (HRV), blood pressure (BP) and peripheral blood flow computed with *Laser Doppler flowmetry* (Laser Doppler flow (LDF) signal) [15, 16]. The distinct spectral peaks have been studied to identify the underlying sources. In circulatory setting, high frequencies in ECG, BP and LDF with a peak around 1 Hz has been verified to represent the *heart rate* (HR), while the peak in the subband 0.2-0.4 Hz is due to *respiratory function*. The HRV signal can be computed from the ECG signal, showing variation of HR introduced by the *autonomic nervous system* (ANS) [23]. The peak around 0.15-0.4 Hz is believed to come from *parasympathetic* (vagal) activity. Together with BP and LDF, lower frequency observations include peaks around 0.1 Hz,

called *Mayer waves*, with origin still disputed [14]. The peaks have been recorded with forced sympathetic nervous activation [24, 25], however other studies have not shown a link [26]. These peaks have previously been linked to blood pressure regulation and *myogenic* (muscle) response, with peaks around 0.04 Hz suggesting *neurogenic* activity [15]. Very low frequencies with a peak around 0.01 Hz have been seen in HRV and LDF signals, and hypothesized to be *metabolic activity*, but with causation still discussed and not yet clarified [14]. HRV fluctuations, assessed in time and spectral domains, have been shown to decrease after *myocardial infarction* and in several other conditions [23]. This is further reported after *coronary artery bypass grafting* (CABG) surgery, but with gradually return to preoperative values after six months. Patients with loss of HRV complexity due to CAGB surgery has an increased mortality rate when compared with normal HRV patients [27], albeit not when fluctuations are normalized within a few months [23]. This reduction might be caused by a combination of *anatomical manipulation*, *anaesthesia* during surgery, *cardioplegia* and *extracorporeal circulation. skrive om ldf og abp her hva som skjer før og etter operason?*

The circulatory oscillations in blood flow signals was computed using the *continuous wavelet transform* (CWT) with a *Morlet* wavelet in [15]. Both CWT and the *discrete wavelet transform* (DWT) were used as feature extraction of ECG signals in [28] and HRV signals in [29, 30]. DWT was also used for feature extraction of arterial blood pressure (ABP) signals in [31, 32]. Wavelet analysis has also been used as feature extraction in combination with classification schemes. CWT was used for ECG in [33] and *Surface Electromyography* (SEMG) signals in [34], and DWT for ECG in [35], all performing classification with *artificial neural networks* (ANN). DWT was used on multivariate 8 and 12-lead ECGs in [36], with different *mother wavelets* and by computing the *variance* and *covariance* of wavelets coefficients for *dimensionality reduction*. Classification between healthy subjects and a specific heart condition was then done using *linear discriminant analysis* (LDA) and *quadratic discriminant analysis* (QDA). In [37], LDA, *principal component analysis* (PCA) and *independent component analysis* (ICA) were used together with standard *t-test* as dimensionality reduction after calculation DWT of ECGs. Then automated diagnosis of *coronary artery disease* (CAD) affected patients was solved with several classifiers, namely *Support Vector Machine* (SVM), *Gaussian Mixture Model* (GMM), *Probabilistic Neural Network* (PNN) and *K-Nearest Neighbor* (KNN). PCA was also used as feature extraction and dimensionality reduction of large ECG signals in [38–40], and of large breast cancer datasets in [41]. In the latter, PCA was also applied for *data visualization*, followed by *k-means* clustering. The *feature selection filter methods* t-test and *analysis of variance* (ANOVA), among others, for dimensionality reduction, were reviewed in [42]. T-test, *chi-test* and *ReliefF* features filters were compared with a new *Biomarker Identifier* filter for gene expression data identifying potential biomarkers for lung cancer [43]. The ReliefF method was also employed on mammogram images with large dimensionally and low number of observations for feature selection of automatic tumor classification [44]. In [45] ReliefF was used in conjunction with three classification algorithms, specifically the KNN, SVM and naive Bayes.

1.4 Objectives

During the process of the thesis, the problem at hand has been specified, and the following work has been conducted:

- Extraction and noise removal of three biomedical signals; ECG (to compute HRV), ABP and LDF, in four distinct periods; pre surgery, right after surgery, after extubation (removal of the endotracheal tube), and after more than 12 hours, from ten heart surgical patients subjected to CABG.
- Provide low frequency content in the range 0.004-2 Hz of circulatory oscillations in the biomedical signals, using the CWT for graphical interpretations, to help with the research paper by medical students at St. Olav's University Hospital [14] .
- Perform PCA of the CWT coefficients from the biomedical signals for data visualization and possible clustering of different predefined periods or classes.
- Implement *feature engineering* of CWT coefficients, using PCA and feature selection techniques, for dimensionally reduction and extraction of feature subsets containing important information of circulatory changes.
- Use these features for classification of aforementioned classes, employing different classifiers such as *decision tree*, *discriminant analysis*, KNN, and SVM.

1.5 Limitations

The limitations of this study are identified to be the possible errors contained in the biomedical signals, before these are subjected to the subsequent analyses. The ECG *electrodes* had to be removed prior to surgery, and so the placements may not be identical for the pre and post operative recordings, potentially leading to different amplitudes and electrode contact noise. Other types of noise included low frequency motion artifacts because of patient and equipment movement, overlapping in the frequency domain with spectral information of interest making filter removal difficult. Therefore extraction of data from the different periods was performed manually seeking nondistorted samples. The postoperative measurements were obtained within three hours after surgery, within an hour and a half after extubation, and at least 12 hours after surgery, but with minor variations among patients. Variations also included interventions such as fluids, medications etc. These were given by the on-duty medical personnel, as a response to the patients' critical condition. Hence medications that might influence the oscillations, such as *morphine* or the $\alpha - / \beta$-blocker *labetalol*, could affect the final results. Other situations that could have corrupted some signals include one patient going to the toilet during preoperative recording, one patient having a pacemaker-rhythm until extubation, and finally for one patient the transducer fell of the bed during the recording and was as a result not at heart-level for the remaining recording time.

Other limits include choices made throughout the subsequent analysis of the medical signals. The CWT was performed only for the Morlet wavelet, while there exists several other wavelets. New wavelets can also be produced, customized for a particular signal or application. The DWT and the FT can also be used to extract frequency information from oscillating signals. For the subsequent machine learning computations, several other techniques exists beyond those employed in this thesis. Feature extraction was performed only for PCA, and feature selection only for one filter method. For classification, there are numerous other algorithms, all with different assumptions of the data yielding individual results.

1.6 Outline

This thesis is structured into ten chapters. Chapter 1 just presented serves as a short introduction to machine learning, motivation for cardiovascular exploration and presentation of the problem including previous work, specific objectives and detected limitations. Chapter 2 gives some basic medical theory of the human circulatory system and associated biomedical signals. Wavelet analysis is briefly clarified in Chapter 3, with further scrutinize on the CWT used in the analysis. Chapter 4 has a more extensive introduction to machine learning, followed by explanation of related problems of dimensionality reduction and *supervised classification* in Chapter 5 and Chapter 6, respectively. The data collection from patients and subsequent analyses are described in Chapter 6. Chapter 8 presents the results of the analyses, which are then discussed in Chapter 9. Finally the conclusions are written in Chapter 10.

2 Medical Background

This chapter will provide basic insight to the human circulatory system, its components and functions, and related medical signals that are used to monitor the condition of the system.

2.1 Human Circulatory System

The human circulatory system is a complex organ system comprised of the heart pumping blood for circulation through the vessels in the body, providing nutrients for the cells and removing waste products, to help fight diseases, stabilize temperature and pH, and maintain homeostasis [46, 47]. The circulatory system is usually divided into the *lymphatic system* and the *cardiovascular system*. The former in an open system of lymphatic vessels transporting *lymph fluid*, and is an important part of the *immune system*. The cardiovascular system, or simply the vascular system, is a closed system of blood vessels and the heart transporting blood to all types of tissue in the body [48]. A simplified version can be seen in Figure 2.1. The vascular system consists of two loops, the *pulmonary* circulation bringing blood to and from the heart via the lungs, and the *systemic* circulation going to the rest of the body. The blood vessels are categorized into five general types. *Arteries* are strong elastic vessels capable of carrying blood away from the heart under high pressure. These subdivide into thinner *arterioles*. The arterioles further divide into the thinnest *capillaries* vessels that connects the smallest arterioles with the smallest *venules*. These merges into thicker *veins* leading blood back to the heart for another circulation.

Blood is a connective tissue composed of a liquid matrix, or *blood plasma* with *proteins*, smaller *molecules* and *electrolytes* (liquid *ions*), and formed elements of *red blood cells (erythrocyte), white blood cells (leukocytes)* and *platelets (thrombocytes)* [47]. The systemic circulation is divided into two, the *macrocirculation* between organs and the *microcirculation* of the smallest arterioles and capillaries vessels. Due to friction between blood and vessels, the blood pressure decreases throughout the circulation [48]. The arterial end of capillaries exert *hydrostatic pressure* on the surrounding tissue fluid, permitting filtration of oxygen carried by *hemoglobin* protein, the majority of red cells, from the lungs, and nutrients to move in high levels to body cells. Pressure continues to

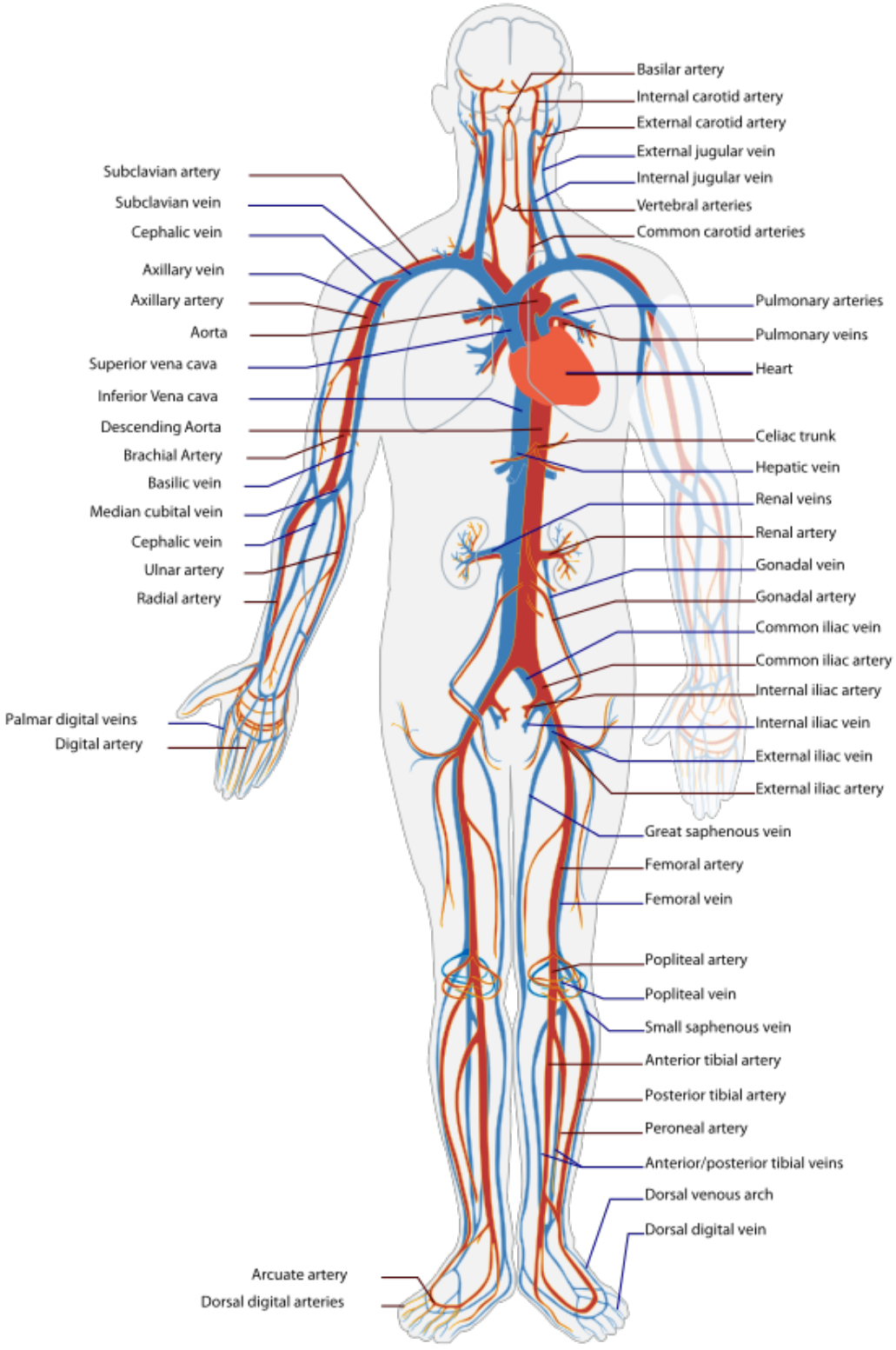
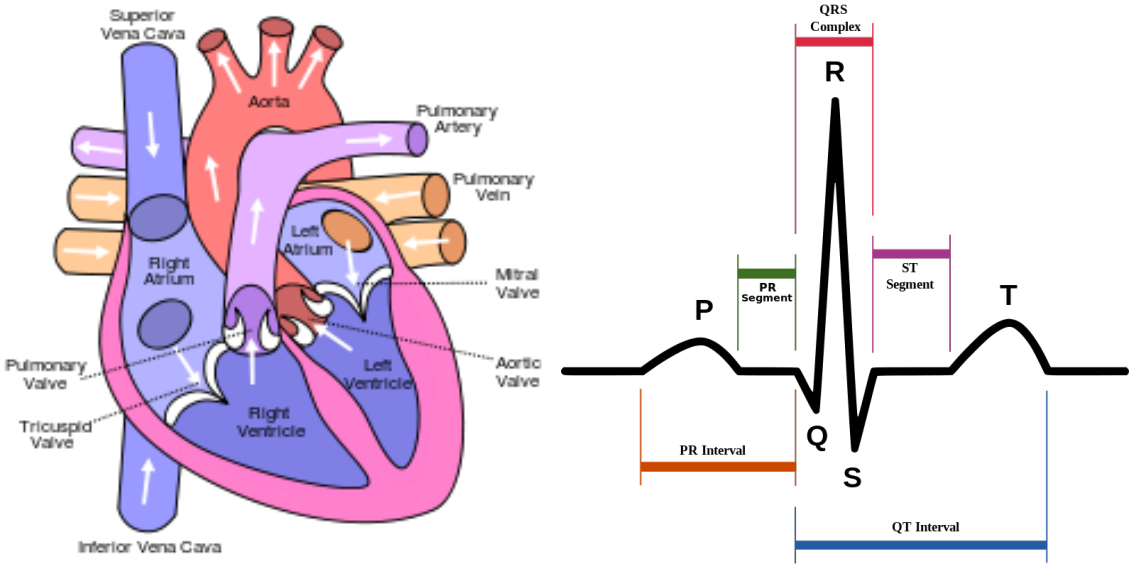


Figure 2.1: The simplified human cardiovascular system. Red indicates oxygenated blood carried in arteries and blue indicates deoxygenated blood carried in veins. Capillaries joining the arteries and veins are omitted. Accessed from [46].

decrease in the venules end, causing *osmotic pressure* inwards which reabsorbs fluid of waste materials. Too large plasma proteins remain in blood through the exchange and transport waste, some of which are filtered in the kidney and liver. Acid bicarbonate form carbon dioxide,CO_2, to maintain pH, and is excreted by the lungs. More fluid leave than return to the capillaries, and this excess is returned to the blood via the lymph system. Other functions of the blood includes *thermoregulation* by redistributing excess heat, *immunity* by providing white blood cells to to sites of injury or invasion by disease-causing agents, and homeostasis using platelets and clotting proteins to minimize blood loss when vessel are damaged [47].

The human heart is located to the left in the *thoracic* cavity of the *thorax*, or chest [48], resting on the *diaphragm*, and the anatomy is displayed in Figure 2.2 (a). The *coronary circulation* refers to blood flow to the *cardiac muscles*. The heart has four major functions, namely collecting the blood from all parts of the body, sending the blood to the lungs, collecting the blood from the lungs, and pumping it back to all parts of the body [49]. Blood low on oxygen enters the *right atrium* from the *inferior* and *superior vena cava*. Here it is passed through the *tricuspid valve* leading blood only one way, and into the *right ventricle*. Then it is further sent through the *pulmonary valve* and with the *pulmonary artery* through the lungs' capillaries receiving oxygen and releasing carbon dioxide, the opposite process of earlier. The *respiratory system* interacts with the circulatory system by infusing blood with inhaled oxygen and then exhale carbon dioxide. The blood returns to the heart through the *pulmonary valve* to the *left atrium*, before passing the *mitral valve* to the *left ventricle* and leaving through the *aortic valve* to the *aorta* for another circulation. The electric activity of cells can be summarized as follows. Tissue is built up by cells. They contain liquid water with molecules and ions. *Extracellular liquid* is outside the cell and *intracellular liquid* is inside, separated by a *membrane* allowing only some ions and molecules to pass, hence creating an *active potential*. This is known as the *Sodium-potassium pump*. The membrane has *resistance* and acts a *capacitor*, and the pump can be modelled as an electric circuit. The cells are *depolarized* building up voltage. Ions are passed through the membrane, causing the "circuit" to *repolarize* and energy is produced. For the cardiac muscles, an active potential causes the cardiac cells to contract, which reduces the atrial and ventricular volume, respectively. This results in an increase in pressure, which is higher before the valve than behind, leading it to open and the blood to pass, before closing. This pressure changes to open and close the valves results in the pumping role of the heart. The contraction part of the heart is called the systole, which is followed by a rest period called the diastole. Cardiac muscles, or *myocardio fibers*, differ from other muscle tissue in the body. Unlike other muscle fibers which are controlled by the ANS and excited by the *motor nerves*, cardiac muscles possess *automaticity* by generating its own stimulus [50]. The *nodal cells* of the heart initiate, synchronize and regulate the heart contractions by generating periodic action potentials. They do not have a rest potential, but constant ion leak, eventually depolarizing spontaneously at regular intervals when the cell membrane potential exceeds a certain threshold. After each depolarization, nodal cells produce a repolarization. These are longer than for other muscle tissue. Nodal cells are organized in two general groups, the *sinoatrial* (S-A) node stimulating the atria and

the *atrioventricular* (A-V) node stimulation the ventricles. Since they generate stimuli themselves, they are called the natural *pacemakers* of the heart. Further, cardiac muscle possess *contractility*, that is fibers respnd with maximal efficiency systole no matter the stimuli, where usually the magnitude implies the contraction. The cardiac muscle develops a myogenic contraction, i.e. a self-initiated contraction. Other properties include *excitability*, or the ability to respond to stimuli, and *conductibility* which mean being capable of transmitting the electric impulses.



(a) Heart anatomy (b) ECG sinus

Figure 2.2: Illustration of the human heart anatomy (a) and the normal ECG sinus (b). Accessed from [46, 51].

2.2 Medical Signals

Different signals can be used to assess the functions of the cardiovascular system and measure vital signs. ECG is a group of signals measuring the electric activity of the heart, and constitutes the most informative clinical signal commonly used in the diagnosis of the cardiovascular system [49]. This produces a periodic signal, with each cycle having distinct characteristics. This is referred to as a *sinus rhythm*, with five wave peaks, illustrated for a healthy subject in Figure 2.2 (b). The P wave is caused by depolarization of the atrium which contracts to fill the ventricle with blood. The *QRS complex* shows the depolarization of the ventricles, more precisely the *septum* which is the wall separating the left and right ventricle. The repolarization of the atrium is weak, and obscured by the QRS complex. The *T wave* shows up as the repolarization effect of the ventricles. The repolarization can be distinguished from the depolarization in the cardiac action potential from the fact that the repolarization wave front is significantly longer in time duration than the depolarization wave front. The repolarization also has a much smoother action potential gradient, which incorporates a smaller gradient

in the time derivative of the cell membrane potential. Sometimes a *U peak* can also appear, representing a portion of the ventricular repolarization. The depolarization and repolarization process during each heart cycle generates local electric potential differences, which can be measured *noninvasive* on the exterior of the body, using electronic recording equipment on the skin rather than the heart directly. Various internationally accepted electrode placements are in existence. The voltage difference between electrodes can be computed, and is referred to as *leads*. The ECG measures heart cycles of relaxation and contraction of the cardiac muscle, and thus gives the pulse or heart rate. This is sensitive to internal and external stimuli that can cause decrease or increase. Two different mechanisms can be distinguished; intrinsic due to changes in the S-node such as stretching or temperature, and extrinsic from ANS regulation. Various deviations in the typical heart rhythm are often caused by either impulse generation malfunctioning or conduction distortion. These deviations from the normal normal functionality of the cardiovascular system can be linked to specific pathological conditions, either genetic or due to malfunctions such as infections, lack of oxygen or obstruction of vessels supplying blood supply to the heart itself [49].

BP determination is one of the most important measurements in all of clinical medicine [52], and is defined as the force that blood exerts against the inner walls of blood vessels throughout the vascular system [48], with thicker vessels composed of multiple wall layers. As such it is used study hemodynamics, i.e. the fluid dynamics of blood flow, and gives primary information about the performance of the cardiovascular system [53]. More commonly it refers to pressure on the arteries (ABP) in the systemic circulation supplied by the *aortic branches* from the aorta [48]. Pressure rise and fall according to the cardiac cycle, with maximum systolic pressure during the ventricular contraction, and minimum diastolic pressure during the in-between rest period. The blood vessels distend as blood is pumped from the ventricles, but contact after, which produce a pulse which can be felt in an artery near the surface of the skin. Smooth artery and arteriole muscles are innervated by the sympathetic nervous system. *Vasomotor fibers* receive impulses to stimulate blood pressure, by either contract and reduce vessel diameter called *vasoconstriction*, or relax and increase diameter called *vasodilation*. Together with blood vessel friction, this determines the the arterial resistance and is known as *peripheral resistance* (PR), which again affects blood pressure. *Stroke volume* (SV) is defined as the volume of blood discharged from the ventricle with each contractions. *Cardiac output* (CO) is the volume discharged from the ventricle per minute, or SV multiplied with HR [47]. BP can be calculated by multiplying CO and PR, and normal arterial pressure is maintained by regulating these two factors. Abnormally high and low blood pressure is referred to as *hypertension* and *hypotension*, respectively. Measuring BP can be executed both *noninvasive* and *invasive*. There are numerous noninvasive approaches [52]. These are without complication and less painful for the patients, but may yield lower accuracy than invasive measurements. Most commonly a *sphygmomanometer* is used, where a cuff is placed on the upper arm occluding the *brachial* artery, and then inflated, manually or automatic, to above the systolic pressure and gradually deflated. Variables of the procedure include placement of the cuff, size of the cuff, which arm, body position, pretest situation, etc. The manual *auscultatory*

method uses a stethoscope to listen on the arterial on the elbow just below the cuff. Other manual methods use *mercury* where blood pressure affect the height of a column of mercury, or *aneroid* where pressure is registered by a mechanical system of metal bellows that expands as the cuff pressure increases and a series of levers that register the pressure on a circular scale. The digital methods use oscillometric measurements and electronic calculations rather than auscultation. BP is reported as systolic pressure over diastolic pressure, and measured in millimeters of mercury (mmHG), eg. 120/80 mmHg [47]. Invasive direct *intra-arterial* pressure measurement is usually more accurate, but with higher possibility of complications, and reserved to critically ill patients undergoing high-risk and major [53]. This allows for beat-to-beat measurement and gives a continuous signal, rich in pathological information such as heart rate, systolic, mean and diastolic pressure, important aspects in cardiology [31]. The measurement procedure will be explained in greater detail in Section 7.1.

HRV is a physiological signal indicating the influence of the ANS on the heart rate [23]. It is the variation of the heart rate, measured by the variation in the time between heartbeats, i.e. beat-to-beat interval. As such, the frequencies are lower than that of the heart rate around 1 Hz. Moreover, as already mentioned in Section 1.3, the higher respiratory frequencies indicate parasympathetic activity and lower frequencies indicate sympathetic activity. The power ratio called *sympathovagal balance* can thus be detected by HRV, and changes will further show imbalance. It has been shown to be a predictor of mortality after myocadinal infarction and CABG. The HRV signal can be extracted from other signal, such as ECG and BP. The peaks of either the R wave in the QRS complex or the systolic pressure are good measures representing one heartbeat, and the time interval between successive beats can easily be found from the time between either of these peaks. Hence, HRV is often named RR variability from the R interval measurements of ECG.

LDF is a noninvasive hemodynamic measure of the blood perfusion in the microcirculation [54]. Currently, LDF does not given an absolute measure of blood perfusion. This limits its use in clinical setting, however, LDF have been used in much research work. The laser light used must be *monochromatic*, using a single frequency. When light is scattered by moving red blood cells it will be frequency shifted depending on the movement of the object, the direction of the incoming light and the direction of the scattered light. This is the *Doppler effect*. The frequency shifts caused by Doppler scattering will then result in a frequency broadening of the originally monochromatic light. The backscatter can be detected to obtain the current of different frequencies, or the *optical Doppler spectrum*. This can further be linked to the *Doppler power spectrum*. *Electric fields*, E-fields, with a variety of frequencies can be simulated, and the power spectral density of the detector current generated by these E-fields can be related to the frequency distribution of the E-fields. Finally the concentration of moving red blood cells (CMBC) and the perfusion can be estimated from the Doppler power spectrum. LDF contains some practical considerations. Measurement volume is difficult to define, as it is affected by the tissue optical properties, and changes with changes of the desired quantity. It has a very small sampling volume, and thus sensitive to *heterogeneities*

or spatial variation in blood perfusion and optical properties of the tissue. Also, the perfusion signal can change a lot if the measurement site is changed, and therefore microscopic blood vessels are usually measured. LDF assumes scattering objects other than red blood cells to be static for tissue at rest, but this is not entirely true as *biological zero* means constant motion due to thermal energy. Moreover, since the motion of scattering objects rather than blood perfusion itself is measured, it is important blood perfusion actually dominates the perfusion signal. As such, LDF is sensitive to tissue movement causing motion artifacts. Finally the time-varying part of the *photocurrent* is bandpass filtered in order to reduce noise, but the filter must be selected carefully to not remove important information. More mathematical details of LDF can be obtained in [54].

3 Wavelet Analysis

This chapter aims to give an introduction of *Wavelet Analysis*, with a focused explanation of the continuous wavelet transform, a mathematical tool providing a time-frequency representation of a signal with very good resolution of both properties.

3.1 Development from Fourier Transform

The *Fourier Transform* (FT) is another mathematical tool, and one of the most widely used, for transforming a signal from the time domain to the frequency domain [55]. It is also reversible, and possible to convert back to the original time domain with an inverse version of the transform function. The transform and its inverse are defined as

$$S(\nu) = \mathcal{F}[s(t)](\nu) = \int_{-\infty}^{\infty} s(t) \mathrm{e}^{-i2\pi\nu t} \mathrm{d}t \tag{3.1}$$

$$s(t) = \mathcal{F}^{-1}[S(\nu)](t) = \int_{-\infty}^{\infty} S(\nu) \mathrm{e}^{i2\pi\nu t} \mathrm{d}\nu. \tag{3.2}$$

A time signal $s(t)$ can be expressed as a combination of sinusoids $\mathrm{e}^{i\omega t}, \omega = 2\pi\nu$, both sines and cosines (sine with phase $\pi/2$), as seen in (3.2). The frequency spectrum in found by (3.1) which gives the amplitude for each frequency that converges with the signal. However, the lack of time localization of power variations led to the development trying to overcome this issue [56]. Using the so called *short-time* Fourier Transform (STFT), a generalization of the Gabor transform with Gabor atoms/functions [57], the problem can be partly solved by sliding a fixed length window $w(t)$ introducing temporal changes in spectral response [58]. Further improvement was then achieved in the work on seismic wave analysis, by using a ripple or short wave, i.e. a wavelet, instead of sinusoids, together with varying window lengths [59–61]. Later this was formalized in the context of quantum physics, and known contemporary as CWT. Wavelet analysis is the collection of independent discoveries from various disciplines, previously deemed disconnected, as well as subsequent collaborations and advancements. Besides the aforementioned development, previous work includes study in geophysics, the reaction of the ear to sound, Brownian motion and harmonic analysis. The now first known use of wavelets was Haar sequences, a set of rectangular basis functions from the work

on orthogonal systems, a special case of compact orthonormal basis functions today known as the discrete wavelet transform. Wavelet analysis has since been further developed, containing several subcategories suitable for different problems. The framework of wavelet analysis has unified many results with a mathematical foundation, also containing existing signal processing applications as *subband coding* and *quadratic mirror filter* (QMF), the latter being the *state-of-the-art* tool for DWT computation. As seen from the development, wavelet is used in a many areas, also including medical signals [62] and pattern recognition.

