Tampere University

Mira Haapala

ANALYSIS OF PULSE WAVE PARAMETERS USING SUPERVISED MACHINE LEARNING METHODS

Faculty of Information Technology and Communication Sciences
Master of Science Thesis
March 2019

ABSTRACT

Mira Haapala: Analysis of pulse wave parameter using supervised machine learning methods
Master of Science Thesis
Tampere University
Master's Degree Programme in Electrical Engineering
March 2019

Impedance plethysmographic signals recorded with electrodes attached on the human body provide information on hemodynamics and thus the condition of the arteries. The main objective of the thesis is to study if the quantitative analysis of the bioimpedance signals provides additional information on the risk for cardiovascular diseases compared with clinical parameters currently used in the assessment of the cardiovascular risk of an individual. This thesis aims to answer three main research questions: 1) are pulse wave parameters able to evaluate the condition of the arteries, 2) could pulse wave parameters provide information equal to the clinical data, and 3) could the impedance measurements be utilized for cardiovascular risk stratification.

This thesis analyzes the bioimpedance signals and clinical data collected in the Cardiovascular Risk in Young Finns Study (YFS). The test subjects were 30–45 years old when the data was collected. In this thesis, both frequency and time domain features including pulse wave decompositions are computed from the pulse waves extracted from the bioimpedance signals and their dependence on clinical phenotypes based on YFS data is evaluated. The YFS data contains demographic information (sex, age), anthropometric data (body mass index (BMI)), clinical information (smoking, hypertension, antihypertensive medication), clinical physiologic data (pulse wave velocity (PWV), blood pressure, heart rate, flow-mediated dilation (FMD)), laboratory analyses (fasting insulin and glucose, lipids of the blood) and imaging data (intima-media thickness (IMT), presence of atherosclerotic plaques in the internal carotid artery). The data was measured from 1853 test subjects, but after removal of test subjects with interrupted measurement or with low signal-to-noise ratio, there are 1738 test subjects used in this thesis.

Besides the linear regression analysis, which was implemented to study the association between individual pulse wave signal derived features and clinical reference values, following supervised machine learning methods: linear and quadratic discriminant analysis, support vector machines, naïve Bayes, AdaBoost, Random Forest and *k*-nearest neighbor are applied to answer the objectives of this thesis. A cross-validation and forward selection are applied to find the most relevant pulse wave features that most accurately classify the test subjects. The results are evaluated with receiver operating characteristics (ROC) curve analysis.

This thesis uses three different labeling methods to determine the ground truth for each subject being at low or high risk for cardiovascular diseases: 1) selected cardiovascular risk factors, 2) abnormal body mass index (BMI), blood glucose, triglycerides, high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol or total cholesterol, and 3) existence of atherosclerotic plaque, hypertension or antihypertensive medication.

As a result, the calculated pulse wave parameters provide independent information from the clinical data about the condition of the arteries because the combination of pulse wave parameters and clinical data provided the best classification results in most of the cases. However, the calculated pulse wave parameters alone do not provide as good information as the clinical data, which is shown by the fact that the classifying result with only clinical data was better than the classifying result with only pulse wave parameters.

As a conclusion, risk stratification improves when the clinical data and the pulse wave parameters are combined. However, the analysis methods of signal processing should be optimized for the bioimpedance measurements. Further in order to verify the classification performance of the developed methods, the data should contain wider spectrum of people, from those who have diagnosed cardiovascular diseases to those who do not have such diseases.