3.2 Continuous Wavelet Transform

The CWT of a real one-dimensional signal $s(t)$ producing a two-dimensional time-frequency space is defined by

$$S(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \psi^*\big(\frac{t-b}{a}\big) s(t) \mathrm{d}t, \tag{3.3}$$

where ψ^* denotes the complex conjugate of the *mother wavelet* ψ on the open (b,a) half plane, $b \in \mathbb{R}, a > 0$ [56]. This is the basic complex prototype function which is dilated and translated to from a set of basis functions that the signal can be decomposed of, as the sinusoids of the Fourier transform. The dilation creates short duration, high frequency, and long duration, low frequency, functions clearly suitable to represent nonstationary signals with power variations [28], which again are localized with the translation. When the analyzed signal correlate well with the dilated and translated mother wavelet, this is represented by high power in the time-frequency domain called a *scalogram*. The shifting and scaling of $\psi(t)$, with parameters b and a respectively, can be written as

$$\psi_{a,b}(t) = a^{-1/2} \psi\big(\frac{t-b}{a}\big) \tag{3.4}$$

and thus the transformation formula can be simplified to

$$S(b,a) = \int_{-\infty}^{\infty} \psi^*\big(\frac{t-b}{a}\big) s(t) \mathrm{d}t \tag{3.5}$$

The process is illustrated in Figure 3.1 (a) including a random signal. However, this is shown for DWT for simplicity. For CWT, the mother wavelet is dilated and translated in a more continuous matter, resulting in redundant information but with graphical visualization as a suitable application. Digressing from CWT, DWT on the other hand uses a specific set of basis functions dilated from the mother wavelet with special properties. The wavelets are not overlapped when shifted, resulting in nonredundant information and excellent preservation of energy in the original signal, a desired property in *lossy compression*. Back to the CWT, given the *admissibility condition*

$$c_\psi = \int_{-\infty}^{\infty} \frac{|\Psi(\nu)|^2}{|\nu|} \mathrm{d}\nu < \infty \tag{3.6}$$

where $\Psi(\nu)$ is the FT of $\psi(t)$, then the wavelet transform is invertible from the following intervention

$$s(t) = \frac{1}{c_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S(b,a) \psi_{a,b}(t) \frac{\mathrm{d}a\mathrm{d}b}{a^2} \tag{3.7}$$

This implies the DC component must vanish, so for smooth $\Psi(\nu)$ then $\Psi(0) = 0$. Consequently, continuous mother functions $\psi(t)$ of interest should, as oppose to long sines and cosines, be bandpass filters of short oscillatory waves that decay sufficiently fast to provide good time resolution, and thus have finite energy fulfilling

$$\int_{-\infty}^{\infty} |\psi(t)|^2 \mathrm{d}t < \infty. \tag{3.8}$$

The Morlet wavelet is a popular analytic complex wavelet used for ECG signals in continuous wavelet analysis [62]. Another popular CWT wavelet is the *Mexican hat*, both shown in Figure 3.2 (a) and (b). The corresponding FT of the wavelets are seen in Figure 3.2 (c) and (d), respectively. The Morlet wavelet is given by an exponential carrier multiplied with a slightly shifted Gaussian window giving the envelope, and defined in time and frequency, respectively

$$\psi_(t) = (\mathrm{e}^{-i2\pi\nu_0 t} - \mathrm{e}^{2\pi^2\nu_0^2\sigma^2})\mathrm{e}^{-\frac{t^2}{2\sigma^2}} \tag{3.9}$$

$$\Psi(\nu) = \sqrt{2\pi\sigma^2}\left(\mathrm{e}^{2\pi^2\sigma^2(\nu-\nu_0)^2} - \mathrm{e}^{-2\pi^2\sigma^2\nu^2}\mathrm{e}^{-2\pi^2\sigma^2\nu_0^2}\right) \tag{3.10}$$

with $\Psi(0) = 0$. For a discrete signal $s(n)$ digitalized at ΔT samples per second, $t = n\Delta T$, the CWT using Morlet wavelet is [28]

$$S(b,a) = \frac{\Delta T}{\sigma\sqrt{2\pi a}} \sum_{n=-\infty}^{\infty} s(n)\mathrm{e}^{-\frac{(n\Delta T-b)^2}{2a^2\sigma^2}} \mathrm{e}^{-i2\pi\nu_0\frac{n\Delta T}{a}} \tag{3.11}$$

The frequency band can be divided into octaves, i.e. a doubling of frequency, which again are partitioned into voices $\mathcal{V}$, so that scale $a_0 = 2^{1/\mathcal{V}}$. With M being the product of number of octaves and number of voices, then scales are given by $a_k = 2^{k/\mathcal{V}}, 1 \le k \le M$. The continuous translation parameter b is also discretize to $u, -\infty \le u \le \infty$, so that

$$S(l,k) = \frac{\Delta T}{\sigma\sqrt{2\pi 2^{k/\mathcal{V}}}} \sum_{n=-\infty}^{\infty} s(n)\mathrm{e}^{-\frac{(n-1)^2}{2^{l+(2k/\mathcal{V})}\nu_s^2\sigma^2}} \mathrm{e}^{-i2\pi\frac{\nu_0}{\nu_s}2^{-k/\mathcal{V}}(n-l)} \tag{3.12}$$

where sampling time $\Delta T = 1/\nu_s$. Dilation of the mother wavelet $\psi(t)$ divides the time-frequency cell into resolution cells, seen in Figure 3.1 (b). For the STFT these are of equal since, whereas they change for wavelet analysis depending on scaling parameter a. The relationship between scaling a and frequency ν can be obtained by calculation the exact response to a sinusoid. Thus combining (3.3) and (3.9), and using $s(t) = \mathrm{e}^{i2\pi\nu t}$, then gives

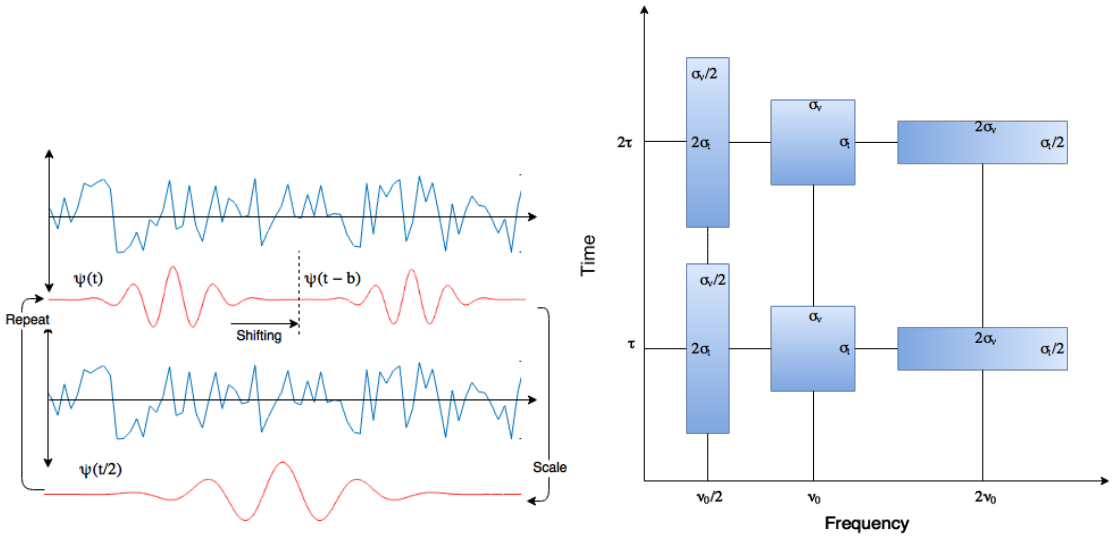
$$S(a,b) = \sqrt{a}\mathrm{e}^{i2\pi b\nu}\mathrm{e}^{-2\pi^2a^2\sigma^2\left(\nu-\frac{\nu_0}{a}\right)^2}, \tag{3.13}$$

peaking at $a = \nu_0/\nu$. This means low scales result in high frequency, and high scales result in low frequency. This is general for all wavelets, since the FT of mother wavelets

are $1/\sqrt{a}\psi(t/a) = \sqrt{a}\Psi(av)$, and so expansion in time equals contraction in frequency and vice versa. The resolution cells are centered at frequency

$$\nu = \frac{\nu_0}{a}, \tag{3.14}$$

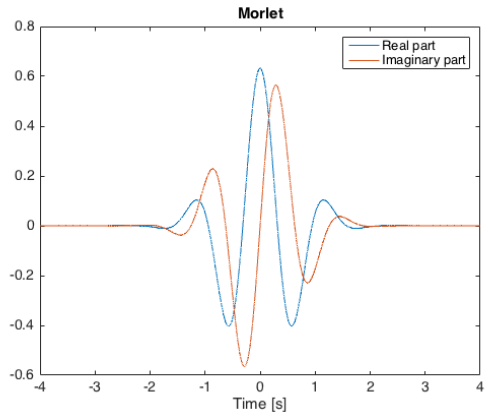
which is the frequency-scale relationship. The original mother wavelet has time and frequency widths given by standard deviation spread, respectively σ_t and σ_ν. Dilating with scale a results in $a\sigma_t$ and σ_ν/a. The reason behind this is due to the uncertainty principle of signal processing, stated mathematically $\sigma_t\sigma_\nu \geq 1/4\pi$. This means resolution of time and frequency will negatively affect each other because of the issue with instantaneous frequency. The lower bound and equal resolution for time and frequency simultaneously is obtained with a Gaussian window, used in the Gabor transform. Increasing the spread of the wavelet gives good frequency resolution but poorer time resolution. Decreasing the spread, using a small window will give good time localization, but at the expense of frequency resolution. The σ parameter for Morlet wavelet is the undilated scale 1 time width of the mother wavelet, used to control the time-frequency resolution. From the preceding response, it is inversely proportional to the mother wavelet dilation in frequency domain, $1/\pi a\sigma$. So for low scales (high frequency), resolution is improved for large values of σ, and conversely low values improves low frequency resolution. Finally the initial frequency ν_0 of the wavelet is chosen to ensure the band-pass property of the mother wavelet. In practice these can be achieved by different means. Either $\nu_0 = 0.5\nu_s$, or $2\pi\nu_0$ is set so that the ratio of highest maximum of $\psi(t)$ to the second highest is 0.5, $\omega_0 = (2/\ln(2))^{0.5} \approx 5.3364$, so usually $\nu_0 = 5/2\pi$ [15, 56, 62].



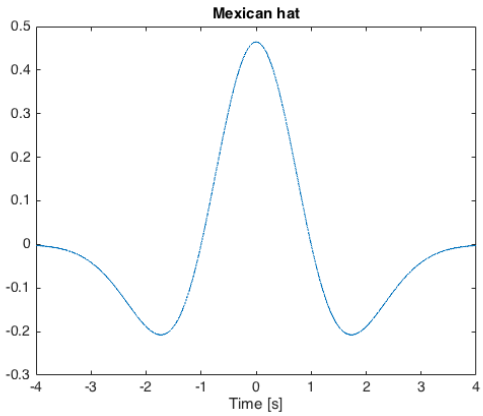
(a) Scaling and shifting

(b) Resolution cells

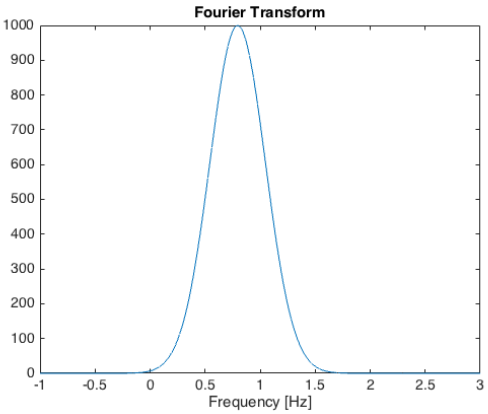
Figure 3.1: The scaling and shifting operation of DWT (a) and the variable length resolution cells of CWT (b), the latter reproduced from [28].



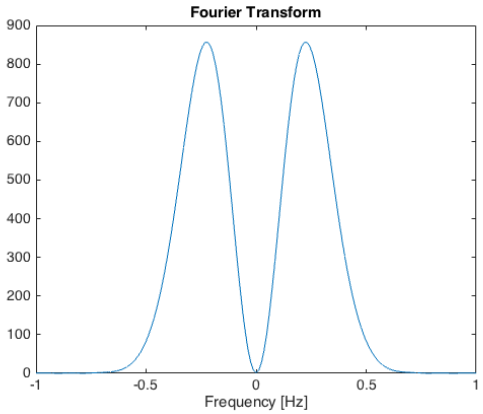
(a) Morlet



(b) Mexican hat



(c) FT of Morlet



(d) FT of Mexican hat

Figure 3.2: The time domain plot of the faster Morlet (complex) (a) and slower Mexican hat (b) oscillating wavelets with scale 1 and ($\omega_0 = 5$). The corresponding FT of the Morlet (c) and Mexican hat (b) wavelets.

4 Machine Learning

This chapter will serve as an introduction to machine learning and define some fundamental probability theory which will be used in the following chapters.

4.1 Framework

Artificial intelligence (AI) is related to philosophy and psychology in wanting to understand *intelligence*, but takes one step further attempting to *build* intelligent entities [63]. The goal is to mimic human reasoning and flexibility seeking optimal choices based on observations. It differs from plain algorithms in a calculator or definitions in an online dictionary, where explicit information from humans are stored in memory. Instead, intelligence is displayed by the ability to learn from this disposable knowledge and make actions based on this. As touch upon in the introduction to this thesis, machine learning refers to a computers ability to learn from observations. It is a subfield of computer science [3] encompassing the study of algorithms for different learning problems, many of which are associated with AI [64]. As famously quoted by Arthur Samuel, *"machine learning is the field of study that gives computers the ability to learn without being explicitly programmed."* [65]. It is closely related to and mostly overlaps with pattern recognition growing out from engineering, both heavily employing statistical theory. As such they both intersect further with statistical learning, itself growing out of classical statistics with a computational approach [2]. Other terms include data mining [4] and *deep learning*, but the latter refers to using several hidden layers in artificial neural networks which is a specific area of machine learning inspired by biological neural networks.

Machine learning and the aforementioned interchangeable fields apply computational sciences to producing a set of tools which overall goal is to model and understand some complexity in measured datasets. It is this process that is referred to as *learning from data* [1], and it can be divided into different settings, *supervised* and *unsupervised* learning, still with intermediate situations of *reinforced* learning [64, 66]. Supervised learning or *predictive modelling* refers to the situations where there for an observation is a number of measurement inputs, $x_i, i = 1, 2, ..., m$, and for each observation there is an associated known labeled output y_i [2]. The goal is to produce a model that relates

the inputs to the output, and then be able to accurately predict future outcome for new input observations, e.g. classification and regression analysis. Thus inputs are often divided into a *training set* for model building and a *test set* for prediction accuracy evaluation. Conversely, unsupervised learning has no known labeled associated output, and the goal for the model is to understand underlying patterns and relationship between a set of input measurements, e.g. cluster analysis. The inputs in both these learning settings are known interchangeably by various terms, including just *measurements*, *predictors*, *attributes* and *independent variables* [2]. In pattern recognition literature inputs are also known as features [1], a word defined as "a distinctive attribute or aspect of something." The outputs in supervised learning is also known as *responses*, *targets* and *dependent variables*.

The computed model is the mathematical explanation of a system given by measurement vectors $\mathbf{x_i}$ from several observations, $\mathbf{X}$. Imagine the true model is given by the function f. The hypothesis is that a function h can be found from a hypotheses set $\mathcal{H}$, and by learning from a large and representative training set $\mathcal{T}$, h will be a good approximation of f [64]. This is known as inductive learning, as oppose to deduction [6]. Instead of starting with a theory and hypotheses, seeking a guaranteed conclusion through many observations, the observations are used to find an hypothesis and build a theory from that, which is not always true. Correlation does not imply causation, and validity depends on data being either *experimental* from some manipulation or plain observations obtain without any control. But induction have the possibility to learn from far less knowledge. Still, data in not enough as some assumptions are needed. The methods calculating a model, also called *learners* or *inducers*, could be either *parametric* or *nonparametric* [2]. The former has strong assumptions about the form or shape of the function f, and then solves for the fixed number of parameters defining f. This makes such methods stable and computationally simpler, but biased with a constraint on $\mathcal{H}$ and the number of functions possible to estimate [1]. The latter is more unbiased, with weaker and not explicit assumptions about f. As such, these methods require more data to fit the model, making them unstable and computationally demanding with a flexible and increasing number of parameters.

4.2 Probability Theory

Machine learning builds upon statistics, which study collection, analysis, interpretation and presentation in order to understand some data. Probability theory again is the foundation to statistics. The selected equations and theory of probability presented in this section are taken form [67–70], and can help to deal with uncertainty and are thus used in the machine learning methods explained in the following chapters.

A categorical random variable Y with values $y_1, y_2, \ldots, y_n$ can be grouped in classes $\mathcal{Y}_c, c = 1, 2, \ldots, C$. The probability of a class $\Pr(\mathcal{Y}_c) = \pi_c$[1] can be estimated by the experi-

[1] Not to be confused with the mathematical constant π.

mental or empirical probability

$$\hat{\pi}_c = \frac{n_c}{n}, \tag{4.1}$$

and thus a frequency of occurrence ratio (relative frequency) where n_c are the number of values y_j of class $\mathcal{Y}_c$ and n is total number y_j values. Given another random variable X, the sum rule says the marginal probability of X is given by the summing out the other variable Y, $\Pr(X) = \sum_Y \Pr(X, Y)$. This can be factored by conditioning on X and from the product rule $\Pr(X, Y) = \Pr(Y|X)\Pr(X)$, which is symmetric. By using this property, *Bayes' theorem* yield

$$\Pr(Y|X) = \frac{\Pr(X|Y)\Pr(Y}{\Pr(X)}. \tag{4.2}$$

If X is a numeric variables with values $x_1, x_2, \ldots, x_n$ and corresponding probabilities $p_1, p_2, \ldots, p_n$, the lower order first raw moment statistic or population *mean* is given by the expected value

$$\mu_X = \mathrm{E}[X] = \mathrm{E}_X = \sum x \Pr(X = x) = \sum_j x_j p_j. \tag{4.3}$$

An estimate of the population mean can be given by the unbiased[2] sample or empirical mean

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} x_j. \tag{4.4}$$

The second central moment or population *variance* of X is the expected value of the squared deviation from the mean given by

$$\sigma_X^2 = \mathrm{var}(X) = \mathrm{E}[(X - \mu_X)^2], \tag{4.5}$$

and is a measure for the spread of data. The estimate can be given by the unbiased[3] sample variance

$$s^2 = \mathrm{var}(X) = \frac{1}{n-1}\sum_{j=1}^{n} (x_j - \bar{X})^2, \tag{4.6}$$

The *standard deviation* is the squared root of the variance

$$\sigma_X = \sqrt{\mathrm{E}[(X - \mu_X)^2]}, \tag{4.7}$$

and thus the average distance from the mean to all the data values of the variable, and the unbiased estimate is the squared root of the unbiased sample variance

$$s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n} (x_j - \bar{X})^2}. \tag{4.8}$$

[2]A statistic is said to be an unbiased estimate of a given parameter when the mean of the sampling distribution of that statistic can be shown to be equal to the parameter being estimated.

[3]Normalization $\frac{1}{n}$ for variance is biased for small n

This is also a measure of the spread, and always non-negative, with low values signify data points close to the mean and high values signify data points far from the mean and each other. How two random variables X_1 and X_2 change together or *correlate*[4] can be given by the *covariance*

$$\sigma^2_{X_1X_2} = \text{cov}(X_1, X_2) = \text{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})] \tag{4.9}$$

where $\text{cov}(X_1, X_2) = \text{cov}(X_2, X_1)$ and $\text{cov}(X, X) = \text{var}(X)$, i.e. variance of a variable is thus the covariance of itself. Positive value indicates variables change similar together, negative value means opposite behaviour, while zero covariance is uncorrelation. From the definition of sample variance the sample covariance is

$$\text{cov}(X_1, X_2) = \frac{1}{n-1}\sum_{j=1}^{n}(x_{1_j} - \bar{X}_1)(x_{2_j} - \bar{X}_2). \tag{4.10}$$

For a vector of random variables $\mathbf{x} = X_1, X_2, \ldots, X_m$ the covariance for each combination is given by the *covariance matrix*

$$\begin{aligned}\Sigma &= \text{cov}(X_i, X_j) = \text{E}[(X_i - \mu_i)(X_j - \mu_j)^T] \\ &= \begin{bmatrix} \text{E}[(X_1-\mu_1)(X_1-\mu_1)] & \text{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \text{E}[(X_1-\mu_1)(X_m-\mu_m)] \\ \text{E}[(X_2-\mu_2)(X_1-\mu_1)] & \text{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \text{E}[(X_2-\mu_2)(X_m-\mu_m)] \\ \vdots & \vdots & \ddots & \vdots \\ \text{E}[(X_m-\mu_m)(X_1-\mu_1)] & \text{E}[(X_m-\mu_m)(X_2-\mu_2)] & \cdots & \text{E}[(X_m-\mu_m)(X_m-\mu_m)] \end{bmatrix}.\end{aligned} \tag{4.11}$$

If a random variable X is normal or Gaussian distributed, it is given by the (univariate) probability density function

$$\mathcal{N}(X|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \text{e}^{\frac{(X-\mu)^2}{2\sigma^2}}. \tag{4.12}$$

This can be generalized to the multivariate normal distribution for an m-dimensional vector $\mathbf{x}$ of normally distributed variables X, given by

$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \text{e}^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}, \tag{4.13}$$

where μ is an m-dimensional mean vector, Σ is an $m \times m$ covariance matrix and $|\Sigma|$ is the determinant of the covariance matrix. Another distribution is the *Chi-squared* distribution, given by

$$\mathcal{X}^2(X|F) = \frac{(\frac{X}{F})^{\frac{F}{2}-1}\text{e}^{-\frac{X}{2}}}{2\Gamma(\frac{F}{2})}, \tag{4.14}$$

[4] Correlation is a score of statistically dependency between two random variables (or sets). Pearson's product-moment (population) correlation coefficient for two random variables X_1 and X_2 is given by $\rho_{X_1,X_2} = \text{corr}(X_1, X_2) = \frac{\text{cov}(X_1,X_2)}{\sigma_{X_1}\sigma_{X_2}}$ [71]

where $X \geq 0$, $\Gamma(F) = (F-1)!$ and F is the degree of freedom.

Statistics also includes higher order moments. *Skewness* is the third order standardized moment, and Pearson's moment coefficient of skewness is given by

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mathrm{E}[(\mathrm{X}-\mu)^3]}{(\mathrm{E}[(X-\mu)^2])^{3/2}}. \tag{4.15}$$

It is a measure of the asymmetry of the probability distribution or the mass concentration of the distribution, either positive to the left with longer right tail of the distribution curve, or negative with mass to the right with longer left trail. The sample skewness estimate is given by

$$b_1 = \frac{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{X})^3}{\left(\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{X})^2\right)^{3/2}}. \tag{4.16}$$

Kurtosis is the forth order standardized moment, defined by

$$\gamma_2 = \frac{\mathrm{E}[(\mathrm{X}-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2}, \tag{4.17}$$

and a measure of the *peakiness* of a probability distribution. *Leptokurtic* distribution refers to positive kurtosis and are normally more peaked than the normal distribution, while *platykurtic* refers to negative kurtosis and flatter distributions. *esokurtic* means kurtosis around zero. The sample excess kurtosis is given by

$$g_2 = \frac{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{X})^4}{\left(\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{X})^2\right)^2} - 3. \tag{4.18}$$

Lastly the *Central Limit Theorem* (CLT) states that a set of independent random variates, each with arbitrary probability distribution with a given mean and a finite variance, tends to a normal probability density function for large sample size. Thus data which are influenced by many small and unrelated random effects are approximated to normal distribution as sample size increase.

5 Dimensionality Reduction

This chapter will illustrate the concept of features and the impact of dimensionality for the corresponding feature space from data. Then follows different associated manipulations like principal component analysis and feature filter methods used for dimensionality reduction, a specific problem in the field of machine learning.

5.1 Feature Processing

The measurements of data can be obtain and manipulated in different ways, before potentially being utilized in an algorithm to explaining a certain pattern or relationship. In fact, feature preprocessing plays a major part and is key to successful modelling [6]. For instance, when collecting data from patients, one can use features such as age, weight, blood pressure, smoker or non smoker etc. All these measurements of one patient can then be called a m-dimensional *feature vector*, and put together with data from other patients to form a *feature set* or matrix. In other words several observations of some given features are collected to make an m-dimensional input or feature space $X \in \mathbb{R}^m$, and an algorithm can learn from these examples. The feature vector can also be all or an excerpt of samples in a finite digital signal, e.g. ECG. Treating this as the *raw data* first measured, it can be further processed making new feature sets to either optimize a learning algorithm or introduce new understanding. This feature processing, sometimes also called *feature extraction* as a whole [72, 73], can be categorized in multiple ways, with analogous names. *Feature transformation* is the process of creating new features from the original features, while *feature selection* or *subset selection* is choosing a subset of features from the original features [74]. Feature transformation can be further divided into *feature construction*, also called *feature generation* [73], and *feature extraction* with a more narrow definition. The former refers to creating new features where the dimensionality of the feature set could either increase or decrease, but the latter is strictly a dimensionality reduction process. By definition, feature selection is also solely a dimensionality reduction process. Regardless of category, all feature processing methods can be described mathematically by $\mathbf{X}^* = g(\mathbf{X})$, where the function s can be any transformation, either linear or nonlinear, or selection with an attach optimization criterion, from the original feature set $\mathbf{X}$ to a new feature set $\mathbf{X}^*$ [1, 75].

Always when measuring some data or signal, it can be given by

$$signal = source + noise, \tag{5.1}$$

meaning that there is potentially an underlying source, pattern or relationship in the measured signal contaminated by some noise, random values or outliers not fitting this pattern. The optimal feature set should therefore be computed to represent the source. Often the measured signal is high dimensional, e.g. when obtaining a continuous signal of some kind, and a multiple, indirect measurement of a source which cannot be quantified directly with the given sensors or equipment [76]. Thus this noise should be removed by unveiling a lower dimensional manifold explaining the source. For instance, in an unsupervised setting, there might be a global variation but with noisy local variations. In a supervised setting, the noise would be data that distorts the separability of *a priori* classes. Also the noise can be redundant information, and a intrinsic dimensionality or subspace of lower dimensionality contains the variability of high dimensional original measure data due to correlation between input variables [77]

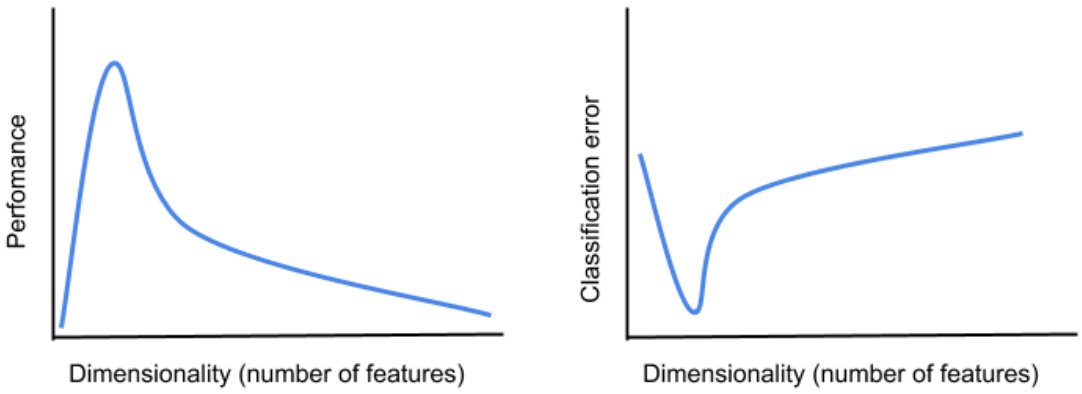
Feature processing methods for dimensionally reduction can be used in different ways, including [72, 75, 76]

- Data compression: strictly preserving the original variability of a signal in a lower dimensionality.
- Data visualization: Possible graphical understanding of large datasets in lower dimensionality, with detection of outliers or removal of noise.
- Data preprocessing: An integral part of supervised modelling, employed as a mean to obtain an optimal smaller feature set for increased computational speed and accurate prediction, i.e. *generalization*, by avoiding the *curse of dimensionality* and *overfitting*.

5.2 Curse of Dimensionality and Overfitting

The curse of dimensionality refers to difficulties that can occur in high dimensionality, and how algorithms can turn intractable while working well in lower dimensions [78]. In statistical learning this can best be visualised by Figure 5.1, illustrating how the learning performance increases, and contrarily classification (or learning) error decreases, as the number of features, i.e. dimensionality, increases. By introducing new features to understand and further scrutinize some data, the performance of a learning algorithm will usually improve, but only to a certain point, before eventually adding more features simply distorts the problem. Hence, there is an optimal dimensionality and corresponding feature set.

The intuition from a three-dimensional world is that the more features at disposal, the easier it will be to learn and distinguish patterns in some data [6]. However, this

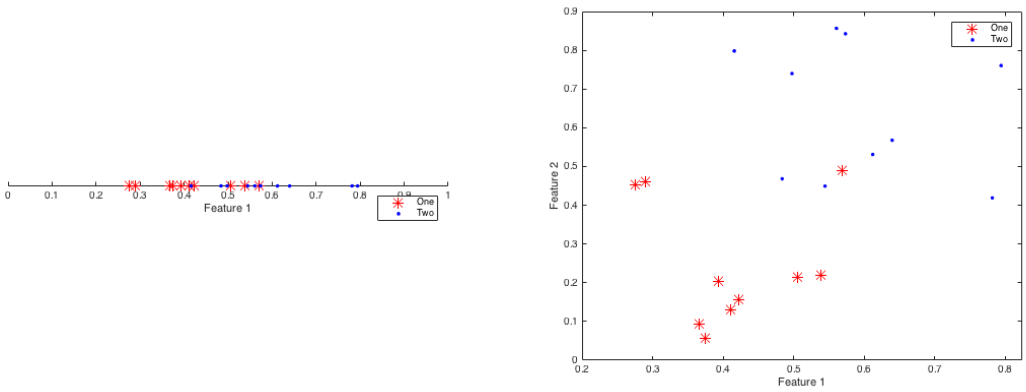


(a) Performance

(b) Classification error

Figure 5.1: Illustration of performance (a) and classification error (b) as a function of dimensionality (number of features). Reproduced from [79].

becomes increasingly untrue for higher dimensionality. Starting with one arbitrary feature, the data could be random with no system, seen partly in the example data in Figure 5.2 (a) [79]. The data tends to have either high or low values, but there are overlap between the two arbitrary classes “One” and “Two”, and not possible to differentiate with a threshold value *classifier*. Adding another feature, the sparsity increases and clusters start to appear, as seen in the feature plane (b) in the same figure, showing the distinction between the two classes.



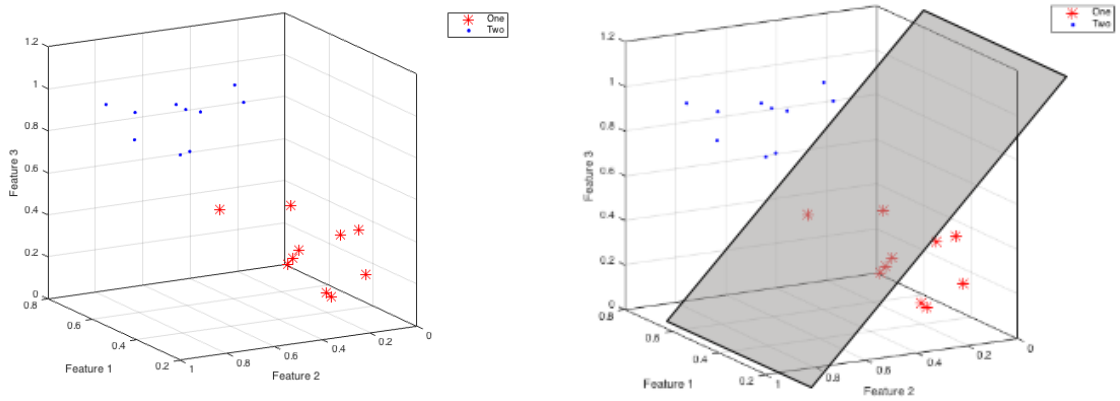
(a) 1-dimensional feature line.

(b) 2-dimensional feature plane.

Figure 5.2: Example data showing feature line (a) versus feature plane (b). Reproduced from [79].

Introducing yet another feature, the two classes are separated in two non-overlapping clusters, seen in Figure 5.3 (a). This problem can thus be perfectly classified with a

hyperplane[1], i.e. a two-dimensional plane in the three-dimensional feature space, as illustrated in (b).



(a) 3-dimensional feature space

(b) 3-dimensional feature space with a hyperplane

Figure 5.3: Example data showing feature space without (a) and with (b) a classification hyperplane. Reproduced from [79].

But, appending new features increases the dimensionality of the feature vector, and with fixed-size observations exponentially decreases the density of the input space [79]. Thus observations need to grow exponentially when adding features in order to avoid the curse of dimensionality. If n observations suffice to reflect a one-dimensional feature space of unit interval size, then n^2 observations are needed for two dimensions, i.e. n^m observations for m features.

Also, when increasing the feature space and consequently the sparsity, the observations are not distributed uniformly [3, 6, 79]. Imagine a unit circle inscribed in a unit square, with observations uniformly distributed inside this square and thus the circle as well. Increasing the dimensionality m of these geometrical shapes, creating hypersphere and hypercubes respectively, the volume of the hypercube is always $1^m = 1$, whereas the volume of the hypersphere with radius 0.5 (unit circle) can be given by

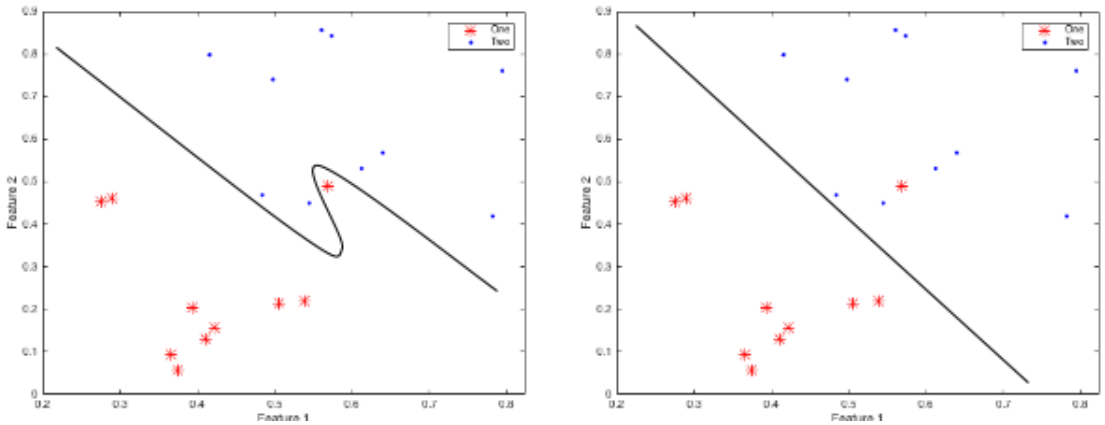
$$V(m) = \frac{\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2}+1)} 0.5^m, \tag{5.2}$$

where $V(m) \to 0$ as $m \to \infty$. Thus observations will tend to lie outside the hypersphere in the corner of the hypercube, and furthermore be harder to classify as feature values may differ greatly when potentially being located in opposite corners of the hypercube.

The goal of a classifier is to generalize based on relatively few relevant observations to make correct predictions about uncertain events, but this becomes exponentially

[1]A hyperplane is a subspace of one dimension less then the original vector space [80]. Strictly speaking a hyperplane must pass through origo, whereas an affine set or affine hyperplane need not [1]

harder for higher dimensionality and leads to overfitting as a direct cause of the curse of dimensionality [79]. Classification error and generalization can be measured in both biased and variance, tendencies to learn constantly the same wrong thing and random noise irrelevant to the source, respectively [6]. Figure 5.4 (a) shows that a complicated nonlinear classifier in a two-dimensional feature space works just as well as the the simpler linear hyperplane in three-dimensions in 5.3 (b). So adding the last feature leads to sparsity and learning special instances and loss of generality. However, the simpler linear classifier in two-dimensions in Figure 5.4 (b) could perform better then a nonlinear classifier, because overall it generalizes better for new observations, but this in not always true. Therefore overfitting can occur for both simple and complex algorithms, and for lower dimensions with high observations and higher dimensions with sparse observations. So simplicity, or complexity, does not imply prediction accuracy [6]. In general, nonlinear classifiers not generalizing well should be model based on fewer features, whereas less computational linear models generalizing easily can perform well for higher dimensionality [79].



(a) Nonlinear classifier (b) Linear classifier

Figure 5.4: Example data showing 2-dimensional feature space with non-linear (a) and linear (b) classifiers. Reproduced from [79].

5.3 Principal Component Analysis

Principal Component Analysis (PCA) is closely related to the preceding work of Singular Value Decomposition (SVD), and its earliest form is widely attributed to the independent work of [81] and [82], before being further developed to its modern form, which could be found in [83]. It is also closely related to numerous other methods, among them Factor Analysis (FA), Canonical Correlation Analysis (CCA), Kosambi-Karhunen–Loève transform (KLT) in signal processing or the Hotelling transform in multivariate quality control, and Proper Orthogonal Decomposition (POD) in mechanical engineering [83, 84], as statistical techniques are often "invented" independently in various field. PCA is today used in several areas like agriculture, biology, chemistry, climatology, demogra-

phy, ecology, economics, food research, genetics, geology, meteorology, oceanography, psychology and quality control [83]. It has been called one of the most valuable results from applied linear algebra [85] and a backbone of modern data analysis [86].

PCA is a versatile unsupervised learning *latent variable method* [66], used for *multivariate analysis* and data visualization, and feature extraction as a preprocessing step to supervised learning. Starting out with some high dimensional complex observations, it is a nonparametric method to compress data by reduction of dimensionality and possibly separate a hidden structure of correlated variables from outliers and noise [85, 87]. PCA uses a vector space transform that projects an input space to a linear subspace of lower dimensionality containing only "principal components" that maintain most of the variability of the data.

First, consider a data set of measurements given by

$$\mathbf{X}_{i,j} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix} = \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \cdots \\ \mathbf{x_m} \end{bmatrix} \in \mathbb{R}^{m \times n}, \mathbf{x_i}^T \in \mathbb{R}^n, \tag{5.3}$$

with m rows of variables or features and n columns of observations, also expressed as m row vectors. To reduce the dimensionality, the aim is to find a new a new set of vectors, or a new *basis*[2], that will explain the data in a new optimal way. X can be linearly transformed to another $m \times n$ matrix Z, so that for some $m \times m$ matrix Λ,

$$\mathbf{Z} = \mathbf{WX} = \begin{bmatrix} \mathbf{w_1} \cdot \mathbf{x_1} & \mathbf{w_1} \cdot \mathbf{x_2} & \cdots & \mathbf{w_1} \cdot \mathbf{x_n} \\ \mathbf{w_2} \cdot \mathbf{x_1} & \mathbf{w_2} \cdot \mathbf{x_2} & \cdots & \mathbf{w_2} \cdot \mathbf{x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{w_m} \cdot \mathbf{x_1} & \mathbf{w_m} \cdot \mathbf{x_2} & \cdots & \mathbf{w_m} \cdot \mathbf{x_n} \end{bmatrix}, w_i, x_j \in \mathbb{R}^m \tag{5.4}$$

where X is written as n columns vectors being *projected* to the columns of the new basis W written as m rows vectors, by the standard Euclidean inner dot product[3] of the corresponding vectors [87]. PCA uses variance σ in the original basis to decorrelate and define the new basis of *orthonormal*[4] vectors as directions in which variance is maximized.

There is no absolute scale for noise, instead all noise is evaluated relative to the measurement signal, and a common measure is the *signal-to-noise ratio* (SNR) given by

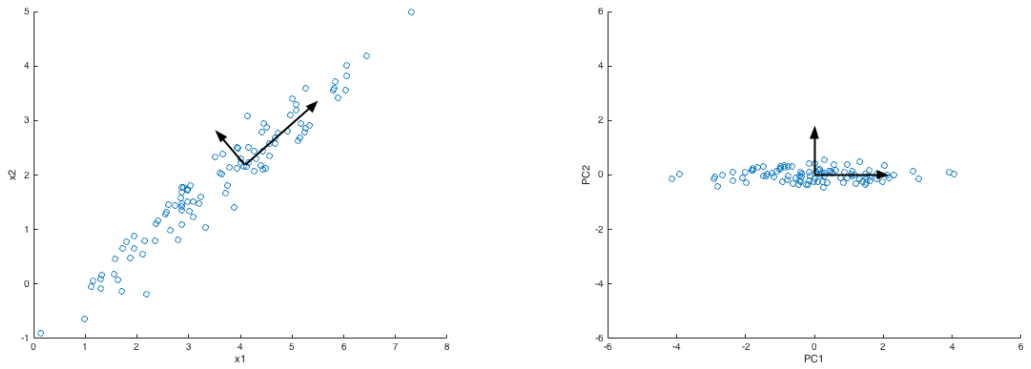
$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}, \tag{5.5}$$

[2] A basis for $\mathbb{R}^m$ is a set of vectors which (1) spans $\mathbb{R}^m$, so any vector in this m-dimensional space can be written as linear combination of these basis vectors, and (2) are linearly independent, i.e non of the basis vectors can be written as linear combination of the others [86].

[3] $[x_1, x_2, \cdots, x_n] \cdot [w_1, w_2, \cdots, w_n] = x_1 w_1 + x_2 w_2 + \cdots + x_n w_n, \mathbb{R}^m$ [88]

[4] Two vectors are orthogonal, i.e. perpendicular, if they form a right angle, i.e. their inner product is zero. Further they are orthonormal if they are orthogonal and unit vectors (magnitude of 1).

and hence a ratio of variance [85]. Clearly a high score indicates low noise contamination and vice versa. Figure 5.5 (a) shows some example data of two arbitrary features $\mathbf{x_1}$ and $\mathbf{x_2}$ scattered against each other, producing a two-dimensional feature space. Each point representing an intersection between variables of the two features are called a *score*. Included are also the corresponding directions of orthogonal maximum variance, first the largest overall and then secondly perpendicular to the first. This is then geometrically rotated and scales by the new basis W of vectors first principal component (PC1) and second principal component (PC2), and represented in a new coordinate system in Figure 5.5 (b) where the PCs are axes forming a right angle. PCA is thus an m-dimensional ellipsoid fitted to the data with axes of original maximum variance direction. Moreover, the ratio of the direction vectors is by definition the SNR, and is consequently maximized by assuming the direction of maximum variance to be the dynamic of interest. Furthermore, variables that do not not depend on other should be kept. This can be understood by again inspecting Figure 5.5. The two features are linearly dependent, but with some noise in direction of lowest variance orthogonal to maximum global variance. If this noise increased, the scatter/cluster would be "fatter" or "rounder", thus indicating uncorrelation and loss of redundancy between the features. Contrary, if the noise decreased, the scatter would form a line demonstrating dependency, correlation and redundancy, and only one of the features would by needed picturing the very idea of dimension reduction.



(a) Feature space

(b) Principal component space

Figure 5.5: Example data in arbitrary two-dimensional feature space and two-dimensional principal component space (first two principal components)

Back to the computation, data is first *mean centered* by calculating the sample mean given by (4.4) for each feature row vector and then subtract this average from each variable n in the given vector. So, given original measured vectors $\tilde{\mathbf{x}}_\mathbf{1}, \tilde{\mathbf{x}}_\mathbf{2}, \ldots, \tilde{\mathbf{x}}_\mathbf{m}$ the resulting vectors are $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m}$. From (4.10) and the subtraction of mean, covariance of vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ is given by

$$cov(\mathbf{x_1}, \mathbf{x_2}) = \frac{1}{n-1} \mathbf{x_1} \mathbf{x_2}^T, \tag{5.6}$$

and is strictly non-negative [85]. From (5.3) each feature is expressed as a vector $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m}$, and the covariance between all the mean centered dimensions can be calculated with the following covariance matrix

$$\begin{aligned} \text{cov}(\mathbf{X},\mathbf{X}) = C_{\mathbf{XX}} &= \frac{1}{m-1}\mathbf{X}\mathbf{X}^T = \frac{1}{m-1}\begin{bmatrix} \mathbf{x_1}\mathbf{x_1}^T & \mathbf{x_1}\mathbf{x_2}^T & \cdots & \mathbf{x_1}\mathbf{x_n}^T \\ \mathbf{x_2}\mathbf{x_1}^T & \mathbf{x_2}\mathbf{x_2}^T & \cdots & \mathbf{x_2}\mathbf{x_n}^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x_n}\mathbf{x_1}^T & \mathbf{x_n}\mathbf{x_2}^T & \cdots & \mathbf{x_n}\mathbf{x_n}^T \end{bmatrix} \\ &= \begin{bmatrix} var(\mathbf{x_1}) & cov(\mathbf{x_1},\mathbf{x_2}) & \cdots & cov(\mathbf{x_1},\mathbf{x_n}) \\ cov(\mathbf{x_2},\mathbf{x_1}) & var(\mathbf{x_2}) & \cdots & cov(\mathbf{x_2},\mathbf{x_n}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\mathbf{x_n},\mathbf{x_1}) & cov(\mathbf{x_n},\mathbf{x_2}) & \cdots & var(\mathbf{x_n}) \end{bmatrix} \in \mathbb{R}^{n\times n}. \end{aligned} \tag{5.7}$$

As previously discussed, the goal of PCA is to

1. Maximize the signal and reduce SNR, thus maximize variance.
2. Have uncorrelated feature vectors, thus minimize covariance.

Inspecting this covariance matrix, it is clear that variance is located on the diagonal and total variance is the trace[5] of the matrix. Conversely, covariance is located off the diagonal. Ergo, the algebraic goal of PCA is to

1. Maximize the diagonal elements.
2. Minimize the off-diagonal elements.

and thusly seek a diagonal covariance matrix, $C_{\mathbf{ZZ}}$, of the transformed matrix Z from (5.4). Assuming the new basis W is an orthonormal matrix,

$$\begin{aligned} C_{\mathbf{ZZ}} &= \frac{1}{n-1}\mathbf{Z}\mathbf{Z}^T = \frac{1}{n-1}(\mathbf{W}\mathbf{X})(\mathbf{W}\mathbf{X})^T = \frac{1}{n-1}(\mathbf{W}\mathbf{X})(\mathbf{X}^T\mathbf{W}^T) \\ &= \frac{1}{n-1}\mathbf{W}(\mathbf{X}\mathbf{X}^T)\mathbf{W}^T = \frac{1}{n-1}\mathbf{W}\mathbf{D}\mathbf{W}^T, \end{aligned} \tag{5.8}$$

where $\mathbf{D} = \mathbf{X}\mathbf{X}^T = (\mathbf{X}^T)^T\mathbf{X}^T = (\mathbf{X}\mathbf{X}^T)^T$ and thus a symmetric[6] $m \times m$ matrix. By the theorem[7] in linear algebra that symmetric matrices is diagonalized by a matrix of its orthonormal *eigenvectors*[8],

$$\mathbf{D} = \mathbf{E}\Lambda\mathbf{E}^T, \tag{5.9}$$

[5] Sum of all diagonal elements in a square matrix

[6] Proof of theorem that matrices are symmetric if and only if it is orthogonally diagonalizable provided in [85]

[7] Proof of theorem that symmetric matrices are diagonalized by a matrix of its orthonormal eigenvectors provided in [85]

[8] Eigenvector calculation defined in [86]

where E is a $m \times m$ orthonormal matrix of orthonormal eigenvector columns from D, and Λ is a diagonal matrix with corresponding *eigenvalues*[9] λ of D as diagonal entries. Selecting the new basis, the rows of W, to be the linearly independent eigenvectors of D found in the columns of E, $W = E^T$, and by theorem[10] of inverse orthogonal matrices, $E^T = E^-1$, eventually gives

$$\begin{aligned} C_{\mathbf{VV}} &= \frac{1}{n-1}\mathbf{WDW}^T = \frac{1}{n-1}\mathbf{E}^T\mathbf{E}\Lambda\mathbf{E}^T(\mathbf{E}^T)^T = \frac{1}{n-1}\mathbf{E}^T\mathbf{E}\Lambda\mathbf{E}^T\mathbf{E} \\ &= \frac{1}{n-1}\mathbf{E}^{-1}\mathbf{E}\Lambda\mathbf{E}^{-1}\mathbf{E} = \frac{1}{n-1}\Lambda. \end{aligned} \tag{5.10}$$

The eigenvalues, also called *latents*, are the diagonal elements of Λ, which measures the variance. So, the corresponding eigenvectors of this matrix can be arranged in order of explained variance of the original data, and thus the principal component *coefficients* or *loadings* are these eigenvectors in order of prominence.

Matrix D has rank $\rho \leq min(m, n-1)$, i.e. number of eigenvectors (of X), and is limited by either m dimensions or n observations in the $m \times n$ matrix X. If all data can occupy and variance be explained by a subspace of dimensionality $\rho \leq min(m, n-1)$, then $min(m, n-1) - \rho$ orthonormal vectors of zero variance not affecting the final solution can be added to maintain the constraint of orthogonality [85]. Mathematically feature extraction is given by

$$\begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_m} \end{bmatrix} \xrightarrow{g} \begin{bmatrix} \mathbf{z_1} \\ \mathbf{z_2} \\ \vdots \\ \mathbf{z_l} \end{bmatrix}, l \leq m \tag{5.11}$$

where g is any transformation from the original feature set $\mathbf{X}$ to new features in $\mathbf{Z}$. With PCA, the dimensionality is not necessarily reduced. Going back to the example in Figure 5.5, all the variance of two original features were explained by two principal components as well. However, examining the eigenvalues revealed that 98% and 2% variance were explained by direction coefficients of PC1 and PC2, respectively. Thus, most of the original information, defined as variance in PCA, would still be accumulated in only PC1 scores for future visualization or input to supervised learning, and PC2 scores could be omitted for dimension reduction and noise removal. The principal component scores are rows of matrix Z, which is the representation of the original data $\mathbf{X}$ in the new coordinate system from the rotation and scaling of principal component coefficients in W, given by (5.4).

Mean centering is performed on the data to ensure the largest eigenvalue and corresponding first principal component is in the direction of maximum variance [89], and axes of the fitted ellipse are indeed the principal components. The centering can be observed in the example in Figure 5.5 (b). By definition of the covariance, subtracting

[9] Eigenvalue calculation defined in [86]

[10] Proof of theorem that the inverse of an orthogonal matrix is its transpose provided in [85].

mean changes the covariance matrix. Other statistical normalizations for standardization or scaling of the input data can be done depending on knowledge of the data. Feature variables with large variance will dominate, so unit variance scaling can be wielded by dividing each variable in one dimension with the corresponding standard deviation [90]. Feature normalization can improve the result when each feature is of different scaling with unequeal units. Several assumptions are made throughout the PCA computation that can end in non-satisfactory results. Covariance through correlation does not imply redundancy [72]. Also in Figure 5.5 (a) or (b), imagine a second parallel linear cluster. Obviously the clusters would be correlated, but not redundant. PCA tries to decorrelated data, or remove second-order dependencies, whereas this would be higher order dependencies [91]. As a linear transformation, the hidden pattern in the measured data might be of a nonlinear nature and go undetected. A nonlinear transformation can be done prior, turning to a parametric approach termed *kernel* PCA [85]. PCA assume orthogonality, but after finding the first direction of maximum variance, the next direction of maximum variance does not necessarily form a right angle with the first one. Also the variance may not be explaining most information. Consider a circular cylinder of radius r and length ζ, and $r < \zeta$, but in slices just separating them. Maximum variance would erroneously be in direction of height, missing information about the slices. *Independent Component Analysis* (ICA) is a special case of PCA, searching for independent components using higher order fourth moment statistic *kurtosis*, rather than finding orthogonal maximum variance components, and assuming independent sources of non-normal distribution of measured variables. By contrast, PCA requires probability distributions of the exponential class, e.g. normal and exponential distribution, the only ones described entirely by mean and variance which are the only statistics used in the analysis. Back to the dimension of X, observations should be greater than dimension, i.e. $n > m$. It has been suggested that PCA in so called *high dimensions low sample size* (HDLSS) situations might not be good practice, not yielding satisfactory results [92, 93]. By the CLT in probability theory, features in one dimension with finite variance of any distributions tends to a normal probability density function for large sample size. And larger number of observations will minimize the probability of errors, maximizing population estimation accuracy and increase generalization of the learning [94], hence avoiding overfitting due to the curse of dimensionality.

5.4 Feature Selection

Feature selection refers to dimensionality reduction methods aiming to find an optimal subset of the original measured features. Generally it can be given by

$$\begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_m} \end{bmatrix} \xrightarrow{g} \begin{bmatrix} \mathbf{x_{i_1}} \\ \mathbf{x_{i_2}} \\ \vdots \\ \mathbf{x_{i_d}} \end{bmatrix}, d < m, i \in 1,2,\ldots,m \tag{5.12}$$

where g is any function selecting d features from the m original $\mathbf{x_i}$ features according to a certain criterion, so that $d < m$ [2, 75]. Feature selection methods are generally divided

into three categorised based on different strategies and associated properties[72, 73]

1. Filter:
    - Independent of learning model (local performance evaluation)
    - Computationally less-expensive
    - Resilient to overfitting
2. Wrapper:
    - Dependent of learning model (predictive performance evaluation)
    - Computationally more-expensive
    - Prone to overfitting
3. Embedded:
    - Built-in (embedded) in learning model (predictive performance evaluation and incorporated in training process)
    - Computationally less-expensive than wrapper methods
    - More resilient to overfitting than wrapper methods

Filter methods are less complex with low risk of overfitting, nevertheless the removal from the learning algorithm can often result in a suboptimal feature subset. Both filters and wrappers can search for all combination in the feature space, although filters are mostly linked with filter ranking methods. Here features are evaluated individually, with a statistical measure such as Person's correlation coefficient or a test statistic from classical statistical tests, using all observations for training, and ranked thereby. But in general, methods not optimizing the predictor of a learning model are labeled filters and used as a first step of feature selection, hence the name. And so hybrid methods exists. Wrapper methods on the other hand use a learning algorithm for evaluation, measuring predictive score for some feature subset at the time. Embedded methods too uses a predictive performance evaluation, but combine the subset search procedure with training of the algorithm. Since these two strategies are connected to a predictive model, *cross-validation*, which will be explained in Section 6.1, can be added to the evaluation. Regardless of the three categories, methods can be either *univariate*, e.g. ranking methods, or *multivariate*, e.g. Relief filtering and wrapper or embedded *greedy algorithms* for forward or backward elimination seeking some local optimization to improve globally. Only applying single features evaluation for selection, chances are the subset is highly redundant, or missing relevant features with jointly predictive power, albeit irrelevant alone. Finally, feature extraction can be done in cascade with feature selection for obtaining an optimal dimensionality reduced feature set. For instance, PCA could be used either before or after a selection methods. Going back to the example data from Figure 5.5 yet again, this two random features scattered in (a) could have been found with feature ranking, and redundancy and noise could be further removed selecting PC1 of largest variance. Alternatively, and also this time picturing a close second parallel linear cluster in (a), after the same rotation shown from (a) to (b), classification would be possible from only feature PC2. Drawing a straight horizontal line given by some threshold value would separate the two clusters in (b), and feature selection could be done to choose PC2 scores, even though the clusters were so close together direction of highest explained variance would be given by PC1 coefficients.

5.4.1 Filter Methods

Feature selection can be performed to rank the given feature set by some scoring measure. This can done both unsupervised and supervised. For example, features can be ranked unsupervised by the variance within each dimension. Supervised feature filter could be done by statistical tests. Each feature is grouped by the class labels, and then a test static can give the score or associated probability of classes having significantly different mean values. In binary class situations, statistical tests include *student's t-test*, *chi-squared test* or *Fisher's exact test*. For multiclass problems, these can be decomposed into several two class problems, or solved directly [95]. *Analysis of Variance* (ANOVA) is a collection statistical models for multiple populations. All these aforementioned statistical tests are univariate, finding the discriminative score between classes for one feature at a time. The choice of statistical test depend of the data. t-test for two population means is a univariate parametric hypothesis test to investigate the significance of the difference between the means of two populations, assumed to be normally distributed [96]. Variance is unknown and can be equal or not, depending on the method. The test statistic has t-distribution from which probabilities are calculated. The *Kruskall–Wallis rank sum test of K populations*, or *H-test*, is a nonparametric ANOVA test which does not assume normal populations. The test statistic is given by

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{F} \frac{R_i^2}{x_j} - 3(N+1), \tag{5.13}$$

where R_j is the rank sum (all samples are ranked by size) of sample number j with size x_j, N is the combined sample size and F is the degrees of freedom. H follows a chi-squared, $\mathcal{X}^2$, distribution with K-1 degrees of freedom. The null hypothesis of equal means is rejected when H exceeds the critical value, given by known values for the chi-squared distribution of various *statistical significance values* and degrees of freedom. When comparing multiple means, *multiple comparison tests* can be used *post-hoc*, or *ad-hoc* to a statistical test to find which means actually differs [97], such as the *Bonferroini correction*.

The supervised ReliefF nonlinear multivariate feature filter algorithm is based on the KNN supervised learner [72], to be explained in Section 6.4. The method evaluates a feature by how well its value distinguishes samples that are from different groups, but are similar to each other [43]. It can either be performed on all features with an exhaustive search, or greedy by heuristic select some features and evaluate. Before doing this the order of the features are randomized. The feature input space is made up of n feature vector observations $\mathbf{x_j}$, each with m elements or dimensions, producing the input matrix $\mathbf{X}_{i,j}$. For one feature vector at a time, which is a point in the feature space, the K closest points of the same class c, $\mathbf{x}_{c_k i}, k = 1, \ldots, K$ (nearest hits) and the K closest points for each different class l $\mathbf{x}_{l_k i}$ (nearest misses) are found using the KNN algorithm with *Manhattan distance*, or L^1-norm. For each feature, a ranking index, or weight, is

computed given by *pseudo code* in [98], or equivalently

$$\Theta(j) = \sum_{i=1}^{m} \left(\sum_{l \neq c} \frac{p_l}{1 - p_c} \sum_{k=1}^{K} \frac{|x_{i,j} - x_{l_k i,\bar{j}}|}{(\max_i x_i - \min_i x_i) nK} - \sum_{k=1}^{K} \frac{|x_{i,j} - x_{c_k i,\bar{j}}|}{(\max_i x_i - \min_i x_i) nK} \right) \tag{5.14}$$

where $\bar{j}$ means not j since nearest feature vector cannot be the feature evaluated. From (5.14), each feature x_j is scored as the sum of weighted differences in different classes and the same class. If the feature is differentially expressed, it will show greater differences for samples from different classes, thus it will receive higher score Θ, and vice versa [43]. Therefore, features can be ranked by the scores, from best with high score to worst with low (possibly negative) score. Ranking depend of the number of K nearest neighbors, which again is dependent on the data. For $K = 1$ estimated ranking can be unreliable for noisy data, while K comparable to the number of observations can result in important features not being found. A thorough explanation of ReliefF can be found in [98].

6 Supervised Classification

This chapter contains preprocessing steps and performance evaluation of supervised learning, followed by an overview of some distinct classification families, describing the core elements in each. This chapter is neither a full catalog nor a complete theoretical explanation, instead more information could be acquired in [1–4, 64, 99].

6.1 Preprocessing and Perfomance Evaluation

The components of learning an application can be given by [6]

$$Learning = Representation + Evaluation + Optimization \tag{6.1}$$

Before building a supervised learner, feature engineering is crucial, as explicated in chapter 5. The hope is to obtain a feature set for which the pattern recognition tasks are easier and faster to solve. This is especially true for situation with more dimensions m than observations n, $m > n$, which often occurs in computational biology areas as genomics and medical recording, together with *regularization* meaning providing more information about the specific problem to algorithms, all in order to circumvent overfitting [1, 92]. With a *representative* multidimensional feature set $\mathbf{X} \in \mathbb{R}^m$ of independent variables, and correct corresponding response or target variables $\mathbf{y} \in \mathbb{R}$, training can commence. The purpose is to build a model which generalizes input and output relationship, to increase the model performance by accurate prediction of independent new data. However, the training is sometimes merged together as with embedded feature selection. The responses can be either quantitative, i.e. numerical (continuous and discrete) variables for regression settings predicting a response or qualitative, i.e. categorical labels, for classification making decisions for new observations [2]. A pool of algorithms should be trained [6, 100], producing several models or a classifier set making up a larger hypothesis space searching for the true function f explaining the relationship between input and output. Classifiers can be divided into different families based on how the model is build. As previously mentioned, there is a distinction between parametric and nonparametric models. Moreover, decision theory or information theory rooted in probability theory should be involved in the classification process. From decision theory, finding the estimate $\hat{f}$ can be broken in two stages [2].

Inference involves learning the relationship between $\mathbf{X}$ and $\mathbf{y}$, how the change together and generally find the joint probability distribution $p(\mathbf{X}, \mathbf{y}$ [2, 3]. The other step involves the *prediction* or *decision*,

$$\hat{\mathbf{y}} = \hat{f}(\mathbf{X}). \tag{6.2}$$

Given true classification $\mathbf{y} = f(\mathbf{X}) + \epsilon$, and considering $\hat{f}$ and $\mathbf{X}$ fixed, then the expected value of the squared classification error reveals

$$\mathrm{E}[(\mathbf{y} - \hat{\mathbf{y}})^2] = \mathrm{E}[\big(f(\mathbf{X}) + \epsilon - \hat{f}(\mathbf{X}))^2] = \underbrace{(f(\mathbf{X} - \hat{f}(\mathbf{X}))^2}_{reducible} + \underbrace{\mathrm{Var}(\epsilon)}_{irreducible}, \tag{6.3}$$

meaning prediction will always contain some error. $\mathbf{y}$ is a function of ϵ, which cannot be be predicted by X since this error is introduced because it may include unmeasured variables not contained in X, or it may contain unmeasurable variation. However, prediction error in $\hat{f}$ can be reduced by improving the reducible addend using the most appropriate statistical techniques. *Bayes error rate* is the lowest possible error rate any classifier can achieve. Probabilities should be included in the decision part, and also more general criteria as misclassification rate or a loss/cost function. General *discriminant* models solve the inference and decision problems together by mapping new observation inputs x directly into decisions, without probabilities. Another way is *generative* models that infer the *posterior* probabilities using Bayes' theorem in (4.2) and model the distribution of input and output. Lastly *discriminative* models learn the boundary between classes and infer posterior probabilities directly.