Keywords: bioimpedance signal analysis, impedance plethysmography, impedance cardiography, pulse wave, machine learning, pulse wave analysis

TIIVISTELMÄ

Mira Haapala: Pulssiaaltoparametrien analysointi valvottujen koneoppimismenetelmien avulla
Diplomityö
Tampereen yliopisto
Sähkötekniikan diplomi-insinöörin tutkinto-ohjelma
Maaliskuu 2019

Elektrodien avulla ihmiskehosta mitatut impedanssipletysmografiset signaalit antavat tietoa hemodynamiikasta ja siten valtimoiden kunnosta. Tämän opinnäytteen pääasiallinen tavoite on tutkia, antaako impedanssisignaalien kvantitatiivinen analyysi lisäinformaatiota riskistä sairastua sydän- ja verisuonisairauksiin verrattuna kliinisiin parametereihin. Tässä työssä pyritään vastamaan kolmeen tutkimuskysymykseen: 1) pystyvätkö pulssiaaltoparametrit arvioimaan valtimoiden kuntoa, 2) antavatko pulssiaaltoparametrit informaatiota, joka vastaa kliinisen datan informaatiota ja 3) voidaanko impedanssimittauksilla tehdä arvioida riskiä sairastua sydän- ja verisuonitauteihin.

Tämä opinnäytetyö analysoi bioimpedanssisignaaleja ja kliinistä dataa, jotka on kerätty Lasten Sepelvaltimotaudin Riskitekijät -tutkimuksesta (LASERI). Tutkimushenkilöt olivat datan keräyshetkellä 30–45-vuotiaita. Tässä työssä laskettiin sekä aika- ja taajuustason piirteitä että pulssiaaltohajotelmia pulssiaalloista, jotka on määritelty bioimpedanssisignaaleista. Lisäksi näiden piirteiden riippuvuutta arvioitiin tutkimuksessa kerättyyn muuhun tietoon perustuviin kliinisiin ilmentymiin. LASERI-data sisältää informaatiota demografisista (ikä, sukupuoli), antropometristä (painoindeksi, BMI), kliinisistä (tupakointi, verenpainetauti, verenpainelääkitys), kliinisfysiologisista tiedoista (pulssiaallon kulkunopeus, verenpaine, sydämen syketaajuus, virtausvälitteinen vasodilaatio), laboratoriotuloksista (paastosokerin ja -insuliinin pitoisuus, veren rasva-arvot) sekä kuvantamistutkimuksen tuloksista (intima-media-kerroksen paksuus ja plakista kaulavaltimossa). Tutkimusdataa oli alun perin 1853 henkilöltä, mutta analysointikelpoista dataa oli saatavilla 1738 henkilöltä.

Lineaarisen regressioanalyysin, missä tutkittiin yksittäisen pulssiaaltoparametrien ja kliinisten viitearvojen välistä yhteyttä, lisäksi käytettiin seuraavia valvottuja koneoppimismenetelmiä tutkimuskysymyksiin vastaamiseen: lineaari- ja neliödiskriminanttianalyysi, tukivektorikone, naiivi Bayes-luokittelija, AdaBoost-luokittelija, satunnaismetsää ja *k*-lähintä naapuria. Ristiinvalidoinnilla ja eteenpäin askeltavalla algoritmilla etsittiin oleellisimmat pulssiaaltopiirteet, jotka luokittelevat koehenkilöt parhaiten. Tuloksia arvioitiin ROC-käyrän (receiver operating characteristics) avulla.

Tässä työssä käytettiin kolmea erilaista nimiöintitapaa koehenkilöiden luokittelemiseksi matalampaan ja korkeampaan riskiin sairastua sydän- ja verisuonitauteihin: 1) ennalta määrättyjen sydän- ja verisuonisairauksien riskitekijät, 2) epänormaali painoindeksi (BMI), verensokeri, HDL-kolesteroli-, LDL-kolesteroli-, triglyseridi- tai kokonaiskolesteroliarvo ja 3) ateroskleroottisen plakin ilmentymä, verenpainetauti tai verenpainelääkitys.

Tuloksina saatiin, että pulssiaaltoparametrien informaatio on osittain riippumatonta kliinisen datan sisältämästä informaatiosta, kun tarkastellaan valtimoiden kuntoa, koska näiden yhdistelmä antoi parhaimmat luokittelutulokset suurimmassa osassa tapauksia. Kuitenkin pulssiaaltoparametrit eivät anna yhtä hyvää informaatiota kuin kliininen data, sillä luokittelutulokset olivat parempia kliinisellä datalla yksinään kuin pulssiaaltoparametreilla.

Loppupäätelminä voi sanoa, että riskiositus paranee, kun pulssiaaltoparametrit yhdistetään kliinisen datan kanssa. Kuitenkin signaalin käsittelyn analysointimenetelmiä pitäisi optimoida bioimpedanssimittauksille. Lisäksi datan pitäisi sisältää sellaisia enemmän sellaisia henkilöitä, joilla on diagnosoitu sydän- ja verisuonisairauksia kuin myös heitä, joilla niitä ei ole, jotta luokittelutuloksen suoritusta voitaisiin varmistua.

Avainsanat: bioimpedanssisignaalianalyysi, impedanssiplethysmografia, impedanssikardiografia, pulssiaalto, koneoppiminen, pulssiaalto analyysi

PREFACE

I would like to thank my examiner M. Sc. Mikko Peltokangas for giving me many, many very good comments and sharing his codes for the calculation of the pulse wave parameters. I thank my examiner Assistant Professor Antti Vehkaoja for sharing his thoughts and giving comments for this thesis. I would like to send my compliments to my examiner Professor Niku Oksala for giving comments and answering on my questions concerning physiology of the cardiovascular system. I express my gratitude to Professor Mika Kähönen, Professor Terho Lehtimäki and MD Leo-Pekka Lyytikäinen for providing the dataset used in this thesis. I would like to thank D. Sc. Pekka Kumpulainen for helping with the machine learning part by providing the example codes. I send my warmest thanks to my friend Heini for proof-reading this thesis and improving my English. Lastly, I thank my partner Topi for sharing his thoughts and supporting me during this whole process.

Tampere, 12.3.2019

Mira Haapala

CONTENTS

APPENDIX C: CLASSIFYING PARAMETERS SELECTED IN THE EVALUATED SCENARIOS

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AGI	Ageing index
AIx	Augmentation index
AP	Augmented pressure
ASE	American Society of Echocardiography
AUC	Area under receiver operating characteristic curve
AV	Atrioventricular
BMI	Body mass index
BP	Blood pressure
CIMT	Carotid intima-media thickness
DBP	Diastolic blood pressure
DT	Diastolic time
ECG	Electrocardiography, electrocardiograph, electrocardiogram
ESC	European Society of Cardiology
FFT	Fast Fourier Transform
FMD	Flow-mediated dilation
FPR	False positive rate
HDL	High-density lipoprotein
HP	Heart period
ICG	Impedance cardiography, impedance cardiograph, impedance cardiogram
IMT	Intima-media thickness
k-NN	K-nearest neighbor
LDA	Linear discriminant analysis
LOOCV	Leave-one-out cross-validation
LVET	Left ventricular ejection time
IPG	Impedance plethysmography, impedance plethysmograph, impedance plethysmogram
LDL	Low-density lipoprotein
MAP	Mean arterial pressure
mmHg	Millimeter of mercury
PP	Pulse pressure
PPG	Photoplethysmography, photoplethysmograph, photoplethysmogram
PWA	Pulse wave analysis
PWV	Pulse wave velocity
QDA	quadratic discriminant analysis
ROC	Receiver operating characteristic curve
SA	Sinoatrial
SBP	Systolic blood pressure
SNR	Signal-to-noise ratio
SVM	Support vector machine
TPR	True positive rate
YFS	Young Finns Study