In order to properly *evaluate* classifiers, an *objective function* or *scoring function* is needed [6]. This can differ for the internal function used in the classifier, to the external used to validate performance. Internally in all the classifiers explained in the coming sections, a *0-1 loss* or *cost function* is used, since class labels are categorical. Externally, different validation techniques can be used. In data-rich situations, the feature input set is usually divided into three sets; a training set, a validation set, and a test set [1]. There is no exact way to divide the input, but a typical distribution could be 50%, 25% and 25%, respectively. *Model selection* refers to estimating the performance of different classifiers to further select the best. The models are build using the training set, and performance is estimated by the validation set. *Model assessment* means using the final best classifier, and then find the the prediction error (generalization error) by testing on new data, hence the test set. Often data is scarce, and the number of observations are insufficient for the three way split. One of the simplest and most widely used method trying to overcome this issue is *cross validation*, again with different variations. *K-fold* cross validation splits the input data randomly into K roughly equal-sized parts. The model is then built with the training data in the $K-1$ parts, with testing or validation on the kth part. This is repeated K times, for $k = 1, \ldots, K$. As such, each fold is used for test one time, with training by the other folds. The *test error*, i.e. the generalization error, is given by

$$\mathrm{Error}_{\mathcal{T}} = \mathrm{E}[L(\mathbf{y}, \hat{f}(\mathbf{X}))|\mathcal{T}] \tag{6.4}$$

conditioned on the training set $\mathcal{T}$, for a general loss function L measuring the error. The goal for validation is to measure this error, but that is not always possible. A related measure is the expected prediction error

$$\text{Error} = \text{E}[L(\mathbf{y}, \hat{f}(\mathbf{X}))] = \text{E}[\text{Error}_{\mathcal{T}}] \tag{6.5}$$

which averages over everything that is random, including the randomness in the training set that produced $\hat{f}$. This is the error estimated by cross validation, i.e. the average generalization error. The prediction error can further be decomposed in bias and variance, as previously explained. Deciding the value of K for K-fold separation is a study of itself. Setting $K = n$, that is the number of observations, this is referred to as *leave-one-out* cross validation, with testing on one observation at the time, and training on the remaining observations. This cross validation estimate is approximately unbiased, but with high variance because all the training sets become similar to each other. Also this is computationally expensive. More typical compromises are $K = 5$ or $K = 10$ depending on the number of observations. The estimated error usually has lower variance, but could be biased. This often improves for increased size of the training set. Moreover, to increase the stability of the estimate, K-fold cross validation itself can be iterated several times, increasing the number of validations.

When performing supervised feature selection, this must be conducted inside the cross validation loop for proper prediction estimate. Doing otherwise, will result in biased estimates. When screening features before cross validation, the whole data set is treated as training set. The selected features from this set will have an unfair advantage, having already seen the class labels. By later dividing this data into training and test, the latter does not mimic a true independent test set. Instead, feature selection is performed on training data given by the $K-1$ folds, possibly resulting in different feature subset for each cross validation loop. Same goes for feature *resampling* of training data. In situations with imbalanced data sets, i.e. the number of observations for different classes are not of the same order, sampling approaches can be utilized. These can be either oversampling or undersampling. With abundance of data, the latter can be employed by removing observations from the majority class trying to equal out the observations imbalance towards the minority class. As such, this can be done prior to cross validation. However, this is not the case for oversampling, introducing new simulated observations which should only be used for training and model building, and not for testing. Oversampling can be applied when data is scarce, by simple copying of observation in the minority class. More sophisticated methods are the *Synthetic Minority Over-sampling Technique* (SMOTE) in [101] and the *Adaptive Synthetic* (ADASYN) sampling approach in [102]. Both algorithms improve class balance by synthetically creating new observations from the minority class via linear *interpolation* between existing minority class observations [103]. The latter is an extension of the former, where more observations are created in the vicinity of the boundary between two classes, rather than in the interior of the minority class. Another approach to overcome uneven observations for the classes, is to change the cost function. Wrong classification of the minority class can be more heavily penalized, thus reducing these errors in case to avoid an important class never gets misclassified.

To assess the performance of a classifier, different performance evaluation measures can be used. This plays a crucial part, also for *tuning* of a classifier to perform optimally [104]. Traditionally, the overall accuracy is most commonly used to measure performance. However, with imbalanced data, especially in extreme cases with up to 99% of observations belonging to one class, other measures should be included. The *confusion matrix* for a binary 2-class setting is given in Table 6.1.

Table 6.1: Confusion matrix, with frequency count of the predictions.

	Predicted class 1	Predicted class 2
Actual class 1	True class 1 (T1)	False class 1 (F1)
Actual class 2	False class 2 (F2)	True class 2 (T2)

This matrix could be extended for multi M-class problems, producing $M \times M$ matrices. Several evaluation measures can be derived from the confusion matrix, including:

- True class 1 rate: $T1_{rate} = \frac{T1}{T1+F2}$
- Trie class 2 rate: $T2_{rate} = \frac{T2}{T2+F1}$

In binary settings, class 1 and 2 are usually referred as positive and negative class, and *sensitivity* is the true positive rate and *specificity* is the true negative rate.

Finally *optimization* refers to searching for the highest scoring classifier, with different algorithms depending of the scoring function.

6.2 Decision Tree

Decision tree algorithms are a family of supervised learning using a decision tree as predictive model. It is one of the most employed approaches to classification, and used in a variety of disciplines [4]. The input space $\mathbf{X}$ is recursively partitioned, building a *directed rooted* tree. It starts from one root *node* representing the whole data set with no incoming edges and splits into several nodes, or subnodes, all with exactly one incoming edge. Nodes that are split are called *internal*, *decision* or *test*, while *leaf* or *terminal* nodes have no outgoing edges. Other terms include a *parent* node splitting into *children* nodes. A branch is a subsection of the entire tree of nodes. Trees slitting nodes in two are called binary trees, but multiple splitting is also possible. Structure and terminology are illustrated in Figure 6.1.

Decision trees work for inputs and outputs that are numerical or categorical, the latter both binary or multiclass. Thus each leaf represents either a class label or the probability of a response variable have a certain value. Starting at the root, prediction of new observations feature vector x is performed by guiding down the tree which splits certain features based on different decision criteria before ending in a leaf, e.g. a class. This is

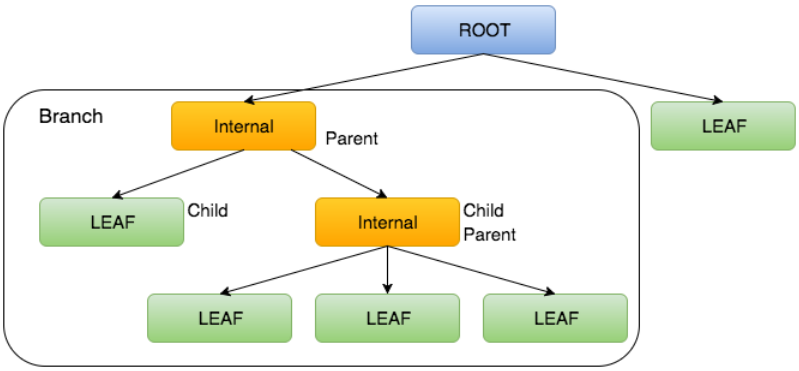


Figure 6.1: Chart of decision tree with terminology. Reproduced from [105].

related to rule induction, where the different decisions are rules ultimately combined to make a rule set. The splitting decision contributes greatly to prediction accuracy, and can by done with several functions. Univariate splitting uses only a single feature. The overall best feature for distinguishing classes given a certain function is split in the root node, then the next best is split in a internal node and so on. Impurity measure is statistical dispersion and defines how well classes are separated, and should satisfy (1) being largest when all classes have equal amount of observations (max impure) and (2) zero when all data belong to one class (pure) [106]. The Gini's Diversity Index works for binary split and is an impurity-based criteria given by

$$Gini(y,n) = 1 - \sum_{c=1}^{C} \pi_c^2, \tag{6.6}$$

where π_c are the empirical class probabilities given by (4.1) for different classes c. The index is highest when all probabilities π_c are equal, and zero for a one class situation with $\pi_c = 1$. This index is computed for each feature, and the one resulting in highest score is split at the feature value yielding the score. Multivariate splitting is more complicated and can also be done by several features participating in the node split, sometimes drastically improving the tree's performance [4]. The tree structure can be viewed as a *combining model* or *committee* approach and a greedy algorithm where each splitting is a model optimizing some local decision criteria given by the splitting function [3].

Decision trees are stable, nonparametric and discriminative models with several advantages [4, 105]. They map both linear and non linear relationships, but linear methods are better for those cases. They also give a graphical representation fairly easy to understand without much background knowledge, which also identifies significant features and their relationship making trees suitable for feature selection [107, 108]. Therefore, less preprocessing is required since outliers have reduces influence. As mentioned, continuous variables are supported, but trees are better fitted for class labels. Furthermore, as trees grows by continuous node splitting so does obscurity and complexity, again affecting overfitting and accuracy. Complexity is usually measured by total number of nodes or leaves, the tree depth and total features used [4]. The goal is to find the optimal

tree, but this is NP-hard[1], and thus only feasible in small problems. Hence, heuristics methods seekinol an approximate of the real solution are necessary, generally separated in bottom-top and the mostly preferred top-bottom strategies. The latter use growing by splitting and *pruning*, i.e. removing nodes, as targets to control complexity and obtain optimal decision tree. Growing will continue until there is one leaf for each observation, i.e. maximum overfitting, unless there is a stopping criteria. This could be, among many others, maximum depth by number of splits or all observations in the training set belong to a single target value y. As seen, loose stopping criteria increases risk of overfitting. However, tight stopping criteria tends to underfit the training set. This is the reason for pruning methods. The stopping is set loose, intentionally solving for an overfit solution. This tree is then reduced by removing branches that are not contributing to the generalization accuracy. Also pruning can turn trees more comprehensible while maintaining adequate accuracy. There are several methods, where one is known as cost-complexity pruning. Here the original tree T_0 is pruned in several stages producing a sequence $T_0, T_1, \ldots, T_R$, where T_R is the root tree. T_{r+1} is found from T_r where leaves are replaced by branches that have the lowest increase in apparent error rate per pruned leaf node, given by

$$\alpha = \frac{\varepsilon(pruned(T,t),B) - \varepsilon(T,B)}{|leaves(T)| - |leaves(pruned(T,t))|}. \tag{6.7}$$

Here $\varepsilon(T,B)$ denotes the error rate of the tree T over the sample B, $|leaves(T)|$ the number of leaves in T and $pruned(T,t)$ the tree obtained by replacing the node t in T with a suitable leaf. Then the generalization error for all each tree T_p is estimated and the best tree selected.

6.3 Discriminant Analysis

Discriminant function algorithms are a family of supervised learning used for both regression and classification, and that model differences between classes by producing a discriminant rule for classification. Given features in $\mathbf{X}$ the goal is to predict values in $\mathbf{y}$ with a function f. The predicted estimate from the input, $\hat{\mathbf{y}}(\mathbf{X})$ will for categorical values assume classes in the set $\mathcal{Y}_c, c = 1,2,\ldots,C$. The loss function can be a $C \times C$ matrix $\mathbf{L}$ which is zero on the diagonal and nonnegative elsewhere with a values representing the cost of prediction error [1]. The expected cost of misclassication is

$$\text{ECM} = \text{E}[\text{L}(\mathbf{y}, \hat{\mathbf{y}}(\mathbf{X}))], \tag{6.8}$$

with joint probability $\Pr(\mathbf{X}, \mathbf{y})$. Conditioning on $\mathbf{X}$ yields

$$\text{ECM} = \text{E}_{\mathbf{X}} \sum_{c=1}^{C} \text{L}(\mathcal{Y}_c, \hat{\mathbf{y}}(\mathbf{X})) \Pr(\mathcal{Y}_c|\mathbf{X}), \tag{6.9}$$

[1] NP-hardness (non-deterministic polynomial-time hard), in computational complexity theory, is a class of problems that are, informally, "at least as hard as the hardest problems in NP [a complexity class used to describe certain types of decision problems.]" [109]

which can be minimized pointwise resulting in the class prediction for input vector $\mathbf{x}$ being given by

$$\hat{\mathbf{y}}(\mathbf{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmin}} \sum_{c=1}^{C} \mathrm{L}(\mathcal{Y}_c, y) \Pr(\mathcal{Y}_c|\mathbf{x}). \tag{6.10}$$

Assuming a loss function $0-1$, meaning zero on the diagonal and cost of misclassification is 1 elsewhere in matrix $\mathbf{L}$, the above equation simplifies to

$$\hat{\mathbf{y}}(\mathbf{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmin}} (1 - \Pr(y|\mathbf{x})) \tag{6.11}$$

or equivalently

$$\hat{\mathbf{y}}(\mathbf{x}) = \mathcal{Y}_c \;\; if \;\; \Pr(\mathcal{Y}_c|\mathbf{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \Pr(y|\mathbf{x}). \tag{6.12}$$

This solution is known as *Bayes classifier*, and states that given an input the classification should be the most probable class, i.e. conditional posterior probability $\Pr(\mathbf{y}|\mathbf{X})$ is needed optimal classification. This serves as an unattainable standard, and many methods try to estimate this conditional probability [2]. By Bayes' theorem in (4.2) combined with the sum rule, the posterior probability for each class is

$$\Pr(\mathcal{Y}_c|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{Y}_c) \Pr(\mathcal{Y}_c)}{\sum_j p(\mathbf{x}|\mathcal{Y}_\zeta) \Pr(\mathcal{C}_\zeta)}. \tag{6.13}$$

Each class conditional density $p(\mathbf{x}|\mathcal{Y}_c)$ can be model as the multivariate normal distribution given by (4.13), yielding a generative model, and class probabilities can be found *a priori*. *Linear discriminant analysis* (LDA), one of the most widely-used classifier, is the special case when the covariance matrix from the distribution function of each class Σ_c is assumed to be equal, $\Sigma_c = \Sigma \forall c$. Since linearity of boundaries are fulfilled also for the monotone transform, comparing two classes $\mathcal{Y}_c$ and $\mathcal{Y}_\zeta$ and using the natural logarithm then becomes

$$\begin{aligned} \ln \frac{\Pr(\mathcal{Y}_c|\mathbf{x})}{\Pr(\mathcal{Y}_\zeta|\mathbf{x})} &= \ln \frac{p(\mathbf{x}|\mathcal{Y}_c)}{p(\mathbf{x}|\mathcal{Y}_\zeta)} + \ln \frac{\Pr(\mathcal{Y}_c)}{\Pr(\mathcal{Y}_\zeta)} \\ &= -\frac{1}{2}(\mu_c + \mu_\zeta)^T \Sigma^{-1}(\mu_c - \mu_\zeta) + \mathbf{x}^T \Sigma^{-1}(\mu_c - \mu_\zeta) + \ln \frac{\Pr(\mathcal{Y}_c)}{\Pr(\mathcal{Y}_\zeta)}, \end{aligned} \tag{6.14}$$

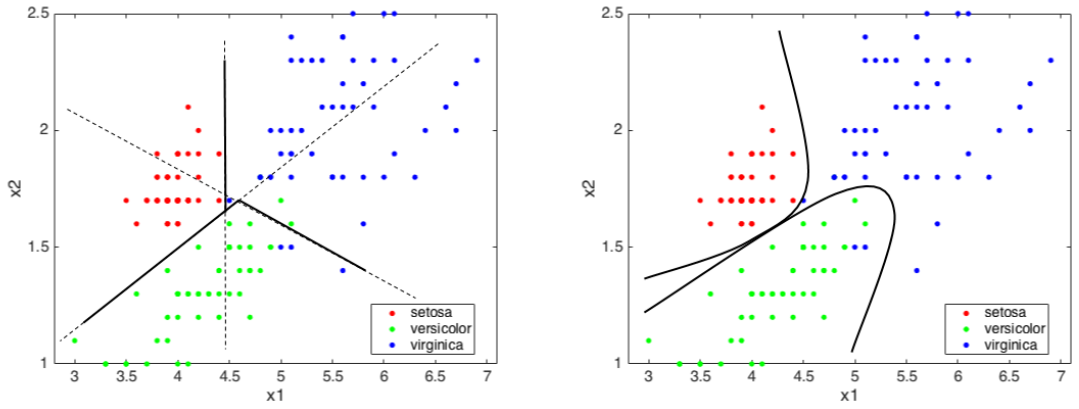
where covariance matrix identity makes normalization factors and the quadratic part in the exponents to cancel. From (6.14) and $\Pr(\mathcal{Y}_c) = \pi_c$ follows the linear discriminant functions for each class $\mathcal{Y}_c \forall c$

$$\begin{aligned} \delta_c(\mathbf{x}) &= \mathbf{x}^T \Sigma^{-1} \mu_c - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c + \ln p(\mathcal{Y}_c) \\ &= \boldsymbol{\beta}_c^T \mathbf{x} + \beta_{c0} \\ \boldsymbol{\beta}_c &= \Sigma^{-1} \mu_c \\ \beta_{c0} &= -\frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c + \ln \pi_c, \end{aligned} \tag{6.15}$$

which are clearly linear in $\mathbf{x}$, and consequently optimal prediction is also given by $\mathbf{y}(\mathbf{x}) = \text{argmax}_c \delta_c(\mathbf{x})$. Here $\boldsymbol{\beta}$ is the weight vector and β_0 is the bias.This is the general notation for discriminant functions, and posterior probabilities could for instance be modeled directly the exponentially expressions [1, 3] Parameter estimates for each class can be given by (4.4) for means, $\hat{\mu}_c = \sum_{y_j=c} x_j / n_c$ and by (4.1) for class probabilities $\hat{\pi}_c = n_c/n$. The pooled covariance matrix estimate can further be computed by $\hat{\Sigma} = \sum_{c=1}^{C} \sum_{y_j=c} (x_j - \hat{\mu}_c)(x_j - \hat{\mu}_c)^T / (n - C)$. The boundaries are given by $\delta_c(\mathbf{x}) = \delta_\zeta(\mathbf{x})$, that is, the set $\{\mathbf{x} : (\hat{\beta}_{c0} - \hat{\beta}_{\zeta 0}) + (\hat{\boldsymbol{\beta}}_c - \hat{\boldsymbol{\beta}}_\zeta)^T \mathbf{x} = 0\}$. The linear boundaries between all classes are computed, and then combined to divide the feature space into regions $\mathcal{R}_c$, as depicted for in the simulated data in Figure 6.2 (a). Between class affine hyperplanes, the stippled lines, together separate the arbitrary feature plane given by the hard lines. A new observation score in this space given by its feature vector will be classified to the region it appears, i.e. the discriminate function δ_c maximizing the vector. When the covariance matrices Σ_c are not assumed equal, exponents do not cancel when comparing classes resulting in the following discriminant functions

$$\delta_c(\mathbf{x}) = -\frac{1}{2} \ln |\Sigma_c| - \frac{1}{2} (\mathbf{x} - \mu_c)^T \Sigma_c^{-1} (\mathbf{x} - \mu_c) + \ln \pi_c \tag{6.16}$$

which yield *quadratic* discriminate analysis. This is depicted in Figure 6.2 (b) for the same example data as in (a).



(a) Linear functions

(b) Quadratic functions

Figure 6.2: Linear (a) and quadratic (b) discriminant analysis. Reproduced from [1, 110].

Here estimates are computed as before, except the covariance matrices are diagonalized by eigenvalue decomposition as in (5.9). So $\Sigma_c = \mathbf{E}_c \Lambda_c \mathbf{E}_c^T$ and then $(\mathbf{x} - \hat{\mu}_c)^T \hat{\Sigma}_c^{-1} (\mathbf{x} - \hat{\mu}_c) = (\mathbf{E}_c^T (\mathbf{x} - \hat{\mu}_c))^T \Lambda_c^{-1} (\mathbf{E}_c^T (\mathbf{x} - \hat{\mu}_c))$ and $\ln |\hat{\Sigma}_c| = \sum_\zeta \lambda_{ce}$, where λ_{ce} are the diagonal elements for diagonal matrix Λ_c. Both LDA and QDA are parametric models, using Bayes discriminant rule maximizing posterior probabilities. They are sensitive to outliers, and assume normal distribution and nonmulticollinearity, i.e. not high correlation between independent variables [111]. However, the methods are simple and robust, often performing as good as more complex models [112]. LDA is a generalization of Fisher's linear discriminant function using another rule maximizing ratio between the pooled

covariance matrix and within class covariance matrices, and which is related to PCA. It is also related to ANOVA (and Multivariate ANOVA (MANOVA)) which can be viewed as the reversed procedure wanting to separate the dependent variables.

6.4 Nearest Neighbor Classifiers

Nearest neighbor methods are a family of supervised learning that can handle both regression and classification, and which stand out from the other algorithms scrutinized in this thesis since no model need to be fit [1]. Instead, these classifiers are instance-based using memory to store the feature space of n data point from training with associated class labels. As such, they are also *lazy*, as oppose to *eager* training, in the sense explicit generalization is delayed until new observations are introduced. *K-nearest neighbor* (KNN) is one of the widely most used classifiers [2], and uses a nonparametric density estimator technique [3]. Given a new input $\mathbf{x}$, a sphere can be centered on this point in the feature space and increased with volume V until containing precisely K training points irrespectively of corresponding class. With n_c point total and K_c point in the sphere of class $\mathcal{Y}_c$, then the following densities is given by conditional probability

$$p(\mathbf{x}|\mathcal{Y}_c) = \frac{K_c}{n_c V} \tag{6.17}$$

and similarly by unconditional probability

$$p(\mathbf{x}) = \frac{K}{NV} \tag{6.18}$$

while class probabilities are still computed from (4.1), $\Pr(\mathcal{Y}_c) = n_c/n$. Applying Bayes' theorem specifies the posterior probabilities

$$\Pr(\mathcal{Y}_c|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{Y}_c)\Pr(\mathcal{Y}_c)}{p(\mathbf{x})} = \frac{\frac{K_c}{N_c V}\frac{N_c}{N}}{\frac{K}{NV}} = \frac{K_c}{K}, \tag{6.19}$$

which are modelled directly without probability distributions resulting in a discriminative method. Equivalently the predictor is given by [1]

$$\hat{\mathbf{y}}(\mathbf{x}) = \frac{1}{K}\sum_{x_{ij}\in N_c(\mathbf{x})} y_j. \tag{6.20}$$

The algorithms takes an m-dimensional input $\mathbf{x}$, which will be a query point x_0 in the feature space along with n training points of pair (x_j, y_j) given by feature vectors $\mathbf{x}_j$ and class labels vector $\mathbf{y}$. Then the K pairs nearest in distance to the query are located, and subsequent classification to the majority class. Thus misclassification is minimized by classifying to the largest posterior probability. The closeness involves some metric, and several distance measures can be used, some of which defined in Table 6.2.

Further, the distances ξ can be *weighted*. Computing for instance $\xi' = 1/\xi$ or $\xi' = 1/\xi^2$, will put greater emphasis on nearer neighbors. Usually each of the features $\mathbf{x}_i$ are standardize to have mean zero and variance 1, since it is possible that they are measured in

Table 6.2: Selected distance measures and corresponding definitions.

Distance measure	Definition
Manhattan/Taxicab (from L^1-norm)	$\|\|x_{ji} - x_{0i}\|\|_1 = \sum_{i=1}^{m} \|x_{ji} - x_{0i}\|$
Euclidean (from L^2-norm)	$\|\|x_{ji} - x_{0i}\|\| = \sqrt{\sum_{i=1}^{m} (x_{ji} - x_{0i})^2}$
Minkowski (from L^Q-norm)	$\|\|x_{ji} - x_{0i}\|\|_Q = \left(\sum_{i=1}^{m} \|x_{ji} - x_{0i}\|^Q\right)^{1/Q}$
Cosine	$\cos(x_{ji} - x_{0i}) = \frac{\sum_{i=1}^{m} x_{ji} x_{0i}}{\sum_{i=1}^{m} x_{ji}^2 \sum_{i=1}^{m} x_{0i}^2}$

different units. In case of a tie, class selection can be chosen at random or the nearest neighbor. To avoid ties, K should not be a multiple of the number of classes C.

KNN is related to *kernel density estimators*, which instead uses fixed volume V and variable K. KNN is a simple, yet accurate classifier with good reported performance for many settings, especially for nonlinear and highly irregular boundaries. However, due to the sparsity and data point being on the edge in higher dimensions, as explain in section 5.2, distance measures often fail in this scenario. Also, the effective number of parameters is n/K, and thus computational complexity increases with respect to data size. Therefore prediction can be slow, contrary to the fast memory-based approach for training. Furthermore, K does not imply performance, which can increase together with K until a certain point before deteriorating. Generalization variance decreases, but bias increases. Performance is data dependent, shown for simulated data with different K in [1]. $K = 1$ gives low bias but high variance, and can sometimes produce the best result. 1-NN is a special case of KNN. The feature space is then divided into *Voronoi cells*, also known by many other names, which determines classification. This is also a special of *prototype* or *centroid* methods with training examples as centers, and further relates to *k-mean* unsupervised clustering. Here data is partitioned into K clusters with n observations belonging to clusters with nearest mean which is the centroid and prototype.

6.5 Support Vector Machine

SVM algorithms are a family of supervised learning that is nonparametric employing discriminative hyperplanes based on *sparse kernel machines* [3]. The goal is to construct decision boundaries that explicitly try to separate data into different classes as well as possible [1]. The m-dimensional feature vectors $\mathbf{x}_j$ are observation points x_j in the feature space. These have associated class labels y_j, thus for n observations this produce n pairs (x_1, y_1),(x_2, y_2),...,(x_n, y_n), with $\mathbf{x}_j \in \mathrm{R}^m$ and $y_j = \{-1, 1\}$. As previously stated, a hyperplane can be defined by

$$\{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta} + \beta_0 = 0\}. \tag{6.21}$$

There could be numerous, suboptimal, hyperplane boundaries separating two classes,

as seen in Figure 6.3 (a) for simulated data with arbitrary classes, all with distances $|\mathbf{x}^T\boldsymbol{\beta}+\beta_0|/||\beta||$ to observations $\mathbf{x}_j$. Letting β be unit vector, $||\beta||=1$, a classification rule induced by $f(x)$ is

$$\mathbf{y}(\mathbf{x}) = \text{sign}[\mathbf{x}^T\boldsymbol{\beta}+\beta_0], \tag{6.22}$$

which is optimal classification of two classes, seen in Figure 6.3 (b). Assuming linearly separable classes, there exists at least one choice of the parameters $\boldsymbol{\beta}$ and β_0 so that for $y_j=+1$, $f(\mathbf{x}_j)>0$ and $y_j=-1$, $f(\mathbf{x}_j)<0$, then $y_j f(\mathbf{x}_j)>0 \quad \forall j$. From this, the hyperplane creating the largest *margin* M between training points from class $+1$ and -1, reduces to the optimization problem

$$\max_{\beta_0,\boldsymbol{\beta},||\boldsymbol{\beta}||=1} M \quad subject\ to \quad y_j(\mathbf{x}_j^T\boldsymbol{\beta}+\beta_0)\geq M, j=1,\ldots,n. \tag{6.23}$$

Replacing the condition with $y_j(\mathbf{x}_j^T\boldsymbol{\beta}+\beta_0)\geq M||\boldsymbol{\beta}||$ reduces the constraint $||\boldsymbol{\beta}||=1$. Since the distances for special observations, or the *support vectors*, are $1/||\boldsymbol{\beta}||$, the margin M is twice this distance as seen in Figure 6.3 (b). The optimization problem is thus equivalently

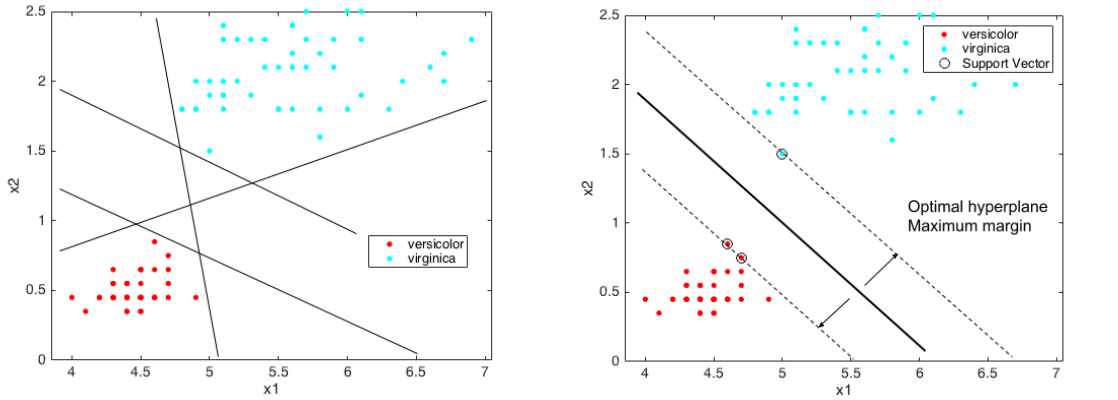
$$\min_{\beta_0,\boldsymbol{\beta}} \frac{1}{2}||\boldsymbol{\beta}||^2 \quad subject\ to \quad y_j(\mathbf{x}_j^T\boldsymbol{\beta}+\beta_0)\geq 1, j=1,\ldots,n, \tag{6.24}$$

where the condition is the *hinge loss function*. For nonseparable cases, with overlap between classes, the optimization problem is still to maximize the margin M, but allowing some points to be on the wrong side. Instead of hard margins ± 1, soft margins or *slack variables* can be defined as $\xi=[\xi_1,\xi_2,\ldots,\xi_n]$, and the new constraint is $y_j(\mathbf{x}_j^T\boldsymbol{\beta}+\beta_0)\geq M(1-\xi_j) \quad \forall j, \xi_j\geq 0, \sum_{j=1}^n \xi_j \leq constant$. By again dropping norm constraint on $\boldsymbol{\beta}$, and defining $M=1/||\boldsymbol{\beta}||$, the optimization problem can be stated as

$$\min||\boldsymbol{\beta}|| \quad subject\ to \quad \begin{cases} y_j(\mathbf{x}_j^T\boldsymbol{\beta}+\beta_0)\geq M(1-\xi_j)\ \forall j, \\ \xi_j\geq 0, \sum_{j=1}^n \xi_j \leq constant \end{cases} \tag{6.25}$$

where points well inside their class boundary do not play a big role in shaping the boundary. Thus only selected observations as support vectors set the boundary as seen in Figure 6.3 (b). This is different form LDA where the decision boundary is determined by the covariance of the class distributions and the positions of the class centroids. The problem in (6.25) is quadratic with linear inequality constraints. Hence it is a *convex optimization problem*, which can be solved by using *Lagrange multipliers* computing estimates $\hat{\mathbf{y}}(\mathbf{x})=\text{sign}[\hat{f}(\mathbf{x})]=\text{sign}[\mathbf{x}^T\boldsymbol{\beta}+\beta]$.

The Support vector (SV) classifier described so far is limited to linear boundaries. By selecting basis functions $\phi_l, l=1,\ldots,L$, the SV classifier is fit to input features $\phi(\mathbf{x}_j)=[\phi_1(\mathbf{x}_j),\phi_2(\mathbf{x}_j),\ldots,\phi_L(\mathbf{x}_j)], j=1,\ldots,n$ and produces the potentially nonlinear function $\hat{f}(\mathbf{x})=\phi(\mathbf{x})^T\hat{\boldsymbol{\beta}}+\hat{\beta}_0$, with still $\hat{\mathbf{y}}(\mathbf{x}=\text{sign}[\hat{f}(\mathbf{x})]$. The SVM is an extension of this idea. From the Lagrange programming to optimal solution, a cost parameter Υ is introduced instead of the constraint, together with multipliers α_j for which $0\geq\alpha_j\leq\Upsilon$, also called



(a) Suboptimal boundaries

(b) Quadratic functions

Figure 6.3: Suboptimal class boundaries (a) and optimal boundary induced by SVM (b).

box constraints. Υ controls the trade-off between the slack variable penalty and the margin. Furthermore, the solution function can be written as

$$f(\mathbf{x}) = \phi(\mathbf{x})^T + \beta_0 = \sum_{j=1}^{n} \alpha_j y_j \Phi(\mathbf{x}_j, \mathbf{x}) + \beta_0, \tag{6.26}$$

where $\Phi(\mathbf{x}_j, \mathbf{x}) = \Phi(\mathbf{x}, \mathbf{x}_j) = \phi(\mathbf{x}_j)^T \phi(\mathbf{x}')$ is the symmetric *kernel function*. As such, the solution only involves the feature mapping ϕ through inner products. Thus this need not be specified at at, only the kernel function Φ, which are able to operate in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space. This is known as the *kernel trick*. The dimension of the enlarged and transformed space is allowed to get very large, infinite in some cases. Here linear boundaries are constructed, producing nonlinear boundaries in lower dimensions. The sparsity in SVM comes from the fact that the kernel is not given for all observations in the training set, but rather the few support vectors on the boundary of each class. The customized kernel should be a symmetric positive (semi-) definite function. Some known kernels are presented in Table 6.3.

Table 6.3: Selected kernel functions, definitions taken from [113].

Kernel	**Definition**
Linear	$\Phi(\mathbf{x}_j, \mathbf{x}) = \mathbf{x}_j \cdot \mathbf{x}$
Gaussian or radial basis	$\Phi(\mathbf{x}_j, \mathbf{x}) = e^{-\|\|\mathbf{x}_j - \mathbf{x}\|\|^2}$
Polynomial	$\Phi(\mathbf{x}_j, \mathbf{x}) = (1 + \mathbf{x}_j \cdot \mathbf{x})^Q$

SVM is fundamentally defined for binary problems, however it can be extended to multiclass problems when classes $C > 2$ [3]. There are different ways of implementing this, with associated advantages and disadvantages. One commonly used approach

called the *one-versus-the-rest* conducts C separate SVMs, where the cth model uses training data of that class as positive examples and the data from the remaining $C-1$ classes as the negative examples. Another approach called the *one-versus-one* trains $C(C-1)/2$ different 2-class SVMs on all possible pairs of classes, and then classify test points according to which class has the highest number of "votes".

SVM supports both regression and classification. It solves for a local solution, which is also a global optimum due to the convex problem statement. The implementation approximate the bound on the generalization error, and depends of the margin M. But this is independent of the dimensionality of the feature space, and therefore mapping the data to higher dimensionality does not lead to overfitting. As such, SVM is resistant against overfitting, but this again depend on the tuning parameters. The regularization parameter Υ introduced by the convex optimization, which defines the margin, must be carefully selected together with choice of kernel and corresponding scaling [114]. This comes at a computational expensive cost.

7 Analysis

This chapter describes the collection of data including noise removal from different patients, and the subsequent analysis. Three biomedical signals were first recorded and transformed to the frequency domain before extracting different features. Using these, visualization and classification of distinct periods in the recordings were performed.

7.1 Data Collection

The study included ten patients scheduled for standard elective, no-emergency CABG surgery. Seven men and three women were included, with age ranging from 47 to 89, mean 69.9 and median 72.5. The surgery was performed at St. Olav's Hospital, Trondheim, Norway. All patients signed an agreement of participation. The study protocol was approved by the regional ethical committee (REK). The study was also added to the protocol registration and results system clinicaltrials.gov, and progress of the study was updated there. The surgery was performed during general balanced anaesthesia with a mixture of intravenous- and gas-anaesthesia, using *Propofol* and *Isoflurane*. The exclusion criteria were non-sinus rhythm, ejection fraction (EF) < 0.5, severe valve disease, right ventricular failure, pulmonary hypertension, and severe postoperative *hemorrhage*, i.e. escape of blood from a ruptured blood vessel.

The biomedical signal recordings were conducted by medical students at St. Olav's University Hospital [14]. The patients were monitored for 60 minutes before start of surgery, and from the end of surgery until the next morning. All three biomedical signals were recorded simultaneously, and fully synchronized. This was verified physiologically by inspection of the ECG and ABP waveforms, with the systolic pressure peak slightly trailing the QRS complex due to the prior ventricle depolarization causing the contraction, volume decrease and pressure increase. 3-lead ECG was recorded using three electrodes to calculated voltage differences, one on each shoulder and one on the left hip, also called the *Einthoven's triangle*. Here the heart is in the center, with zero net potential. In this setup lead I refers to voltage difference between right (negative pole) and left (positive pole) shoulder, lead II between right shoulder (negative pole) and left hip (positive pole), and lead III between left shoulder (negative pole) and left hip (positive pole). The electrodes were connected to a F135 Dual Bio Amp (ADInstruments,

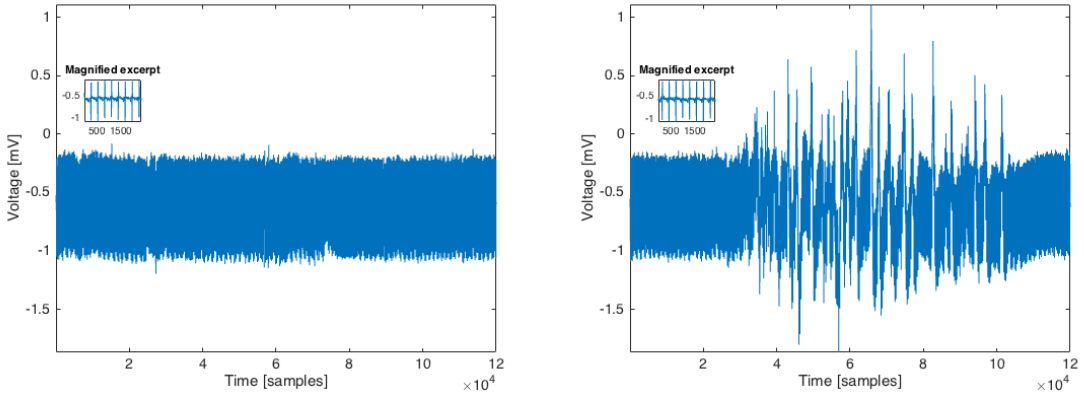
NZ) amplifier using MLA2340 Lead Wires (ADInstruments, NZ). Only the lead II analog voltage signal was used further in the analysis. ABP was obtained invasively by placing an Arterial Cannula 20G (BD, USA) in an artery in the wrist, often *arteria radialis*. Other commonly used arteries include *femoral*, *doralis pedis* or *brachial*. The cannula led to a sterile, fluid-filled infusion set with sodium chloride, NaCl. This was connected to a pressure MLT0670 Disposable BP Transducer (stopcock) (ADInstruments, NZ), where liquid within the infusion tubing is in contact with a diaphragm that moves in response to the transmitted pressure waveform, and converts the movement to an electrical analog signal. The transducer needed to be kept horizontally level with the patient, traditionally, the right atrium. Raising or lowering the transducer relative to the patient will alter the reading. Using a MLAC05 Deltran II cable (ADInstruments, NZ) the transducer were again connected to a FE117 BP Amp (ADInstruments, NZ) amplifier. LDF is usually employed to measures peripheral blood flow of smaller vessels in the microcirculation. This is not restricted to a specific surface on the body, but measurements were chosen to be obtained from the leg, since this placement were anticipated to attract less movement artifacts. A MSP100XP Standard Surface Probe (ADInstruments, NZ) sensor converted the biological signal into an electrical voltage analog signal, and was further connected to a LNL191 Blood FlowMeter (ADInstruments, NZ) amplifier. All three analog signals from the amplifiers were connected with MLACxx BNC to BNC Cables (ADInstruments, NZ) to the data acquisition (DAQ) hardware system PowerLab 16/35 (ADInstruments, NZ). Together with the software LabChart v8.1.5 Windows (ADInstruments, NZ), the signals were digitalized to *Nyquist rate* of 400 Hz, well above the frequency band below 2 Hz to be investigated. However, this rate was selected to correspond with previous papers, reporting frequency components of a human's ECG signal up to 100 Hz [115], consequently requiring a sampling rate of at least 200 Hz due to the *Nyquist theorem*.

(a) Close-up (b) Overview

Figure 7.1: Pictures of the equipment used for acquisition of data.

The LabChart software was used to convert the digital biomedical signals to *MAT*-files, and then accessed with the software MATLAB R2015b (MathWorks Inc., Natick, MA) which was used for the further analyses. By also employing the LabChart Reader v8.1.1 Mac (ADInstruments, NZ) to study the signals, excerpts were extracted from the predefined periods. These were an hour prior to surgery, an hour after surgery, an hour after extubation, and an hour at least 12 hours after surgery, all containing the least amount of noise. Sophisticated noise removal of biomedical signals is a challenging problem in biomedical engineering. This was out of the scope of this thesis, and limited to simpler methods. *Raw* data were recorded without the use of commercial monitors, where analog and digital filter specifications are hard to come by, in order to maintain full control of the frequency content. To further combat the noise, even shorter periods were extracted, but with a length of at least 17 minutes corresponding to $1/(17 \cdot 60)\text{s} \approx 0.001$ Hz, from the definition of temporal frequency

$$F = \frac{1\,[cycle]}{T\,[period\,in\,seconds]}. \tag{7.1}$$



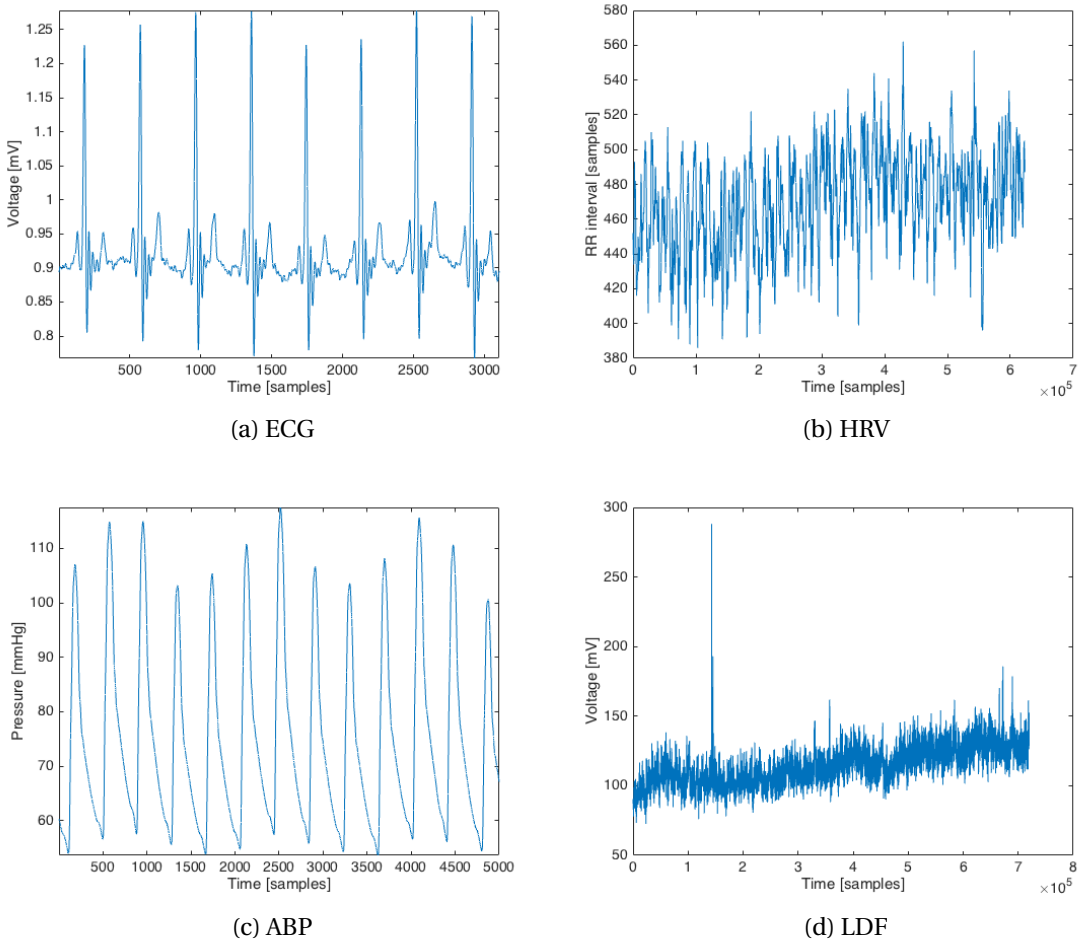
(a) ECG without noise

(b) ECG with noise

Figure 7.2: ECG signal excerpt without (a) and with (b) noise artifacts.

Figure 7.2 shows short excerpts of recorded ECG signal with and without noise due to movement artifacts. Also included are even shorter excerpts magnified inside the figures, showing the actual signal waveform. The time-varying ECG signals were filtered using a 1 dimensional digital *infinite impulse response* (IIR) lowpass *Butterworth* filter, from the Signal Processing Toolbox™, with cutoff frequency at 15 Hz. An even shorter excerpt of the filtered ECG recording is displayed in Figure 7.3 (a), clearly showing several of the repeating periodic *PQRST cycles*. However, the filtering process did not remove all the low frequency movement artifacts which overlapped in the frequency domain with the desired signal information below 2 Hz. Therefore excerpt were chosen to exclude most of the noise, for all three signals simultaneously, to keep signal excerpts synchronous. The ECG signal from one patients was heavily distorted, and erased from the further analysis. From these filtered ECG signals, the HRV signals were extracted. The ECG signals were used instead of ABP because the R peaks were sharper than the

systolic pressure peaks, hence producing better time localization. The R peaks were located in each signal using the Signal Processing Toolbox™, and the subsequent RR time intervals constituted this new signal. As such, the HRV signals were resampled with an irregular sampling frequency, at approximately 1 Hz which is also the average normal HR. The R peak detection parameters were selected by inspection. The signals were mean centered, and peak prominence was set to 1×10^{-4}V = 10 mV off the baseline. Also a 200 sample minimum lag between was peaks were selected, equivalent of max pulse of 120 for 400 Hz sampling frequency of ECG. The HRV computed signal from one patient is shown in Figure 7.3 (b). The ABP signals contained less noise, and were not preprocessed. An excerpt in shown in Figure 7.3 (c). The LDF signals contained a great number of low frequency outliers overlapping the frequency range of interest. Noise removal proved difficult, and were adequately resolved employing a manual solution due to the low number of signals. Outliers were detected by normalizing each signal against its maximum value, and then detect peak values over a certain threshold using the Signal Processing Toolbox™. The outlier peaks were replace with a part of signal before, instead of being set to zero or mean value, all of which resulting in manipulated "wrong" solutions. But this way, the outliers did not introduce a significant power overlapping in the frequency domain. The LDF signal from one patient in shown in Figure 7.3 (d), including one major outlier.



(a) ECG

(b) HRV

(c) ABP

(d) LDF

Figure 7.3: ECG short excerpt showing cardiac cycle (a), HRV signal from one patient computed from ECG signal (b), ABP short excerpt showing continuous BP with systolic maximum and diastolic minimum pressure (c), and the LDF signal from one patient (d).

7.2 CWT

The CWT signals of the biomedical time signals were computed using a *fast Fourier transform* FFT based function from the Wavelet Toolbox™. The FFT is an algorithm for efficiently computing the *discrete Fourier tranform* (DWT)[1], but CWT can be computed by using wavelets instead of sinusoids [56]. To simplify the calculations, and avoid running out of memory, the ABP and LDF signals were downsampled to sampling rate 10 Hz, still with a Nyquist rate above the frequency range of interest. This was managed by *decimation*, extracting only every 40th sample, hence $400Hz/40 = 10$ Hz, using the Signal Processing Toolbox™. To keep signals synchronized, and achieve identical CWT computation, all signals were of course resampled to the same sampling rate. The HRV already had a low sampling rate, albeit irregular, and the average was calculated to

[1]DWT simple corresponds to the FT of discrete values, using sums instead of integrals in (3.1) and (3.2).

facilitate the CWT analysis. The Morlet wavelet was tested due to previously reported popularity for analysis of complex signals. The resulting CWT waveform of both the ECG and ABP signals revealed a peak for the HR frequencies around 1 Hz, and the Morlet wavelet was therefore used for the rest of the analysis. The center frequency of the Morlet wavelet was set accordingly to wavelet theory, $\omega_0 = 5$. The final frequency band to be scrutinized was set to 0.004-2 Hz. The corresponding scales for wavelet calculation were given by the frequency-scale relationship in (3.14) from wavelet theory. Also, for frequency resolution, 32 voices were selected per octave. The calculations resulted in a time-frequency domain, or scalogram, with time localization for power variations. Starting out working on this thesis, the exact medical study to be conducted at the St. Olav's Hospital were not fully decided, lingering between two options. As such, the CWT was selected for its ability to produce time resolution, and the graphical representation property producing plots in the work of [14]. However, based upon the chosen medical study presented in the previous section, it was decided to only work in the frequency domain for the remaining analyses. This was calculated by averaging the absolute value of the complex power coefficients computed for all time instances for a given frequency, and repeated for all the frequencies. This also greatly reduced the number of data samples. The CWT was also chosen and kept for the further analyses because of the great resolution of low frequencies, and was thus well suited for representing the biomedical signals in this study. A natural logarithmic scale were used for the frequency content in the visual scalograms due to the resolution variation across the spectrum. The CWTs from each patients were normalized for comparisons between the patients. This was performed by finding the greatest global power value for each patient, looking in all the four periods, and dividing each sample by this maximum. In context of the machine learning analyses yet to be presented, the CWT is regarded as feature transformation processing, going from the original recorded time domain observations to the average scalogram frequency domain observations. CWT from this point on will refer to the average CWT of the biomedical signals, and thus only contain frequency content of average power coefficients.

7.3 PCA

PCA feature extraction was performed using the Statistics and Machine Learning Toolbox™. Each frequency sample from the CWT analysis were treated as a feature, and thus the CWT signals equalized a feature vector. Going back to the example data in Figure 5.5, this could then be though of as originating from two peaks in a frequency spectrum being used as features, thus creating the scatter scores for several observations. CWT for each of the three biomedical signals were computed for four periods per patients, for nine patients, thus totalling 36 independent feature vector observations. These vectors were combined as column vectors, producing a feature input matrix $\mathbf{X}$ per biomedical signal. As a consequent, each row was then observations of the same feature, treated as one dimension. These dimensions were kept original, but also normalized against the corresponding standard deviation using it as *divisor* for each variable producing new normalized *quotients* vectors. Thus PCA were conducted on the original and normalized

data, and for both cases, the PC coefficients and PC scores were computed. It is reported that PCA should not be used in settings, often biomedical, with fewer observations than feature dimensions [92]. Since the number of frequency samples, or feature dimensions, were greater than the number of observations, the PCA was also done on only every tenth frequency sample in order to have less features than observations. The PC scores vectors were extracted as feature vectors and used for feature extraction before being utilized in supervised learning models, to be explained in the next section. A threshold was set to 95%, extracting PC vectors until they together explained at least this amount of the variance in the original CWT data input. This was computed by first finding the total variance, given by the trace of the diagonal eigenvalues matrix, and the variance of PC coefficient vectors, or eigenvectors with associated eigenvalues indicating the variance for that vector. The PC scores were again found from the PC coefficients. For data visualization, the scores, or the rotated values in the original feature space, where scattered in a plot for the PC coefficients yielding the greatest variance. This was done in a supervised manner, using a priori knowledge with a class label vector for each observations, in order to differentiate between the classes in the PC feature space when looking for a clustering effect. Keep in mind, the PCA itself was still strictly unsupervised. The plotting was done either in a two dimensional PC plane or in a three dimensional PC space, depending on the variance explained by the PCs. When total explained variance by the first two PCs were above 90%, these two were scattered against each other using the Statistics and Machine Learning Toolbox™. When the total variance was below 90%, the third PC was also included and plotted using a function from [116]. The centroid from observations for each class using Euclidean distance was also computed using the k-means algorithm in the Statistics and Machine Learning Toolbox™, and used to differentiate between classes visually. PCA were performed three times, one for each signal. The feature vectors form the three signals could have been conjoined, but this would further increase the dimensionally while still only have 36 observations. Appending observations from each signal as new observations would result each dimension having observations of different scaling, even though all CWT data is given in power.

7.4 Classification

The classification procedure included feature processing, classification, validation and model tuning. The original CWT data for all the signals were of relative high dimensionality, with more features than observations. Besides using PCA, other feature processing methods were conducted. First, the frequency band 0.004-2 Hz was divided into smaller subbands, thus the CWT average coefficient feature vectors for each observation were divided into smaller vectors. The subband selection was based on the work in [15], with each subband intended to represent different physiological mechanisms as explained in Section 1.3. The following limits were used

- Subband 1: 0.004 – 0.0095 Hz
- Subband 2: 0.0095 – 0.021 Hz
- Subband 3: 0.021 – 0.052 Hz
- Subband 4: 0.052 – 0.15 Hz

- Subband 5: 0.15 – 0.6 Hz
- Subband 6: 0.6 – 2 Hz

yielding approximately the same number of frequency coefficients in each of the six subbands. Feature construction were performed by computing new features from the original CWT feature vectors $\mathbf{x}_j$. For each subband, for all periods and all three signals, six new features were computed using the samples in that subband. The four statistical moments mean, variance, skewness and kurtosis were calculated using the corresponding sample estimate equations in section 4.2. Total power was calculated by squaring and adding all the power samples in each subband $\mathcal{S}_b, b = 1,\ldots,6$, given by

$$P_{bj} = \sum_{x_j \in \mathcal{S}_b} x_j^2 \tag{7.2}$$

Finally max peak was also found, by fist locating all power peaks in each subband using the Statistics and Machine Learning Toolbox™, and then select the largest. If no peaks were found, max peak was set to zero. At least some of these new features were expected to differentiate between the classes, based on previous work and reports of decomplexification and reduction of power associated with CABG surgery. The new features, from each subband and each biomedical signal were conjoined for each observations. The CWT computation and frequency plotting of time-varying biomedical signals, with subband division and mean, variance, power and peak value calculation, were formalized in a MATLAB application in the work of [117], and utilized in [14] and also intended for future usage.

The generated features were further analyzed by using feature selection filter methods, from the fact that these are not tied one specific classifier, and classification was performed for a pool of classifiers. This was done in order to reduce the dimensionality and increase the performance of the classifiers. For initial feature selection, the power features of each subband from HRV and ABP were evaluated with the H-test in the Statistics and Machine Learning Toolbox™. This was done in order to see if this feature actually differentiated before and after surgery, at a 1% significance level. The nonparametric test was selected after inspection of the distribution of the four classes by *histogram* plotting and the use of *Shapiro-Wilk parametric hypothesis test of composite normality* from [118], both revealing nonnormality. The power values from the four different classes were also box plotted. These showed low variation between classes *B*, *C* and *D*, i.e. data after surgery, which was also revealed by the PCA.The choice of four distinct periods was merely used as a mean to track development on fluctuations in the hours after surgery. As reported from previous paper, decomplexification is excepted to be restored after CABG surgery only several months later. Therefore, classification was performed also for multiclass situation with four original, but with main focus on the binary classification between before and after surgery. The three classes after surgery were merged into one class. For further efficient feature selection, the nonparametric ReliefF algorithm was chosen based on previous reported use and its multivariate approach. Since the number of observations were small, all were used to weight the features. As such all features where evaluated at once, based on the whole feature input

space. The algorithm was tested on the constructed features and the PCA features, for both binary and multiclass class labels, with different K values for the nearest neighbor search trying to obtain an optimal value. Since data came from 36 observations, the search space was set to $2 \leq K \leq 30$.

After obtaining all the new generated features, these and the PCA features were used to build and test different classifiers for both two and four distinct classes. As part of the validation of the classifiers, K-fold cross validation was chosen. It was decided to used six folds, thus dividing the 36 observations into six equally sized parts, each of six random and unique observations. Hence the classifiers were trained on 30 observations, and tested with the remaining six, repeated six times. Inside this loop, feature selection with the ReliefF algorithms was performed each time, using the optimal number of nearest neighbors depending on the features and number of classes. The features selected were dependent on the observations, randomized by cross validation, potentially changing the feature set for each training. The weight or importance factor for each feature was added with its previous weight for new classification function calls. This was done in order to study which features that were performing best overall, discriminating the classes well regardless of random observations from cross validation. The final weight score for each feature for all the different classifiers were visualized in a bar plot. The scores for each feature in each subband were also summed together, for the first five subbands for the HRV signal and for all six subbands of the ABP signal. Further, for the two class situation, oversampling was conducted. Since class A had less observations than the other classes combined, synthetic A class observations were created with the ADASYN algorithm to balance the classes prior to training each model. Some experiments with changing the classification cost function was also conducted, making minority class misclassification more expensive.

Several different classifiers were trained. For each training, only the features with positive weights from ReliefF were used in order to reduce dimensionality. Also, only the very best features (3 and 10) were used for training, in order to use less features than the 36 observations. For each classifier algorithm, all introduced in the theory part, there exist numerous individual tuning parameters to optimize performance beyond those already covered. For the binary decision tree training, different thresholds were used to control the tree depth. Performance validation was obtained for 4, 20 and 100 maximum number of splits. For splitting criteria the Gini index was used. For the discriminant analysis, linear and quadratic functions were trained. The latter was obtained for diagonal covariance matrices. The KNN algorithm had several tuning parameter, training for all possible combinations. The distances were set to Euclidean, cosine and Minkowski, all with equal and squared inverse weighed distances. Further, 1, 5 and 12 number of nearest neighbors were picked. As with KNN, SVM also had several parameters. Training was done for linear, Gaussian and polynomial (of order three) kernels. Different scaling parameters of the kernels were also used. These were set to 1, 5 and auto, the latter using a MATLAB built-in heuristic procedure to select the scale value. Lastly the box constraint was varied, using default values 1 and 10. For both KNN and SVM, which are distance-based methods, features were standardized in each dimension by mean

centering and dividing by standard deviation.

The performance measures were average from each of the six validation runs. Since the classifier performance, i.e. the expected prediction accuracy, varied greatly for each validation, and also when repeating the entire validation loop, this procedure were implemented as default. As such, each cross validation, or the six folds loop, was itself iterated 10 times (restricted from a higher value due to computer power capacity), thus totalling 60 validations. The final averaged overall prediction performance, given by the correct rate, sensitivity and specificity measures, were recorded for each classifier with all the different parameter tunings, and ordered by descending order. From here, the 10 best classifiers with given parameter combination was ordered by correct classification rate in a table. Further in this table were the specificity and sensitivity scores for the classifiers, which also included the highest score for these measures. The classification was also done for only the three and ten best ReliefF weighted features, in order to inspect the results when number of observations were greater that the total number of features. The five best results from this procedure was also presented in a table, ordered by the correct classification rate. And as for classification using all positive weighed features, the specificity and sensitivity was included, with the highest score for both present in the table. Finally the classification was done on the PCA scores features, from the three biomedical signals combined. The results were again presented in a table, with highest scores for each of the three performance measures.

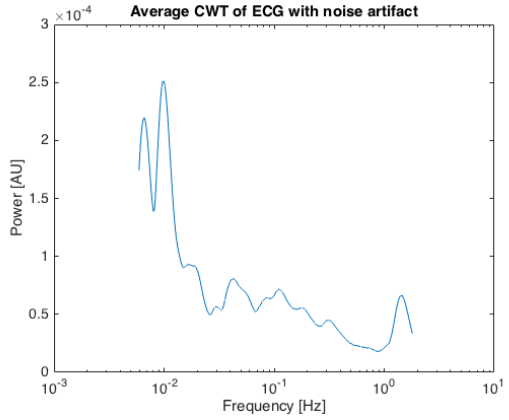
8 Results

All the results from the analyses described in the previous chapter are presented here. First the computed CWT scalograms are shown, with some selected excerpts and the impact of noise removal. Then follows the PCA with scatter plots and class clustering inspection. Lastly the classification and the adjoined feature selection and validation results are included.

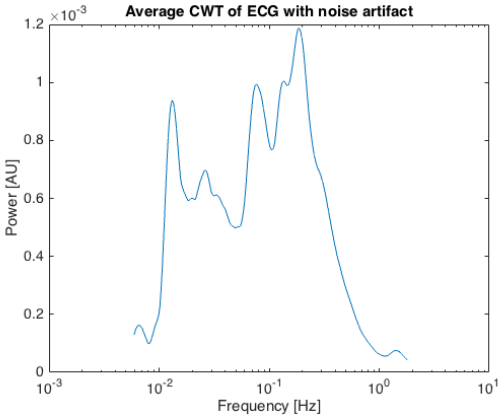
8.1 CWT

The CWT power density in arbitrary units [AU] of all three biomedical signals, for the four predefined periods and for all patients were computed using the Morlet mother wavelet. It resulted in 283 average power frequency samples, and thus 283 original features or dimensions for the following analyses. For the ECG and ABP signals, the CWT revealed the HR peak around 1 Hz, which is shown for ECG in Figure 8.1 (a). This plot also shows power in lower frequencies due to physiological mechanisms as previously reported. Figure 8.1 (a) and (b) correspond to Figure 8.1 (a) and (b), respectively, and clearly illustrate the influence of low frequency movement noise artifact in the ECG signal on the CWT computation. The noise overlaps with frequencies of significance and is difficult to remove. As a result excerpts without these noise artifacts were selected for the CWT analysis, since advanced noise removal was out of the scope for this thesis.

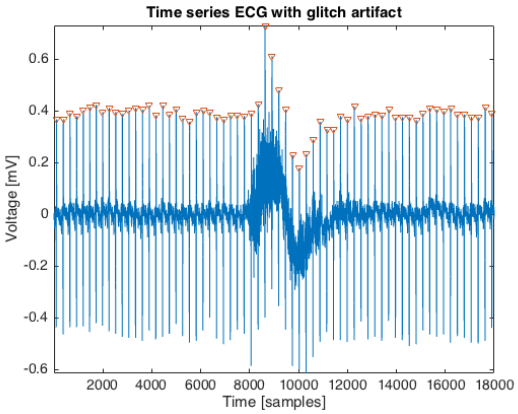
However, the HRV signals were computed from the ECG signals by R peak detection. This worked even though the ECG signals contained some artifact, like short bursts or glitch artifacts, as seen in Figure 8.1 (c). But, by inspection of more distorted ECG signals, the R peak detection introduced some erroneously time localizations as seen in Figure 8.1 (e). However, most of these noisy segments were removed prior to HRV extraction. From the time values localized by the red markers on top of R peaks, RR intervals were computed for two and two subsequent R peaks, representing the variability in HR. In Figure 8.1 (d) and (f), the corresponding CWTs of the HRV signals from the ECGs with a small glitch and severe noise artifact, respectively, are shown. The frequencies are strictly lower than 1 Hz, since it represent the variation of HR which again variates around 1 Hz. As in the CWT of the ECG signal, lower frequencies with constant power and distinct peaks are observed, also in compliance with previous findings.



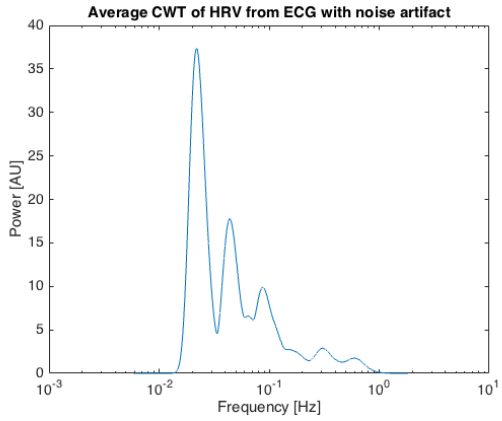
(a) CWT of ECG without noise



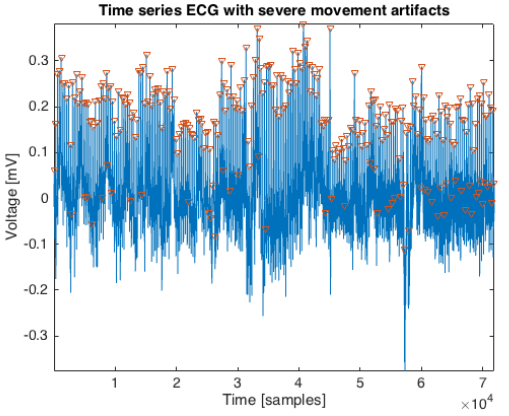
(b) CWT of ECG with noise



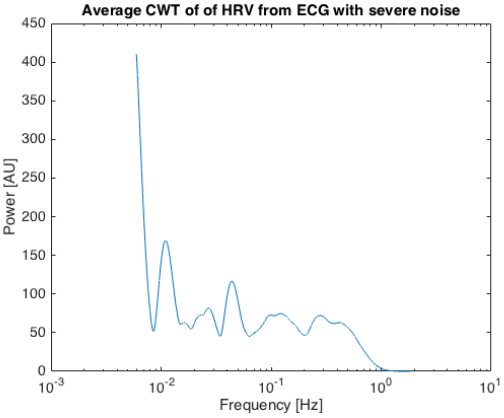
(c) HRV detection



(d) CWT of HRV



(e) HRV detection

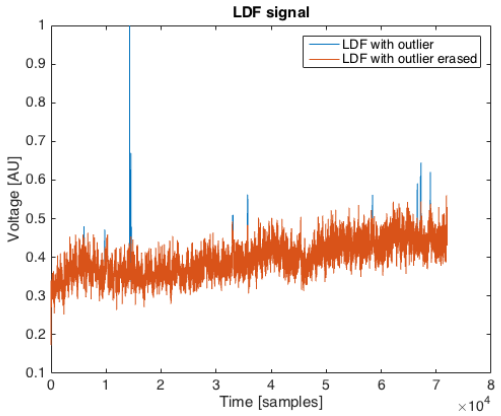


(f) CWT of HRV

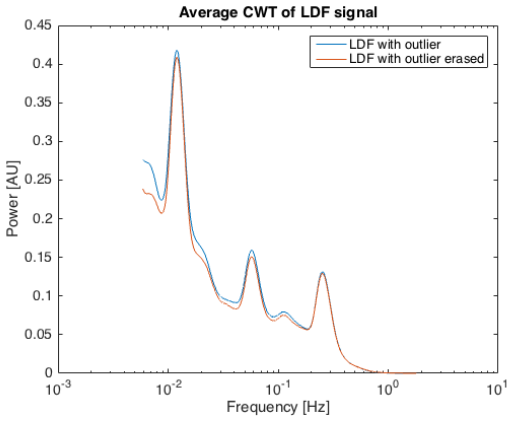
Figure 8.1: CWT of ECG signal excerpt without (a) and with (b) noise artifacts. HRV detection of ECG with glitch artifact (c) and corresponding CWT (d).

The LDF signals contained some outliers, some signals more than others, due to movement noise. The outlier peaks were detected by normalizing the signals, hence the arbitrary units, and then erase peaks over a certain threshold decided manually for

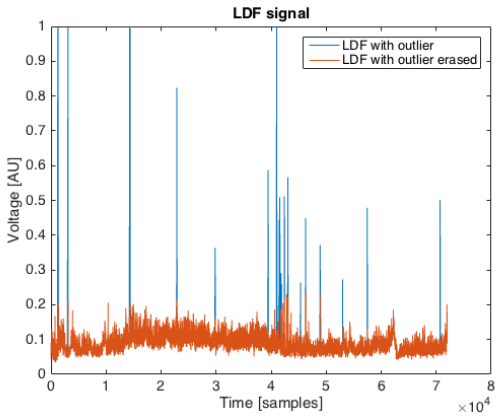
each signals. Figure 8.2 (a) shows the LDF signal from one patient before and after erasing one major outlier, and and the corresponding CWTs are shown in Figure 8.2 (b). From this it is clear that few outliers do not affect the power spectrum significantly. However, Figure 8.2 (c) shows another LDF signal also before and after outliers removal, but containing several high amplitude outliers. Here the corresponding CWTs in Figure 8.2 (d) differ greatly, showing how the low frequency power of the outliers affect the CWT signal. The signal shapes are to some extend identical, except power overall is lower when outliers are erased. Also some peaks become more prominent, because power is computed relative to the maximum frequency power found in the original time signal. This LDF signal had no peak around 1 Hz indicating the HR, although this was shown for LDF signals in work of [15].



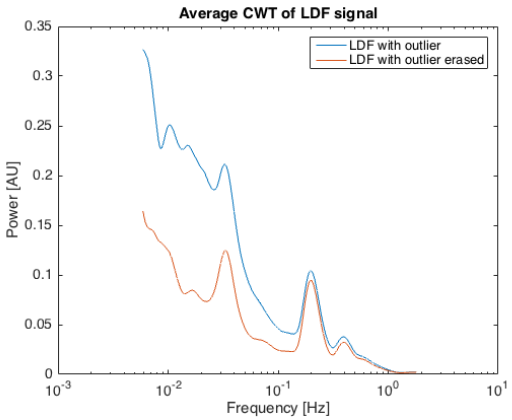
(a) LDF with and without some outliers



(b) CWT of LDF with and without some outliers



(c) LDF with and without many outliers

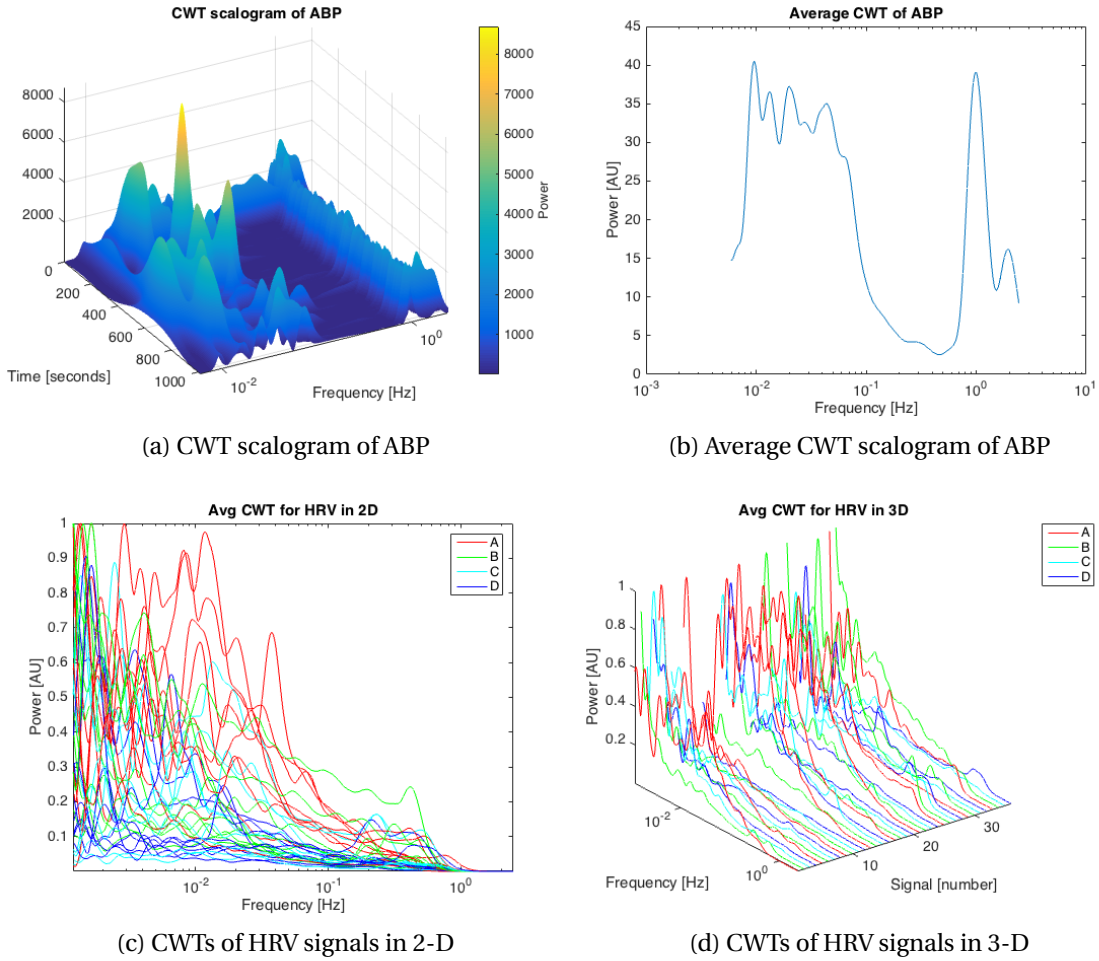


(d) CWT of LDF with and without many outliers

Figure 8.2: LDF signal with and without erasing one major outlier (a) and the corresponding CWTs (b). LDF signal with and without erasing many major outlier (c) and the corresponding CWTs (d).

The time-frequency scalogram of an ABP signal is visualized in Figure 8.3 (a). The scalogram shows the time localization of the frequency content, and how this varies across

the time the signal is recorded. The HR frequency at 1 Hz is persistent throughout the signal, while the lower frequencies have greater variance. The corresponding average CWT of the ABP signal is shown in Figure 8.3 (b). Both plots in this Figure shows how the average CWT is computed, and resembles a cross section of the time-frequency scalogram. The HR frequency is nearly constant, resulting in average high power. The lower frequencies have high maximum power, but with greater variance, and thus the average power approximately equals that of the HR.



(a) CWT scalogram of ABP

(b) Average CWT scalogram of ABP

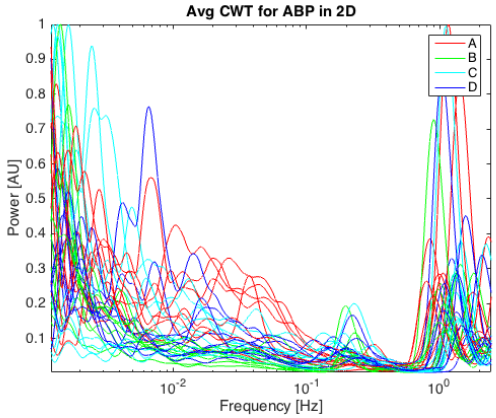
(c) CWTs of HRV signals in 2-D

(d) CWTs of HRV signals in 3-D

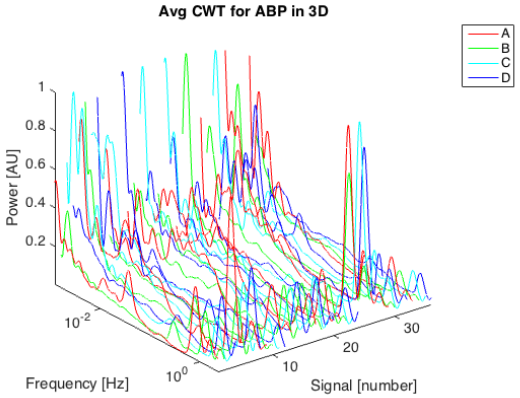
Figure 8.3: CWT scalogram time-frequency domain of a ABP signal (a) and the corresponding average CWT scalogram frequency domain (b). CWTs of all HRV signals plotted in two dimensions (c) and three dimensions (d).

The CWTs of all the signals were plotted in two and three dimensions for initial inspection of the waveforms and power distribution. Each plot contained four CWT signals per patients, one for each predefined period, for nine patient thus totalling 36 CWT signal, i.e. one for each observation. The CWTs for the different time periods were labelled as classes, *A* referring to before surgery, *B* to right after surgery, *C* to after extubation and *D* to at least 12 hours after surgery. All the CWTs for all three biomedical signals

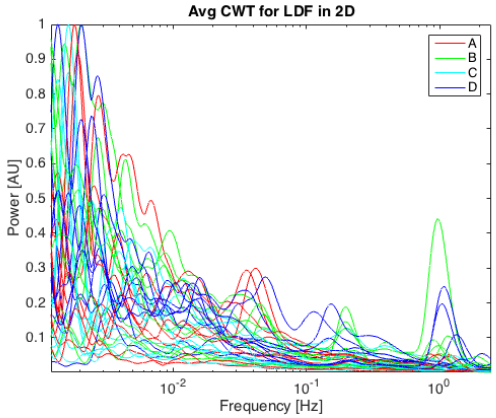
revealed circulatory oscillations in the selected frequency band 0.004 - 2 Hz, with constant power in the lower frequencies accompanied by distinct oscillatory peaks. The CWTs for the HRV signals are shown in Figure 8.3 (c) and (d). In general the power was higher in class *A* before surgery than in the classes after surgery, but still with exceptions. The CWTs of the ABP signals are shown in Figure 8.4 (a) and (b). Here also the power is clearly higher before surgery than after for lower frequencies. The CWTs of all the LDF signals are shown in Figure 8.4 (c) and (d). Some CWT signals record HR frequencies, but not all. Also the clear distinction between pre and post surgery power is not as pronounced here as for the CWTs from the two other biomedical signals HRV and ABP.



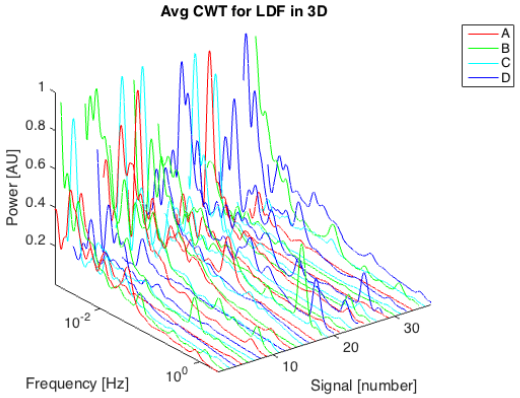
(a) CWTs of ABP signals in 2-D



(b) CWTs of ABP signals in 3-D



(c) CWTs of LDF signals in 2-D



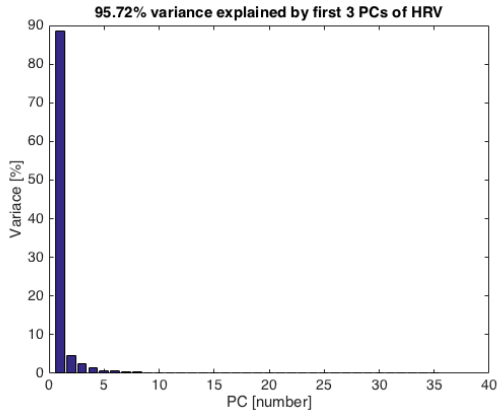
(d) CWTs of LDF signals in 3-D

Figure 8.4: CWTs of all ABP signals plotted in two dimensions (a) and three dimensions (b), and all LDF signals plotted in two dimensions (c) and three dimensions (d).

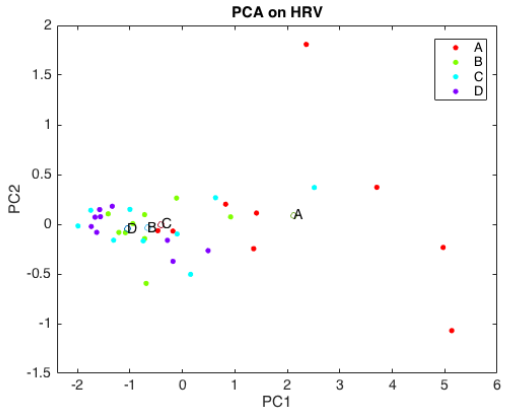
8.2 PCA

PCA was computed on the CWT signals for each of the biomedical signals at a time. The original input dimensionality was 283 features or frequency power samples, for

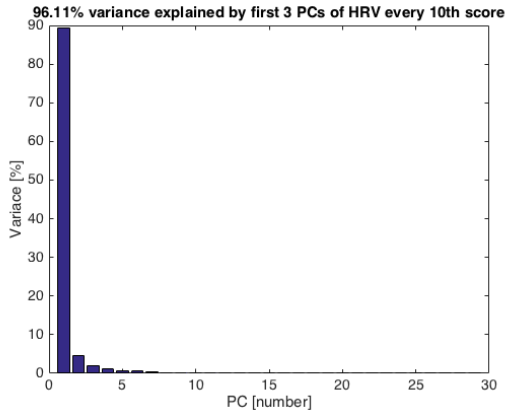
36 observations. PCA was performed both as a preprocessing step for the input to the supervised learning models, and for data visualization looking for clustering of the different classes in lower dimensionality. For the HRV signal input, the first three PC coefficient vectors explained over 95% of the total variance of the data, as can be seen in Figure 8.5 (a). The associated PC scores, or rotation of the original data in maximum variance directions, plotted in Figure 8.5 (b) show no clear clustering of the different classes. But class *A* data tends to be to the right, with greater within-class variance, while the other three classes for after surgery data are clustered together to the left, around origo. The same analysis as just described was also performed on the HRV data using only every tenth frequency power sample, in order to have fewer original features than observations. Starting out with 283 CWT features, this meant using 28 features. This result was almost identical compared with the previous HRV data, with only a slightly increased explained variance from 95.72% to 96.11% for the two PCs, and with same clustering, seen in Figure 8.5 (c) and (d).



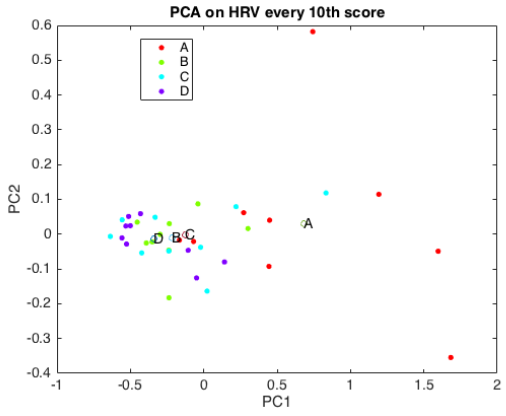
(a) Explained variance



(b) PCA score plot



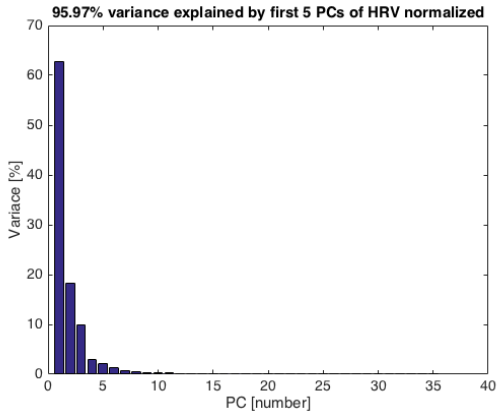
(c) Explained variance



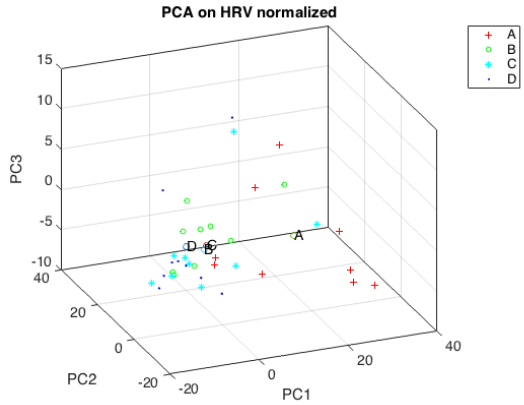
(d) PCA score plot

Figure 8.5: PCA with PC explained variance of HRV (a) and HRV every tenth sample (c), and PC score plot in two dimensions of HRV (b) and HRV every tenth sample (d).

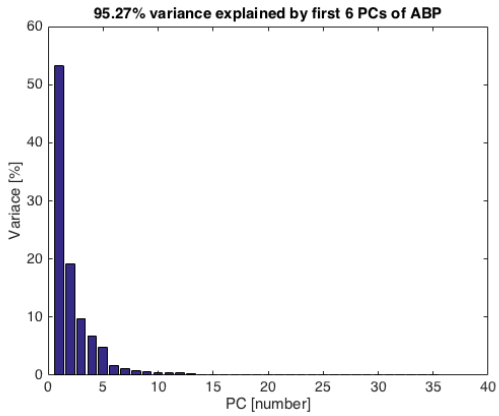
The HRV data was also normalized within each dimension. This however decreased the performance slightly, with the first four PCs needed to explain over 95% of the variance, seen in Figure 8.6 (a). The associated score plot was then done in three dimensions, adding the third most principal vector, because the first two did not explain at least 90% of the variance. Again no clear clustering appeared, as can be seen in Figure 8.6 (b). The distinction between class *A* and the rest persisted, but to a lower extent. PCA was also done on the two other biomedical signals. Results for the ABP signals showed that six PC vectors were needed to explained at least 95% of the original data, visualized in Figure 8.6 (c). The PC score plot was again done in three dimensions, indicating the same as all the previous explained results. There were no clear clustering of the four classes, but still a distinction between class *A* and the rest, illustrated in Figure 8.6 (d). Finally PCA was performed for the LDF signals. As seen in Figure 8.7 (a), 95.62% of the total variance in the LDF CWT input matrix was explained by the first six PCs. The PC score plot in Figure 8.7 (b) showed no clear distinction between any of the classes, with all data appearing in one big cluster.



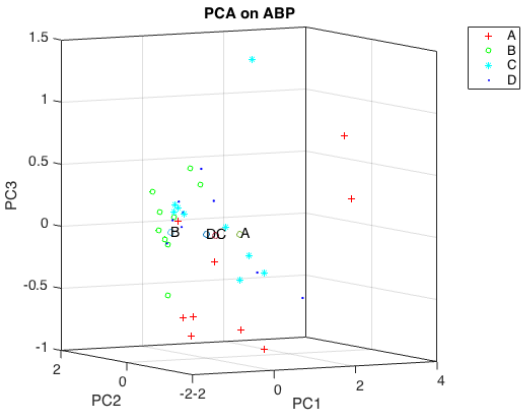
(a) Explained variance



(b) PCA score plot

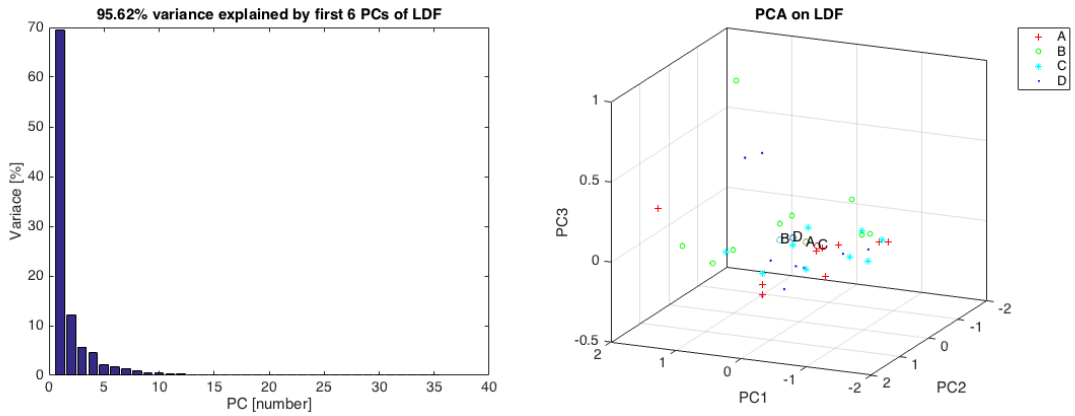


(c) Explained variance



(d) PCA score plot

Figure 8.6: PCA with PC explained variance of HRV normalized (a) and ABP, and PC score plot in three dimension of HRV normalized (b) and ABP (d).



(a) Explained variance

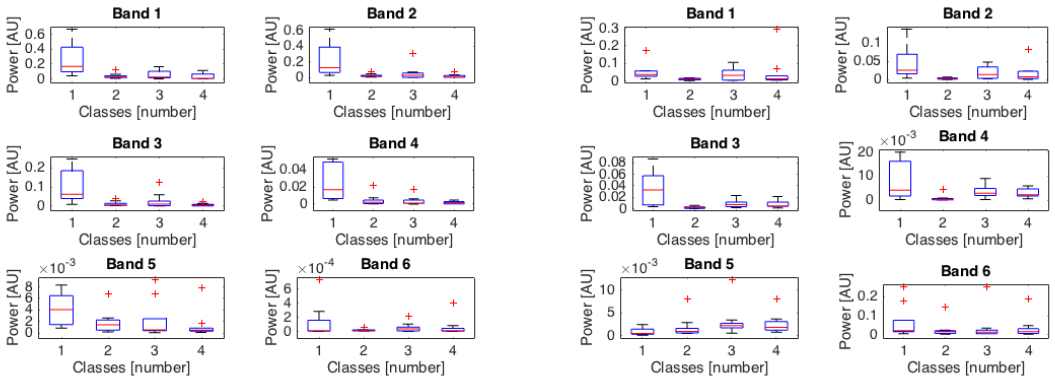
(b) PCA score plot

Figure 8.7: PCA with PC explained variance (a) and PC score plot in three dimension (b) of LDF.

8.3 Classification

The supervised classification process started with a continuation of feature extraction, besides the PCA. Six features from six subbands per signal resulted in 36 new features. This was computed for all three signals, with the resulting vectors concatenated for a total of 108 features or dimensions, for each of the 36 observations. The H-test statistical analysis of the power feature for the HRV and ABP signals are visualized by box plots in Figure 8.8. The red line in the boxes are the median value of the population, thus the power features from corresponding class. The blue line encloses the 25th *percentile*[1], the lower bound also called *first quartile* (Q1), and the upper 75th percentile also known as the *third quartile* (Q3). The *whiskers*, or line through the boxes, extend to the most extreme data point, not considered outliers which are the red crosses. The chi-squared statistic score and probability of class *A* before surgery and the other classes after surgery having equal means for power features, for HRV and ABP, are showed in Table 8.1. A returned value of probability less than 0.001 indicates that the test rejects the null hypothesis of equal means in the classes, at a 1% significance level. This is observed for all bands but 6 for HRV, and for band 2, 3 and and 5 for ABP. This last subband also indicate reversed trend, with power increasing after surgery.

[1]Statistical measure indicating the value below chich a given percentage of samples in a group of samples fall. For example the 50th percentile, or the second quartile (Q2), is the value below which 20% of the samples may be found.



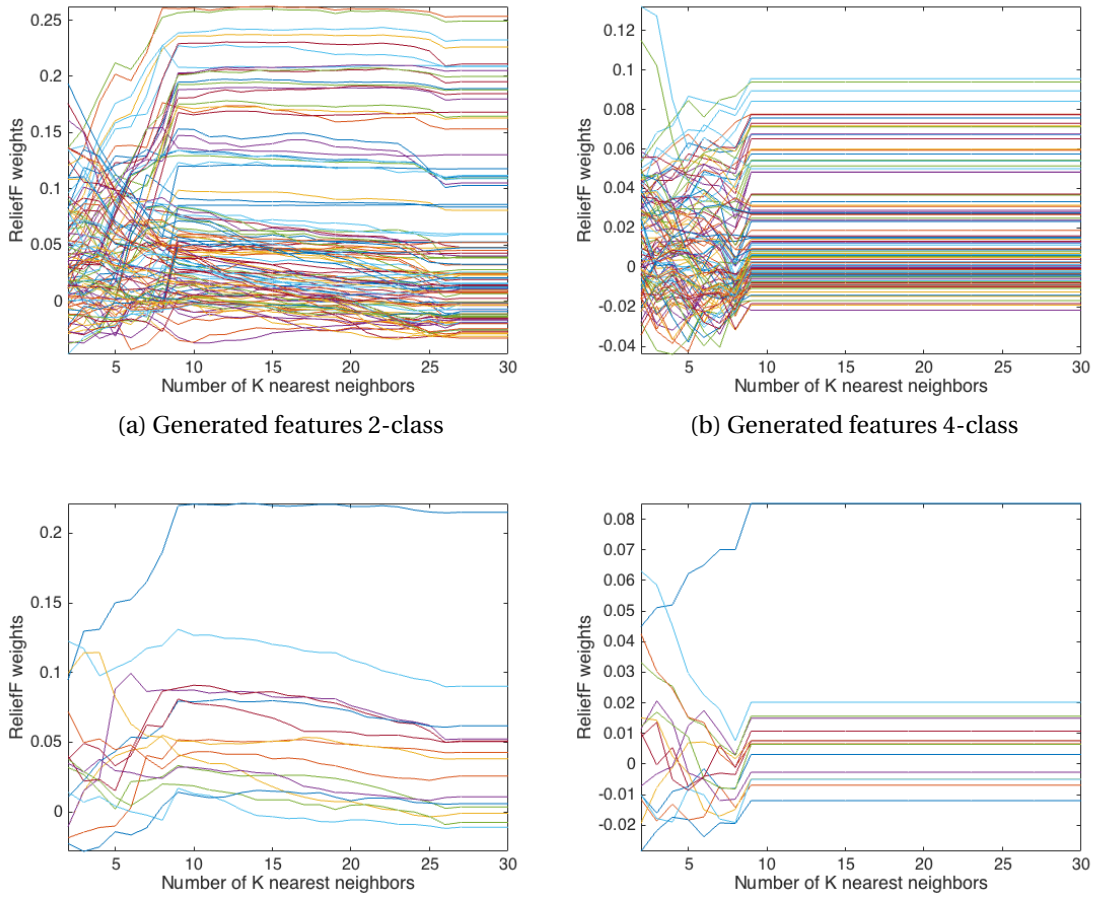
(a) HRV (b) ABP

Figure 8.8: Power box plot for each class and each subband, for the signals HRV (a) and ABP (b).

Table 8.1: Chi-squared statistic and probability of power observations in different subbands, before and after CAGB surgery, having equal means.

Band	**HRV**		**ABP**	
	Chi-sq	Pr >Chi-sq	Chi-sq	Pr >Chi-sq
1	13.4805	2.4106×10^{-4}	2.4107	0.1206
2	13.7501	2.0881×10^{-4}	8.2246	0.0041
3	13.2135	2.7793×10^{-4}	8.8652	0.0029
4	15.1381	9.9924×10^{-4}	3.4037	0.0650
5	7.4077	0.0065	6.6336	0.0100
6	0.0163	0.8983	2.4107	0.1206

When varying the K number of nearest neighbors in the ReliefF algorithm, the computed weights for each feature also fluctuated before stabilizing. This was done for the generated, or constructed features, calculated in the different subbands, and the PC scores features from the PCA of CWT signals. For four different classes, the K value stabilized at nine nearest neighbors. For two classes, the weights stabilized later with $K = 27$. These results are visualized in Figure 8.9.



(a) Generated features 2-class

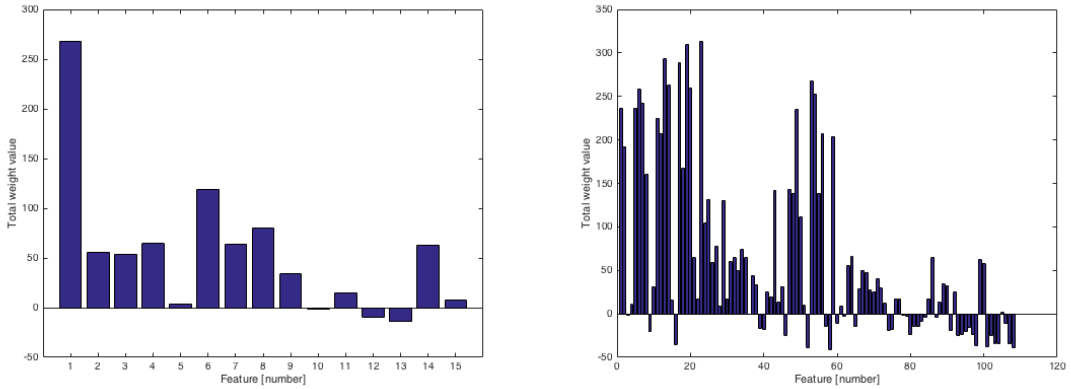
(b) Generated features 4-class

(c) PCA scores 2-class

(d) PCA scores 4-class

Figure 8.9: ReliefF computed weights with varying values of K, for 2-class situation with generated features (a) and PCA scores (c), and 4-class situation with generated features (b) and PCA scores (d).

The total weight score for each feature, from each feature selection process building, all the different classifiers, for two class situation, is displayed in Figure 8.10. In (a), the first three features are the PC scores from HRV, the next six are from ABP, and the rest are from LDF. As such the HRV and ABP signals had mostly positive weights, except the second PC from ABP. The LDF signal obtained three positive weighted PC score, but only one with a significant value, and the rest either close to zero or negative. From the 15 original features, 13 were positive weighted. However, only nine had a higher value, indicating discriminative property between the two classes. In (b), the first 36 are from HRV, the middle 36 from ABP and the last 36 are from LDF. The order is features mean, variance, skewness, kurtosis, power and max peak, for subband 1 up to subband 6, thus 36 features in total per signal. Also here, the features generated from the CWT of the HRV and ABP signals obtained mostly positive weights, while the LDF signal obtain lower and negative weights. Of 108 original features, only 72 features had final positive weighted score. The total weight score for each feature, for the HRV and ABP signals separately, is presented in Table 8.2.



(a) PCA scores

(b) Generated features

Figure 8.10: ReliefF computed weights for each feature, PCA scores (a) and generated features (b), for 2-class situation and K=30, added for each feature selection when building the different classifiers.

Table 8.2: Total weight score for each feature, for the HRV and ABP signals separately.

Feature	**HRV**	**ABP**
Mean	1212	270
Variance	932	174
Skewness	136	-19
Kurtosis	32	35
Power	1193	255
Max peak	754	399

The classification of the different classes was performed with different classifiers, each with several distinct parameter tunings. For four classes, the results were very poor. Performance accuracies went from slightly above 50% correct classification, and down to 25%. However, the main classification was conducted for two classes, before and after the surgery. Here the classifier parameters were tuned to optimize the performance. All the performance accuracies are the expected performances from the K-fold cross validation technique. Table 8.3 contains some of the results for classification with all positive weighted generated features from the ReliefF algorithm. It includes the ten highest correct classification rates. Also the highest specificity and sensitivity scores are included. None of the classifiers obtained perfect classification, with the SVM algorithm achieving overall best result. This classifier also obtained perfect score for the majority class, i.e. the after surgery observations, but with lower score for the minority class. The classifier's specific parameters are included in the parentheses, in the order kernel

function, kernel scaling, and box constraint. The KNN classifier achieved the same types of results as the SVM classifier, only slightly less accurate. The specificity performance were high, but at the same time resulting in lower sensitivity. The parentheses specifies the distance measure, number of nearest neighbors and distance weighting used by the classifier. The decision tree classifier had highest sensitivity score, still with some minority class misclassifications. Also this classifier had relatively high specificity, but not as good as SVM. The parentheses shows the maximum number of splits in the tree. The LDA and QDA classifiers generally performed poorly, with low correct classification rate. Trying to change the cost function in the different classifiers usually did not improve the result. Some obtained perfect sensitivity, but at the price of lower specificity, and thus overall performance.

Table 8.3: Classification performance measures for the ten best classifiers, using all positive weighted generated features from the ReleifF feature filtering,

Classifier (parameters)	**Correct rate**	**Specificity**	**Sensitivity**
SVM (linear, auto, 10)	0.9444	1.0000	0.7778
SVM (linear, auto, 1)	0.9222	1.0000	0.6889
SVM (polynomial, auto, 10)	0.9167	0.9630	0.6889
SVM (linear, 5, 1)	0.9111	0.9926	0.6667
Tree (20)	0.9000	0.9259	0.8222
SVM (Gaussion, auto, 1)	0.8889	0.9259	0.7778
KNN (Euclidean, 12, equal)	0.8778	0.9630	0.6222
KNN (Minkowski, 5, squared inverse)	0.8778	0.9481	0.6667
Tree (4)	0.8722	0.8963	0.8000
KNN (Euclidean, 12, squared inverse)	0.8722	0.9407	0.6667

Classification was also obtained in lower dimensionality, for fewer positive weights than all 72. This did not change much using either only three or ten features in total. The results from three feature classifications are shown in Table 8.4, containing the five highest correct classification rates. Also the overall highest specificity and sensitivity scores are included. The results were generally worse than that from using more features. All the classifiers showed the same tendencies, or ratio between specificity and sensitivity, as previously explained, but with lower classification accuracy. However, the QDA classifier improved, but still with low sensitivity score.
Table 8.5 shows the results from using the PCA scores as features. These were also exposed to the ReliefF algorithm, using only those 11 features with positive weighted score. The table contain the five highest correct classification rates, and the highest scores for specificity and sensitivity. The results are worse than that from using all positive weighted generated features, but slightly better than the other lower dimensionality case

Table 8.4: Classification performance measures for the five best classifiers using only the three best weighted generated features from the ReleifF feature filtering.

Classifier (parameters)	Correct rate	Specificity	Sensitivity
SVM (Gaussian, 5, 1)	0.8778	0.9185	0.7556
QDA	0.8611	0.9259	0.6667
Tree (100)	0.8444	0.8519	0.8222
SVM (Gaussian, auto, 10)	0.8167	0.8370	0.7556
KNN (Euclidean, 12, squared inverse)	0.8111	0.8296	0.7556

using only the very best generated features. Again the SVM classifier get high specificity, but at the expense of sensitivity. And also the QDA had increase performance for fewer features.

Table 8.5: Classification performance measures for the five best classifiers using all positive weighted PC score features from the ReleifF feature filtering.

Classifier (parameters)	Correct rate	Specificity	Sensitivity
KNN (cosine, 9, squared inverse)	0.9111	0.9556	0.7778
QDA	0.9000	0.9481	0.7556
KNN (cosine, 9, equal)	0.8833	0.9333	0.7333
SVM (Gaussian, auto, 1)	0.8722	0.9259	0.7111
SVM (polynomial, auto, 10)	0.8722	0.9407	0.6667
KNN (Euclidean, 5, equal)	0.8333	0.8370	0.8222
SVM (Gaussian, 1, 1)	0.7500	1.0000	0.0000

9 Discussion and Future Work

This chapter serves as the discussion and interpretation of all the findings presented in the preceding chapter. Also suggestions for future work are included, presenting other analysis methods that can be used to further scrutinize the biomedical signals analyzed in this thesis, for deeper understanding of the circulatory system.

9.1 CWT

The CWT was computed on downsampled signals, with the only difference from the original sampled signals being the loss of power amplitude, but with no relative change between different peaks. Using the Morlet mother wavelet, CWT was able to extract circulatory oscillations in the low frequency band $0.004 - 2$ Hz in all the three biomedical signals. The wavelet analysis proved constant power for lower frequencies and distinct peaks for the different subbands, both in line with previous work on this subject. For the ECG and ABP signals, the HR frequencies were visible, which by physiological knowledge and empirical evidence are always present in these signals. When selecting a mother wavelet, it is important that the wavelet represents the characteristics of the signal to be analyzed, in order to extract the correct information. Slowly varying, simple wavelets might not reveal all the frequency content of a complex signal. The Morlet wavelet has been reported as a popular choice for complex signals, and usage in medical settings. Together with the extraction of low frequency peaks and medical background, the Morlet wavelet was therefore assumed to produce the correct frequency content to be analyzed further with other methods.

The CWT of some of the ECG signals were found to contain much low frequency power concentrated around the center of the analyzed frequency band, exemplified in Figure 8.1 (b). The noise was in all likelihood due to motion artifact from patients' inevitable movements, since signals were recorded over an extended period of time. By inspecting the ECG signals first, extracting only excerpts without apparent noise, the corresponding CWTs instead had lower overall power, with clear HR frequencies. However, the HRV signals were computed from the ECG signals with R peak detection. This proved to work despite noise artifacts in the ECG signals, but not for vast noise distortion. As a result, removing the noisy segments made sure the HRV calculations through R peak

detection were valid. They could thus be used further in the other analyses in this thesis. As mentioned, ECG noise removal was done mostly manually, not with advanced filtering methods. A simple low pass filter was implemented, but only to eliminate high frequency noise outside the analyzed frequency band, to make R detection easier. HRV extraction was possible, even in instances with a reasonable amount of noise. Therefore the HRV signal has the potential to be a good feature in analysis of the heart, circulatory system or ANS, in recording settings involving movement. With simple detrending of the signal, the R peaks can more easily be detected correctly, as long as the removed frequencies are known to occur from movement noise. When recording the ECG signals for this study, sampling frequency was set to 400 Hz based on previous reports on the subject. However, new information has since been discovered. As reported by [119, 120], in ECG recordings of normal patients a 125 Hz sampling rate is sufficient. But for recordings from heart transplant patients and different simulated RR intervals, these signals should be digitalized at 1 kHz before being utilized for HRV signal extraction. This is again due to the decomplexification and power reduction between healthy and pathological patients. The reason for this not being researched earlier, was simply that extraction of HRV was not decided until later in the study progress. The resulting HRV signals still did resemble previous findings, with overall power clearly dropping from before surgery to after surgery, and circulatory oscillations reduced and decomplexified.

The ABP signals were not preprocessed. Inspection of the waveforms showed no clear noise artifacts, probably due to the recording of the signals. This was done invasive, with a catheter placed inside an artery and pressure directly converted to electric pulses. Thus the recording was not sensitive to movement artifacts. The CWT of the ABP signals also showed loss of oscillations after surgery, but not over the whole analyzed frequency band. The lowest frequency band, the respiratory frequencies and heart rate had power for all classes. When computing the CWT of the signals, this was averaged for all the time instances associated with different frequencies, visualized in the result for computation of CWT for ABP. This plot clearly shows that there are temporal differences for some of the frequency bands. While heart rate is persistent at all times, lower frequencies varies greatly and thus this dimension is lost in the performed analyses.

The LDF, however, did contain a substantial amount of noise. As with the ECG signal, excerpts with less noise were extracted by inspection. All the three signals were simultaneously recorded, and thus synchronous excerpts were used, finding segments with less noise for both ECG and LDF in particular. For many of the LDF signals the HR frequency peak was not present. This contradicts previous findings. The reason is probably because the signals contained much noise in the recordings. As can be seen in Figure 8.4 (c) and (d), there is extensive power in the lowest frequencies for many of the signals. Further from Figure 8.2 (d), this power is reduced when outliers are erased. As such, this power might be from noise not removed with the outliers detection and subsequent erasure, but rather embedded in the signal itself. The CWT scalogram seems to be given in relative power to the maximum. When the frequency band is divided in many voices, power is concentrated in lower frequencies. Therefore low frequency noise will prohibit any significant peak for the higher HR frequencies. LDF is known to be very sensitive

to movement noise. Previously in [15], the HR frequency and lower frequencies were detected as distinct power peaks with CWT using Morlet. But the analyzed signals were only recorded for 20 minutes. In this study, signals have been recorded for several hours, with much movement more likely to occur. Therefore, the movement sensitivity seems to be a challenge for LDF, especially when recording for long periods of time. Contrary to HRV and ABP, it seems to be no clear distinction between pre and post surgery CWT signals. But due to the noise influence just explained, no conclusions should be based merely on these results.

The CWT can be regarded as feature extraction from the original time domain samples of the biomedical signals, to the time-frequency scalogram domain. Further using only frequency information, by averaging the power coefficients over time, the resulting transformation compresses the original data significantly. Resampling the original data down to 4 Hz sampling rate, or four samples per second, by the Nyquist theorem, frequencies below half this rate will be preserved, i.e. the frequency band 0- 2 Hz. For at least 17 minutes signals, this requires at least $4 \cdot 60 \cdot 17 = 4080$ samples. The resulting CWT in this setting stores the frequency information in only 283 samples. The frequency content from the CWT is thus new features of the circulatory system, which was first assessed with the time-varying biomedical signals. CWT was performed based on previous reported work, extracting circulatory oscillations to better explain changes in the cardiovascular system after being put through challenges, e.g. diseases, medication and surgery.

9.2 PCA

The PCA and data visualization of the CWTs indicated the same results as the simple inspection of the CWTs of all the biomedical signals for each class. For both the HRV and ABP signals, the PC score plots indicated a distinction between pre and post surgery data. But the different classes were not separated entirely into individual clusters. Further, this meant that the information in the original 283 CWT samples were explained by maximum variance, resulting in only three PCs for HRV and six PCs for ABP. However, no new information or a clearer separation of the classes was revealed, as hoped when deciding to implement the analysis method. Still, the dimensionality reduction exercised by the analysis made the new computed features interesting for further use. These PC scores could be used as features in the supervised classification, hoping they would provide enough differentiation property between the classes for before and after surgery. Since the dimensionality was reduced, this could increase the performance of classification, especially for classifier algorithms prone to overfitting. For the ABP signal, a third PC was included to perform a scatter plot in three dimensions, since the first two PCs did not explain 90% of the total original variance. This PCs also had a relative high amount of variance, specifically 10%. As such, this dimension helped to differentiate the classes, as seen in the aforementioned 3-D plot. The PCA of the frequency content for the LDF signal showed no clear distinction between the classes, not even between before and after surgery. But as explained in the discussion for the

CWT result, the signals probably contained severe noise, and no conclusions should be based on the analysis of this signal. An interesting observation is that both ABP and LDF contained 95% of the original variance in the corresponding CWT signals, in the first six PCs. However, only the ABP showed any distinction between class *A* and the rest, suggesting differentiating power in PCA cannot be interpreted only by the amount of variance explained by the PCs.

The PCA was also conducted on the normalized HRV data, as suggested upon reviewing the analysis method. This did not improve the result, instead the performance decrease with reduced explained variance in the most principal components, and even less separation between the classes. This was probably because all the data had the same scaling and same units. All samples in each dimension were power samples from the CWT, and thus normalization did not improve the result. This should only be done to reduce the importance of features with different large scaling compared to the other features. Finally PCA was also done on a HRV signal containing only every tenth sample of the original HRV signal. As such, this new signal had less features than samples, since it has been suggested that PCA in HDLSS situations might not end in desirable results. This did not change the result significantly, hence the original information was still comprised in the downsampled signal. Also, it seems that PCA performance only deteriorate for higher dimensions, as the 283 CWT features are still relatively low, even for only 36 observations. Back to the PCA results in general, the mathematical transform assumes distribution of the exponential family. By the CLT, real life data tends to normal distribution for larger sample sizes. And larger number of observations will minimize the probability of errors and maximizing population estimation accuracy. For the HRV and ABP signal, PCA did resemble the information in the CWT of the two signals. The PC scores differentiated, to some extent, between pre and post surgery observations from only nine patients, and could be an interesting feature extraction technique for circulatory oscillations.

9.3 Classification

The classification process started with feature selection, in search of an optimal feature subset from the original feature set to improve classification performance. The power feature from each subband from the HRV and ABP signals were further scrutinized after the initial inspection of the CWT waveform. For the HRV signal, the box plots shows that there is a clear loss of power from before to after surgery for lower frequencies. This cooperate earlier findings, but has been further proved for several subbands. The H-test also shows that there is a statistical difference at a 1% level in all but one subband. This is the highest subband, but since HRV is sampled at approximately 1 Hz, the power in this subband bear no meaning. It has also been showed that power drops in several subbands for the ABP signal as well, with a statistical significance. This occurs in the middle subbands 2 and 3. An interesting finding is that the power in subband 5 have opposite trend, steady rising from class *A* before surgery to class *D* at least 12 hours post surgery. The box plots shows the difference between the four original classes, while

the statistical test was calculated for class A against the rest. Class B, C, and D shows no clear distinction, as anticipated due to previous findings stating power is not restored until several months after CABG surgery. Due to the uncertainty in the LDF signal, this signal was not used in the statistical test.

For efficiently evaluating all the features, and their discriminative property, the ReliefF feature filter method was employed. This was used to ensure that only important features were used to training the classifiers, to reduce the dimensionality and avoid curse of dimensionality and overfitting. To optimize the performance of this technique, the K parameter for number of nearest neighbors in the algorithm was varied. From this the value was found when the computed weights stabilized. $K = 30$ was used in all the classifications. For each training of a classifier, conducted several times because of the repeating nature of the cross validation scheme, the computed weights for each feature were continuously added. As such, the total weight score for each feature was calculated. From this, the importance of the features could be assessed. The bar plot of all the PC scores shows that the three first scores from HRV achieve positive weights, with decreasing value correlating with lesser explained variance. The middle six comes from ABP, with five positive weights of value well above zero. The last six from LDF are generally poor, with low and negative weights, The bar plot of all the 108 generated features shows that the latter 36, for each subband in the LDF signal, has low or negative weights. Again, this means that the features differentiate poorly between pre and post surgery data. But as pointed out several times, the LDF signal had considerable amount of noise, and the validity of this result is questionable. The 36 features in the middle represents the ABP signal. This shows high value for the features of the middle subbands, as seen from the CWT waveforms and box plots of the power feature. The first 36 represents the HRV signal, with generally high weight values, except the last subband. This however should not be considered, as HRV had almost no oscillations for the heart rate frequencies. The weights for each feature in each subband were summed for HRV in the lower five subbands, and for ABP in all the subbands. This shows that mean, variance, power and max peak are the features that generally discriminate well. Skewness and kurtosis, however, obtain low scores and should not be used in future work classifying circulatory oscillation changes.

The classification of the four original classes resulted in very poor accuracy, with performance equal of and worse than a trivial classifier, randomly classifying classes with an expected accuracy correct rate of 50%. This was understandable from the inspection of the CWT signals and the feature analysis, showing low variation between after surgery classes B, C and D. The main classification of two classes, before and after surgery, achieved generally high classification for some of the classifier algorithms. The SVM classifier obtained highest correct classification rate. This was because of the high specificity score, i.e. the classification accuracy of the majority class after surgery with most observations. The sensitivity score, i.e. the classification accuracy of the minority class for observations before surgery, was on the other hand much lower. This was the general result for all the classifiers, regardless of parameter tunings. Some experiments trying to change the cost function in the classifiers achieved a higher sensitivity score,

but at the expensive of specificity, and thus overall correct classification rate decreased. The decision tree algorithm obtained the highest sensitivity score, but not with perfect specificity, thus an overall lower performance accuracy than SVM. The discriminant analysis, both LDA and QDA, and the only parametric methods, performed poorly compared to the other more flexible nonparametric methods with less assumptions of the data.

The number of features were reduced further to train the classifiers using only the very best discriminating features according to the ReliefF method. As such, the number of features were lower than the number of observations, to ensure the curse of dimensionality was avoided. The number of features were varied between only the three best, and up to ten features, with no significant difference. However, this resulted in lower overall performance for all the classifiers, except the QDA model. This indicates that this classifier works better in lower dimensionality. Further, it shows the other classifiers, especially the SVM, work reasonable well in higher dimensions. The dimensionality is still relatively small, but must be seen in light of the number of observations, which is much lower. The classifiers were also tested in lower dimensionality for the PCA computed features, obtaining generally the same results as the previous low dimensionality case, and thus worse than using more features. This was reasonable, considering the visual display of the PC scores, and the computed weights with the ReliefF filter method. Still, the results are decent considering the low number of observations, and PCA does capture, to some extend, the difference in the circulatory system from before to after surgery.

The evaluation of the classifiers were performed using the K-fold cross validation. Because of the low number of observations, the value was set to $K = 6$. As such, the 36 observations were divided into equally sized folds of six observations. This method is recommended for classifier validation when data is scarce, as the case was for this study. Cross validation computes the expected classification performance, and not the true test accuracy. The K-fold loop is often itself repeated, to further average the performance. This was also employed here, mostly due to the great variance in the performance measures between each training. This was probably due to the different test observations used in each validation, randomly assign by the cross validation method. Even though the mean of the power feature was significantly different in the five lowest subbands for the HRV signal, there were outliers present from inspecting the CWT signals and the PC score plot. With few observations, only nine for the pre surgery class, the observations overlapping with the other class is hard to classify correctly. The ADASYN algorithm was implemented to overcome the imbalance between the classes, but this still did not improve the results significantly. This synthetic observations were introduced in the cross validation loop, and only used to train the classifiers. They were not used for training, since this would bias the classifier, with almost identical observations potentially being in the training and test sets at once. The variance in the classification results also gave no clear pattern for which classifiers were best suited for this data. The SVM mostly produced the best results, with the KNN and decision tree algorithms slightly behind, and the discriminant analysis being the worst. But the different results for the various parameters tunings gave no clear answers for which one

were the better. This seemed more dependent of the training and test sets division in cross validation.

This machine learning classification of observations from circulatory variables are a novel approach to study the differences from before and after cardiac surgery. Previous work is based on statistical tests of data from the time and frequency domain. This thesis extends previous work, extracting several features from the frequency band computed with the CWT. These are exposed to feature processing, and classified with several different algorithms to analyze the changes occurring in the circulatory system from CAGB surgery. The results show no perfect classification between observations in the two classes. When extracting the data, at lest 17 minutes of continuous recording had to be extracted. Some of these still included some noise artifact, which could be the reason for outliers in class A that overlap with class B, C, and D. However, results are promising for further testing. Obtaining more data is always better than further optimizing the current classification algorithms [6]. Training with small numbers of observations has been investigated in [121], with 5-25 independent observations per class. Although the classification models achieve acceptable performance, the *learning curve* can be completely masked by the random testing uncertainty due to the equally limited test observations. A learning curve displays a certain performance measure by varying the number of observations. Results from the aforementioned paper determined that 75-100 observations are usually needed to test a good, but not perfect classifier.

9.4 Future Work

Throughout the progress of this thesis, several choices of analyzing methods have been made. As a result, the are still many more ways in which the data from the medical study can be analyzed. First of, DWT can be employed when extraction frequency information from the biomedical signals. DWT supports many other wavelets, which might perform better than the Morlet wavelet for feature extraction. Also, DWT is not redundant, with high concentration of energy. Thus less frequency samples would be needed for each frequency subband, maybe eliminating the use of extensive dimensionality reduction. Both the CWT and DWT extract time localization of frequency variations. As seen in the computed scalogram for ABP in Section 8.1, lower frequencies varies greatly at different time instances. This dimension is an interesting property for further analysis of the circulatory system. By monitoring subject with ECG during different perturbations, the HRV can be extracted, and frequency changes with time localization can be mapped back to the different perturbations.

PCA was used in this study for feature extraction. This finds the maximum variance in the original data, which reduces the dimensionality and can be employed as features in supervised learning. The method did separate observations from different classes to some extend, but should be repeated on a greater number of observations. Also, adding Gaussian noise to the PC scores, as proposed in [122], could improve the performance of subsequent classification of small and unbalanced data sets. A special case

of PCA is the ICA, instead searching for independent components in the original data. This method could be used to extract features in medical signals for further classification. Also, using this method for different biomedical signals, hidden common features present in the signals could be extracted, for further knowledge of the circulatory system.

The use of machine learning has huge potential in medical settings, and is already a growing field. This is probably the future standard analysis method for finding patterns in medical signals, and subsequent implementation for supervised diagnosis of diseases. The classification performed in this thesis and the calculated features, has the potential to diagnose cardiovascular diseases associated with loss off complexity, and follow up CABG surgical patients to see if circulatory oscialltions are back to normal. Before this is a reality, work in this thesis needs to be repeated, but as stressed repeatedly, including more observations from new cardiac patients. In machine learning, usually more data beats clever algorithms. Also, more tuning parameters are possible, as well as testing other classifiers. The SVM generally performed best. This could be used further, with other feature selection methods as well. A wrapper method could be implemented together with the SVM classifier, to optimize performance.

10 Conclusions

The purpose of this thesis was to assess circulatory variables with different biomedical signals in 10 cardiac surgical patients, and investigate changes in low frequency oscillations from before CABG surgery and in the hours after. The HRV signal was extracted from the ECG recordings. Also, ABP and LDF signals were recorded. The CWT of the signals were obtained in the frequency range 0.004-2 Hz. But time localizations were omitted, instead computing the average frequency information over all time instances, to be further analyzed. Initial inspection of the CWT waveform showed HRV oscillations heavily decreasing after surgery. This was also observed in the ABP signal, but to a lesser extend, with oscillations steadily increasing for the respiratory frequencies after surgery. LDF gave no clear difference between pre and post surgical recordings. Excerpts of all signals were extracted in order for the recordings to obtain the least amount of noise. However, the LDF signals contained considerable amount of movement artifacts. The signals were included in the further analyses, but reveal no information of any importance.

PCA was performed on the CWT data of the three biomedical signals, in order to investigate the information in lower dimensionality. For the HRV and ABP signals, a small difference between pre and post surgical observations could be seen, but with no clear distinction and still overlapping data point. This was inspected by scattering the very most principal components, which included most of the variance in the original data. To further assess the discriminate power in these new extracted features, PC scores were used in the supervised classification.

Feature selection was performed prior to supervised classification. The frequency band computed for the original time-varying signals was divided into six subbands. In each band, several new features were computed. The power feature from the HRV and ABP signals were used in the statistical H-test. For HRV, this proved statistical significant difference in power from before to after surgery, with oscillations decreasing in all subbands, but the highest HR frequency range. The ABP also had statistical significant decrease in power for the middle subbands, and with opposite trend for the respiratory frequencies. ReliefF feature filtering were conducted on all the constructed features computed for each biomedical signal. Positive weighted features indicated discrim-

inate properly between the class for observations prior to surgery, and the class for observations post surgery. This was showed for most of the generated features from the CWT of the HRV and ABP signals. Mean, variance, power and max peak achieved high total weight score, and can be used for future classification of changes in circulatory oscillations.

Using these features, supervised classification with the decision tree, LDA, QDA, KNN and SVM was conducted. Performance measures were assessed using 6-fold cross validation, iterated 5 times. Correct classification rates were generally high, however no perfect classification was recorded. PCA was also used as features, obtaining decent results, but worse than the generated features from each subband. Overall, the SVM classifier obtained highest performance accuracy. But the decision tree and KNN algorithms resulted in highest sensitivity score, which were generally low. Sensitivity measured the classification accuracy of the minority class of observations before surgery, and the low accuracy is probably due to the low number of observations for this class. Furthermore, this very reason is also the most likely explanation for the modest accuracy results. With more observations of the discriminative features found in this thesis, higher performing classification of changes in circulatory oscillations should be possible.

A Manual for MATLAB code

MATLAB files with filename extension .m are included in the attached ZIP file *MATLAB_mathias_falk_master_thesis_2016.*zip. This contain an identically named folder with MATLAB functions and scripts used to performed the analysis presented in this thesis. All the files includes explanatory descriptions, accessible using the *help* function in MATLAB followed by the file name. First the script *initialize.m* is run to clear the workspace, command window and close figures, and also initialize global constants for the further analysis. The ten *dataextXX.m* files in the folder *dataext* are run for data extraction for each patients. This folder needs to contain all the MAT-files with data for each patients, obtainable from the shared folder online found at `https://studntnu-my.sharepoint.com/personal/mathiafa_ntnu_no/Documents/Master2016_mathiasfalk`. This data is also obtainable directly in the MAT-file *patientdata.mat*, in the same folder, with precalulated data. Then the script *dataProc.m* is run. This performs HRV extraction of ECG with the function *HRV.m*, downsampling of ABP and LDF signals, and noise removal of the latter signal with function *deloutliers.m*. From these three new signals, CWT is calculated with the function *CWTscalogram*. The new HRV signal and the resampled ABP and LDF signals are stored in the *patientdata.mat* file for future use. CWT signals are also stored, in the the *CWTdata.mat* file, which can also be obtain directly from the shared folder with precomputed data. For the main analysis, *featProc.m* constructs the new features in each subband, for each signal. The new features, and the CWT data for each signal is organized in matrices containing all observations. Here rows are observations, and columns are features or dimensions, thus the transpose mode compared to notation used in this thesis. Also the PCA is performed with the call of function *PCAcalc.m*, performing plotting and saving scores as feature in a matrix. The PC score plotting in 3-D is done with the function *gscatter3.m*. For classification, the script *classifications.m* are used. Here variables can be specified for the analysis. Variable *featureHyperSub* contain a matrix with all generated features per observation. *scoresPCA* contain all PC scores. Variables *classNum2* and *classNum4* contain vectors with integer number for two and four classes, respectively, used for class labels in supervised settings. The classifications are performed on all classifier combinations stated in the script, using the function *classify.m*. This function again uses the function *ADASYN.m* to perform oversampling in the two-class situation. The feature weights from feature selection are

then plotted.

For the other analyses, the remaining scripts and functions are used. *siganalyis.m* plots excerpts of all the signals, the R detection in ECG, and the CWT results with and without noise in ECG and LDF. *wavanalysis.m* plots the time and Fourier domains of the Morlet and Mexican hat wavelets. *featanalysis.m* plots the ReliefF analysis for several values of nearest neighbors in the feature weighting importance ranking, using function *rffplotK*. The power feature is box plotted in each subbands, and further statistical testing of this feature is done with the function *stattest.m*, itself checking normality of the feature with *swtest.m*. Finally *plotCWT* plots all the CWT signals for each observation, for each signal, in 2-D and in 3-D.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, 2nd ed., P. Bickel, P. Diggle, S. Fienberg, and U. G. andand Scott Zege, Eds., ser. Springer Series in Statistics. New York: Springer, 2009, pp. xi, 14–15, 18–22, 101–111, 150–151, 219, 459–470, 649, 656–658. DOI: 10.1007/978-0-387-84858-7.

[2] G. Casella, S. E. Fienberg, and I. Olkin, Eds., *An Introduction to Statistical Learning, With Applications in R*, ser. Springer Texts in Statistics. New York: Springer, 2013, pp. 1, 15–28, 39–42, 127. DOI: 10.1007/978-1-4614-7138-7.

[3] C. Bishop, *Pattern Recognition and Machine Learning*, M. Jordan, J. Kleinberg, and B. Schölkopf, Eds., ser. Information Science and Statistics. New York: Springer, 2006, pp. 12–24, 36, 38–46, 124–127, 179–199, 291–292, 32–326, 338–339, 424–425, 653–654.

[4] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, 2nd ed., O. Maimon and L. Rokach, Eds. New York: Springer, 2010, pp. vii, 149–170. DOI: 10.1007/978-0-387-09823-4.

[5] L. Hamilton, "Six novel machine learning applications," *Forbes*, 2014. [Online]. Available: http://www.forbes.com/sites/85broads/2014/01/06/six-novel-machine-learning-applications/#2714b82467bf.

[6] P. Domingos, *A few useful things to know about machine learning*, Department of Computer Science and Engineering, University of Washington, Accessed Jul. 21, 2016, Seattle, WA, 2012. [Online]. Available: http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf.

[7] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institue, Tech. Rep., 2011.

[8] B. Guo, D. Zhang, Z. Yu, Y. Liang, Z. Wang, and X. Zhou, "From the internet of things to embedded intelligence," *World Wide Web*, vol. 16, no. 4, pp. 399–400, 2013. DOI: 10.1007/s11280-012-0188-y.

[9] M. Chui, M. Löffler, and R. Roberts, "The internet of things," *McKinsey Quarterly*, 2010. [Online]. Available: http://www.mckinsey.com/industries/high-tech/our-insights/the-internet-of-things.

[10] K. R. Foster, R. Koprowski, and J. D. Skufca, "Machine learning, medical diagnosis, and biomedical engineering research - commentary," *BioMedical Engineering OnLine*, 2014. [Online]. Available: https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-13-94.

[11] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health Information Science and Systems*, vol. 2, no. 3, 2014. DOI: 10.1186/2047-2501-2-3.

[12] D. M. et al., "Heart disease and stroke statistics— 2015 update: A report from the american heart association," *Circulation*, vol. 131, no. 4, e156–178, 2014. DOI: 10.1161/CIR.0000000000000152.

[13] W. H. Organization, *Global status report on noncommunicable diseases 2014*, Key Facts Sheet. Accessed Aug. 21, 2016, 2014. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/.

[14] F. E. T. Axelsson, B. G. Lio, and N. K. Skjærvold, *Circulatory oscillations in post cardiac surgery patients*, Medical research paper, Norwegian University of Science and Technology, Faculty of Medicine, Department of circulation and medical imaging, Trondheim, Norway, 2016.

[15] M. Bracic and A. Stefanovska, "Wavelet-based analysis of human blood-flow dynamics," *Bulletin of Mathematical Biology*, vol. 60, no. 5, pp. 919–935, 1998.

[16] *Frontiers of Blood Pressure and Heart Rate Analysis*, ser. Studies in Health Technology and Informatics. Amsterdam, The Netherlands: IOS, 1997, vol. 35, ISBN: 978-90-5199-312-7.

[17] M. Varela, R. Ruiz-Esteban, and M. J. M. de Juan, "Chaos, fractals, and our concept of disease," *Perspectives in Biology and Medicine*, vol. 53, no. 4, pp. 584–595, 2010.

[18] T. G. Buchman, "Fractals in clinical hemodynamics," *Anesthesiology*, vol. 117, no. 4, pp. 699–700, 2012.

[19] B. Manor and L. A. Lipsitz, "Physiologic complexity and aging: Implications for physical function and rehabilitation," *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, vol. 45, pp. 287–293, 2013.

[20] R. Brunner, G. Adelsmayr, H. Herkner, C. Madl, and U. Holzinger, "Glycemic variability and glucose complexity in critically ill patients: A retrospective analysis of continuous glucose monitoring data," *Critical Care (London, England)*, vol. 16, no. 5, R175, 2012.

[21] Y. L. Ho, C. Lin, Y.-H. Lin, and M.-t. Lo, "The prognostic value of non-linear analysis of heart rate variability in patients with congestive heart failure–a pilot study of multiscale entropy," *PLoS ONE*, vol. 6, no. 4, e18699, 2011. DOI: 10.1371/journal.pone.0018699.

[22] K. K. H. et al., "Predicting survival in heart failure case and control subjects by use of fully automated methods for deriving nonlinear and conventional indices of heart rate dynamics.," *Circulation*, vol. 96, no. 5, pp. 842–848, 1997.

[23] N. Lakusic, D. Mahovic, P. Kruzliak, J. C. Habek, M. Novak, and D. Cerovec, "Changes in heart rate variability after coronary artery bypass grafting and clinical importance of these findings," *BioMed Research International*, vol. 2015, 2015. DOI: 10.1155/2015/680515. [Online]. Available: http://www.hindawi.com/journals/bmri/2015/680515/.

[24] C. Julien, "The enigma of mayer waves: Facts and models," *Cardiovascular Research*, vol. 70, no. 1, pp. 12–21, 2006. DOI: 10.1016/j.cardiores.2005.11.008.

[25] M. P. et al., "Power spectral analysis of heart rate and arterial pressure variabilities as a marker of sympatho-vagal interaction in man and conscious dog," *Cardiovascular Research*, vol. 59, no. 2, pp. 178–193, 1986.

[26] Y. Cheng, B. Cohen, V. Orea, C. Barres, and C. Julien, "Baroreflex control of renal sympathetic nerve activity and spontaneous rhythms at mayer wave's frequency in rats," *Autonomic Neuroscience: Basic and Clinical*, vol. 111, no. 2, pp. 80–88, 2004. DOI: 10.1016/j.autneu.2004.02.006.

[27] N. L. et al., "Outcome of patients with normal and decreased heart rate variability after coronary artery bypass grafting surgery," *International Journal of Cardiology*, vol. 166, no. 2, pp. 516–518, 2013.

[28] A.-H. Najmi and J. Sadowsky, "The continuous wavelet transform and variable resolution time–frequency analysis," *Johns Hopkins APL Technical Digest*, vol. 18, no. 1, pp. 134–137, 1997.

[29] G. Kheder, A. Kachouri, R. Taleb, M. ben Messaoud, and M. Samet, "Feature extraction by wavelet transforms to analyze the heart rate variability during two meditation techniques," in *6th ESEAS Internatinal Conference on Circuits, Systems, Elecronics, Control & Signal Processing*, (Dec. 29–31, 2007), Cairo, Egypt, pp. 374–378.

[30] V. N. Hegde, R. Deekshit, and P. S. Satyanarayana, "Heart rate variability feature extraction using continuous wavelet transform," *Journal of Medical Imaging and Health Informatics*, vol. 5, no. 5, pp. 898–906,

[31] A. Pachauri and M. Bhuyan, "Wavelet transform based arterial blood pressure waveform delineator," *International Journal of Biology and Biomedical Engineering*, vol. 6, no. 1, pp. 15–25, 2012.

[32] M. D. M. et al., "Blood pressure waveform analysis by means of wavelet transform," *Medical & Biological Engineering & Computing*, vol. 47, no. 2, pp. 165–173, 2009. DOI: 10.1007/s11517-008-0397-9.

[33] A. Harkat, R. Benzid, and L. Saidi, "Features extraction and classification of ecg beats using cwt combined to rbf neural network optimized by cuckoo search via levy flight," in *2015 4th International Conference on Electrical Engineering (ICEE)*, (Dec. 13–15, 2015), 2015, pp. 1–4. DOI: 10.1109/INTEE.2015.7416767.

[34] J. Kilby, G. Mawston, and H. G. Hosseini, "Analysis of surface electromyography signals using continuous wavelet transform for feature extraction," in *Advances in Medical, Signal and Information Processing, 2006. MEDSIP 2006. IET 3rd International Conference On*, (Jul. 17–19, 2006), 2006, pp. 1–4. DOI: 10.1049/cp:20060353.

[35] Y. Ozbay, "A new approach to detection of ecg arrhythmias: Complex discrete wavelet transform based complex valued artificial neural network," *Journal of Medical Systems*, vol. 33, no. 6, pp. 435–445, 2009.

[36] E. A. Maharaja and A. M. Alonsob, "Discriminant analysis of multivariate time series: Application to diagnosis based on ecg signals," *Computational Statistics & Data Analysis*, vol. 70, pp. 67–87, 2014. DOI: 10.1016/j.csda.2013.09.006.

[37] D. G. et al., "Automated diagnosis of coronary artery disease affected patients using lda, pca, ica and discrete wavelet transform," *Knowledge-Based Systems*, vol. 37, pp. 274–282, 2013. DOI: 10.1016/j.knosys.2012.08.011.

[38] F. Castells, P. Laguna, L. Sörnmo, A. Bollmann, and J. M. Roig, "Principal component analysis in ecg signal processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, p. 98, 2007.

[39] Y. C. Yeh, T. C. Chiang, and H. J. Lin, "Principal component analysis method for detection and classification of ecg beat," in *Bioinformatics and Bioengineering (BIBE), 2011 IEEE 11th International Conference on*, (Oct. 24–26, 2011), 2011, pp. 318–322. DOI: 10.1109/BIBE.2011.59.

[40] V. Kalpana, S. T. Hamde, and L. M. Waghmare, "Ecg feature extraction using principal component analysis for studying the effect of diabetes," *Journal of Medical Engineering & Technology*, vol. 37, p. 2, 2013. DOI: 10.3109/03091902.2012.753126.

[41] D. Napoleon and S. Pavalakodi, "A new method for dimensionality reduction using k- means clustering algorithm for high dimensional data set," *International Journal of Computer Applications*, vol. 13, no. 7, pp. 41–46, 2011.

[42] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007. DOI: 10.1093/bioinformatics/btm344. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/23/19/2507.abstract.

[43] I.-H. Lee, G. H. Lushington, and M. Visvanathan, "A filter-based feature selection approach for identifying potential biomarkers for lung cancer," *Journal of Clinical Bioinformatics*, vol. 1, no. 11, 2011. DOI: 10.1186/2043-9113-1-11. [Online]. Available: https://jclinbioinformatics.biomedcentral.com/articles/10.1186/2043-9113-1-11.

[44] A. Heshmati, T. Moallem, R. Amjadifard, and J. Shanbehzadeh, "Relieff-based feature selection for automatic tumor classification of mammogram images," in *2011 7th Iranian Conference on Machine Vision and Image Processing*, Tehran: IEEE, pp. 1–5. DOI: 10.1109/IranianMVIP.2011.6121616.

[45] Y. W. et al., "Tumor classification based on dna copy number aberrations determined using snps arrays," *Oncology Reports*, vol. 5, pp. 1057–1059, 2006.

[46] Wikipedia, *Circulatory system — wikipedia, the free encyclopedia*, [Accessed Aug. 20, 2016], 2016. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Circulatory_system&oldid=734940371.

[47] K. M. V. D. Graaff and R. W. Rhees, *Schaum's Easy Outline of Human Anatomy and Physiology*, P. B. Wilhelm, Ed., ser. Schaum's Outline. New York: McGraw-Hill, 2001, pp. 104–105, 108, 115–116, 125–126. DOI: 10.1036/0071-406069.

[48] B. Elling, K. M. Elling, and M. A. Rothenberg, *Anatomy and Physiology Paramedic*. Burlington, MA: Jones and Bartlett Learning, 2003, pp. 35–50.

[49] K. Najarian and R. Splinter, *Biomedical Signal and Image Processing*, 2nd ed. Boca Raton, FL: CRC Press, 2012, pp. 171, 173–174, 176–178, 181–182.

[50] C. S. G. Perera, *Multiple Sensor Data Analysis, Fusion, and Communication for ULTRASPONDER*. Trondheim, Norway, 2009, pp. 12–13, Master's Thesis in Electronics. Norwegian University of Science and Technology Department of Electronics and Telecommunications.

[51] Wikipedia, *Electrocardiography — wikipedia, the free encyclopedia*, [Accessed Aug. 20, 2016], 2016. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Electrocardiography&oldid=735927626.

[52] T. G. P. et al., "Recommendations for blood pressure measurement in humans and experimental animals, Part 1: Blood pressure measurement in humans: A statement for professionals from the subcommittee of professional and public education of the american heart association council on high blood pressure research," *Circulation*, vol. 111, no. 5, p. 700, 2005. DOI: 10.1161/01.CIR.0000154900.76284.F6. [Online]. Available: http://circ.ahajournals.org/content/111/5/697.

[53] S. R. et al., "Accuracy of invasive arterial pressure monitoring in cardiovascular patients: An observational study," *Critical Care*, vol. 18, no. 6, 2014. DOI: 10.1186/s13054-014-0644-4. [Online]. Available: https://ccforum.biomedcentral.com/articles/10.1186/s13054-014-0644-4.

[54] I. Fredriksson, C. Fors, and J. Johansson, "Laser doppler flowmetry," pp. 1–13, 2007, Department of Biomedical Engineering, Linköping University. [Online]. Available: http://www.imt.liu.se/bit/ldf/ldf.pdf.

[55] H. Shatkay, *The Fourier Transform – A Primer*, Department of Computer Science Brown University, Providence, RI, 1995. [Online]. Available: http://www.phys.hawaii.edu/~jgl/p274/fourier_intro_Shatkay.pdf.

[56] D. T. Lee and A. Yamamoto, "Wavelet analysis: Theory and applications," *Hewlett-Packard Journal*, vol. 45, no. 6, pp. 44–47, 1994.

[57] D. Gabor, "Theory of communication. part 1: The analysis of information," *The Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946. DOI: 10.1049/ji-3-2.1946.0074.

[58] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological Society*, vol. 79, no. 1, p. 63, 1998.

[59] A. Graps, "An introduction to wavelets," *IEEE Computational Science and Engineering*, vol. 2, no. 2, pp. 50–61, 1995, ISSN: 1070-9924. DOI: 10.1109/99.388960.

[60] M. Sifuzzaman, M. Islam, and M. Ali, "Application of wavelet transform and its advantages compared to fourier transform," *Journal of Physical Sciences*, vol. 13, pp. 121–123, 2009, ISSN: 0972-8791.

[61] I. Daubechies, "Where do wavelets come from? a personal point of view," *Proceedings of the IEEE*, vol. 84, no. 4, pp. 510–513, 1996, ISSN: 0018-9219. DOI: 10.1109/5.488696.

[62] P. S. Addison, "Wavelet transforms and the ecg: A review," *Physiological Measurement*, vol. 26, R155–R199, 2005. DOI: doi:10.1088/0967-3334/26/5/R01.

[63] S. J. Russell and P. Norvig, *Artificial Intelligence, A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1995, p. 3.

[64] N. J. Nilsson, *Introduction to Machine Learning, And Early Draft of a Proposed Textbook*. Department of Computer Science, Stanford University, Stanford, CA, 2005, pp. xi, 1–2, unpublished work.

[65] P. Simon, *To Big to Ignore, The Business Case for Big Data*. Wiley, 2013, ISBN: 978-1-118-63817-0.

[66] Z. Ghahramani, *Unsupervised learning*, Gatsby Computational Neuroscience Unit, University College London, London, UK, 2004. [Online]. Available: http://mlg.eng.cam.ac.uk/zoubin/papers/ul.pdf.

[67] C. Walck, *Hand-book on STATISTICAL DISTRIBUTIONS for experimentalists*. Stockholm, Sweden, 1996, pp. 3–4, 36, 99, 119, Internal Report SUF–PFY/96–01, Particle Physics Group, Fysikum University of Stockholm. [Online]. Available: http://www.fysik.su.se/~walck/suf9601.pdf.

[68] E. W. Weisstein, *Sample mean*, MathWorld–A Wolfram Web Resource. Accessed 25.07.16. [Online]. Available: http://mathworld.wolfram.com/SampleMean.html.

[69] ——, *Sample variance*, From MathWorld–A Wolfram Web Resource. Accessed 25.07.16. [Online]. Available: http://mathworld.wolfram.com/SampleVariance.html.

[70] ——, *Central limit theorem*, MathWorld–A Wolfram Web Resource. Accessed 26.07.16. [Online]. Available: http://mathworld.wolfram.com/CentralLimitTheorem.html.

[71] ——, *Correlation coefficient*, From MathWorld–A Wolfram Web Resource. Accessed Jul. 27, 2016. [Online]. Available: http://mathworld.wolfram.com/CorrelationCoefficient.html.

[72] I. Guyon and A. Elisseef, *Feature Extraction, Foundations and Applications*, I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, Eds., ser. Studies in Fuzziness and Soft Computing. Berlin: Springer, 2006, vol. 207, pp. 2–16. DOI: 10.1007/978-3-540-35488-8.

[73] T. Wang and S. Nanda, *A tutorial on feature extraction methods*, Accessed Aug. 16, 2016, 2012. [Online]. Available: https://www.phmsociety.org/sites/phmsociety.org/files/Tutorial_PHM12_Wang.pdf.

[74] H. Liu and H. Motoda, *Feature Extraction, Construction and Selection, A Data Mining Perspective*, H. Liu and H. Motoda, Eds., ser. The Springer International Series in Engineering and Computer Science. New York: Springer, 1998, vol. 453, pp. 3–4. DOI: 10.1007/978-1-4615-5725-8.

[75] H. Motoda and H. Liu, "Feature selection extraction and construction," pp. 1–2, Accessed Jul. 30, 2016. [Online]. Available: http://www.ar.sanken.osaka-u.ac.jp/ motoda/papers/fdws02.pdf.

[76] A. Ghodsi, *Dimensionality reduction, A short tutorial*, Department of Statistics and Actuarial Science, University of Waterloo, Accessed Jul. 22, 2016, Waterloo, Ontario, Canada, 2006. [Online]. Available: http://www.stat.washington.edu/courses/stat539/spring14/Resources/tutorial_nonlin-dim-red.pdf.

[77] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press, 1995, p. 8.

[78] R. E. Bellman, *Adaptive Control Processes, A Guided Tour*. New Jersey: Princeton University Press, 1961.

[79] V. Spruyt, *The curse of dimensionality in classification*, Accessed Jul. 21, 2016, 2014. [Online]. Available: http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/.

[80] E. W. Weisstein, *Hyperplane*, From MathWorld–A Wolfram Web Resource. Accessed Jul. 07, 2016. [Online]. Available: http://mathworld.wolfram.com/Hyperplane.html.

[81] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, Series 6, vol. 2, no. 6, pp. 559–572, 1901.

[82] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933. DOI: /10.1037/h0071325.

[83] I. Jolliffe, *Pricipal Component Analysis*, 2nd ed., P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg, I. Olkin, N. Wermuth, and S. Zeger, Eds., ser. Springer Series in Statistics. New York: Springer, 2002, pp. 1–9, 303. DOI: 10.1007/b98835.

[84] Wikipedia, *Principal component analysis — wikipedia, the free encyclopedia*, Accessed Jul. 22, 2016, 2016. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Principal_component_analysis&oldid=730307136.

[85] J. Shlens, *A Tutorial on Principal Component Analysis*, 2nd ed. La Jolla, CA, 2005, pp. 1, 3–5, 7, 11–12, Institute for Nonlinear Science, University of California, San Diego.

[86] A. A. Farag and S. Elhabian, "A tutorial on principal component analysis," 2009, University of Louisville, CVIP Lab.

[87] M. Richardson, "Principal component analysis," pp. 2, 6–8, 2009, Special tocip essay, Mathematical Institute, Oxford University.

[88] J. Renze, C. Stoverand, and E. W. Weisstein, *Inner product*, From MathWorld–A Wolfram Web Resource. Accessed Jul. 25, 2016. [Online]. Available: `http://mathworld.wolfram.com/InnerProduct.html`.

[89] H. E. Nystad, *Comparison of Principal Component Analysis and Spectral Angle Mapping for Identification of Materials in Terahertz Transmission Measurements*. 2015, p. 18, Master's thesis, Department of Electronics and Telecommunication, Norwegian University of Science and Technology.

[90] *Principal component analysis part ii*, Department of Materials Science and Engineering, Iowa State University. Accessed Jul. 26, 2016. [Online]. Available: `http://cosmic.mse.iastate.edu/library/pdf/pcalevel3.pdf`.

[91] J. Shlens, *A Tutorial on Principal Component Analysis*, 3rd ed. La Jolla, CA, 2014, p. 10, Institute for Nonlinear Science, University of California, San Diego.

[92] K. E. Muller, Y.-Y. Chi, J. Ahn, and J. S. Marron, *Limitations of High Dimension, Low Sample Size Principal Components for Gaussian Data*. [Online]. Available: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.380.372&rep=rep1&type=pdf`.

[93] S. Jung and J. S. Marron, "Pca consistency in high dimension low sample size context," *The Annals of Statistics*, vol. 37, no. 6B, pp. 4104–4130, 2009. DOI: `10.1214/09-AOS709`.

[94] J. W. Osborne and A. B. Costello, "Sample size and subject to item ratio in principal components analysis," *Practical Assessment, Research & Evaluation*, vol. 9, no. 11, 2004, Accessed Jul. 28, 2016. [Online]. Available:`http://pareonline.net/getvn.asp?v=9&n=11`.

[95] I. Guyon and A. Elisseef, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, p. 1176, 2003.

[96] G. K. Kanji, *100 Statistical Tests*, 3rd ed. London, UK: SAGE Publications, 2006, pp. 8, 31–33.

[97] J. H. McDonald, *Handbook of Biological Statistics*, 3rd ed. Balitmore, MD: Sparky House Publishing, 2014, pp. 254–255.

[98] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine Learning Journal*, vol. 53, pp. 23–69, 2003.

[99] A. Smola and S. Vishwanathan, *Introduction to Machine Learning*. Cambridge, UK: Cambridge University Press, 2008, pp. 155–196.

[100] J. Brownlee, *8 tactics to combat imbalanced classes in your machine learning dataset – Machine Learning Mastery*, Accessed Jul. 02, 2016, 2015. [Online]. Available: http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/.

[101] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[102] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, (Jul. 1–8, 2008), 2008, pp. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.

[103] D. Siedhoff, *Adasyn (improves class balance, extension of smote)*, MATLAB function from File Exchange. Accessed Jul. 10, 2016. [Online]. Available: https://se.mathworks.com/matlabcentral/fileexchange/50541-adasyn--improves-class-balance--extension-of-smote-.

[104] Y. Suna, M. S. Kamela, A. K. Wongb, and Y. Wangc, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, p. 3362, 2007.

[105] M. Saraswat, *A complete tutorial on tree based modeling from scratch (in r & python) – Analytics Vidhya*, Accessed Jul. 28, 2016, 2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/.

[106] M. Hauskrecht, *Cs2750 machine learning – lecture 19*, University of Pittsburgh, 2003. [Online]. Available: https://people.cs.pitt.edu/~milos/courses/cs2750-Spring03/lectures/class19.pdf.

[107] C. Ratanamahatana and D. Gunopulos, *Scaling up the naive bayesian classifier: Using decision trees for feature selection*, Computer Science Department, University of California, Riverside, CA. [Online]. Available: http://alumni.cs.ucr.edu/~ratana/DCAP02.pdf.

[108] K. Grąbczewski and N. Jankowski, *Feature selection with decision tree criterion*, Department of Computer Methods, Nicolaus Copernicus University, Toruń, Poland, Accessed Jul. 29, 2016. [Online]. Available: http://www.fizyka.umk.pl/publications/kmk/05-Fsel-DT.pdf.

[109] Wikipedia, *Np-hardness — wikipedia, the free encyclopedia*, Accessed Jul. 30, 2016, 2016. [Online]. Available: https://en.wikipedia.org/w/index.php?title=NP-hardness&oldid=723220359.

[110] I. MathWorks, *Discriminant analysis*, Accessed Aug. 10, 2016, Natick, MA. [Online]. Available: http://se.mathworks.com/help/stats/discriminant-analysis.html.

[111] J. Poulsen and A. French, *Discriminat Function Analysis (DA)*, Accessed Aug. 10, 2016. [Online]. Available: http://userwww.sfsu.edu/efc/classes/biol710/discrim/discrim.pdf.

[112] S. Sayad, *Linear discriminant analysis*, An Introduction to Data Mining – Online resource. Accessed Aug. 10, 2016. [Online]. Available: http://www.saedsayad.com/lda.htm.

[113] I. MathWorks, *Fitcsvm*, Accessed Sep. 2, 2016, Natick, MA. [Online]. Available: http://se.mathworks.com/help/stats/fitcsvm.html.

[114] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[115] S. Nayak, M. K. Soni, and D. Bansal, "Filtering techniques for ecg signal processing," *International Journal of Research in Engineering and Applied Sciences*, vol. 2, no. 2, pp. 671–679, 2012.

[116] V. S. Selvam, *gscatter3*, Department of Electronics & Communication Engineering, Sriram Engineering College, Perumalpattu, India, MATLAB function from File Exchange. Accessed Jul. 15, 2016. [Online]. Available: https://se.mathworks.com/matlabcentral/fileexchange/37970-gscatter3.

[117] K. L. Stenerud, *Analysis of low frequency content from the circulatory system*, Unpublished manuscript, Norwegian University of Science and Technology, Department of Electronics and Telecommunications, 2016.

[118] A. BenSaïda, *Shapiro-wilk and shapiro-francia normality tests*, MATLAB function from File Exchange. Accessed May. 25, 2016. [Online]. Available: https://se.mathworks.com/matlabcentral/fileexchange/13964-shapiro-wilk-and-shapiro-francia-normality-tests.

[119] S. Abboud and O. Barnea, "Errors due to sampling frequency of the electrocardiogram in spectral analysis of heart rate signals with low variability," in *Computers in Cardiology 1995*, (Sep. 10–13, 1995), 1995, pp. 461–463. DOI: 10.1109/CIC.1995.482685.

[120] L. Hejjel and E. Roth, "What is the adequate sampling interval of the ecg signal for heart rate variability analysis in the time domain?" *Physiological Measurement*, vol. 25, no. 6, pp. 1405–1411, 2004.

[121] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft, and J. Popp, "Sample size planning for classification models," *Analytica chimica acta*, vol. 760, pp. 25–33, 2013.

[122] I. B. V. d. Silva and P. J. L. Adeodato, "Pca and gaussian noise in mlp neural network training improve generalization in problems with small and unbalanced data sets," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, 2011, pp. 2664–2669. DOI: 10.1109/IJCNN.2011.6033567.