A	area
B	diastolic wave
C	cost parameter
$\mathbf{D}$	regressor matrix

$G_m(\boldsymbol{x})$	weak classifier
I	current
L	length
M	margin
P_i	inflection point
P_1	First systolic peak
P_2	Second systolic peak
T_i	time delay
U	voltage
V	volume
Z	impedance
d	regressor variable
r_i	Lagrange multiplier
w_0	threshold
w	weight vector
x	multidimensional data
y	response variable
$\boldsymbol{y}$	vector of observations
$\boldsymbol{y'}$	vector of fitted values
α_i	Lagrange multiplier
α_m	coefficient
β	regression coefficient
$\boldsymbol{\beta}$	regression coefficient vector
$\boldsymbol{\beta'}$	least-square estimator vector
ε	random error
$\boldsymbol{\varepsilon}$	random error vector
μ	mean vector
ξ	slack variable
ρ	resistivity
$\boldsymbol{\Sigma}$	covariance matrix
σ	conductivity
ω_m	weight

1. INTRODUCTION

Cardiovascular diseases cause half of the deaths of the working-age people in Finland and they are one of the most common causes of death [1]. Atherosclerosis, which is the most common cause of arterial narrowing, may lead to coronary artery disease that is one of the most common cardiovascular diseases in Finland. Therefore, it is important to understand what might increase the risk for cardiovascular diseases. In Finland, Cardiovascular Risk in Young Finns Study (YFS) studies cardiovascular risk from the childhood to the death of the test subjects [2]. This thesis analyzes data collected from the subjects participating in YFS in their thirties and forties.

The analyzed YFS data is divided in two categories: 1) clinical data and 2) bioimpedance signals. The clinical data contains patient history and risk factors, laboratory analyses, (blood lipids, fasting glucose and insulin), anthropometric data (body mass index (BMI)), clinical physiologic data (blood pressures, heart rate, flow-mediated dilation (FMD), pulse wave velocity (PWV)), and imaging data (intima-media thickness (IMT) and presence of atherosclerotic plaques in the internal carotid artery). The clinical data is described in more detailed way in Section 4.1. In addition to this data, impedance plethysmogram (IPG), impedance cardiogram (ICG) and electrocardiogram (ECG) have been obtained from the subjects using CircMon cardiovascular monitoring device. From the IPG and ICG, different pulse wave parameters were calculated, and they are presented in Section 4.2. The main research questions examined in this thesis are: 1) are calculated pulse wave parameters able to evaluate the condition of the arteries, 2) could they provide additional information equal to clinical data, and 3) could the impedance measurements be utilized for cardiovascular risk stratification.

Previous studies have proposed various non-invasive arterial pulse wave measurement and analysis methods for assessing the condition of the vasculature or risk for cardiovascular diseases. The non-invasive measurement methods include photoplethysmographic (PPG) recordings [3]–[5], different kinds of pressure and force transducers such as tonometric sensors and thin-film sensors, as well as bioimpedance measurement. Photoplethysmograph records the volume changes of the blood with light. The volume changes modify the scattering and absorption of the light [6, p. 372]. Pressure and force transducers have to be placed on the top of a superficial artery so that force is applied to the artery wall and thus the sensor is able to measure the changes in arterial blood pressure and the force [6, p. 332]. Bioimpedance measurements are based on the impedance changes of the body such as impedance changes caused by pulsatile dilating arteries [6, p. 366].

In the analysis point of view, the estimates of the arterial pulse wave velocity (PWV) are commonly considered as an indicator of arterial health and the relation of the PWV to the condition of arteries and risk for cardiovascular diseases has been studied [7]–[9]. Different kinds of augmentation indices have been proposed especially for the analysis of pressure pulse waves. Aging index (AGI) [4] based on the 2nd derivative of pulse wave is a traditional analysis method for index finger PPG signals. Besides the direct pulse wave derived features, different kinds of pulse wave decompositions have been proposed for modeling the wave reflection and superposition [10]. However, it is noteworthy that the values of these parameters change with the age, because ageing changes the shape of the arterial pulses (see Figure 1). To the best of my knowledge, an analysis of bioimpedance signals has not been done previously by calculating pulse wave parameters.

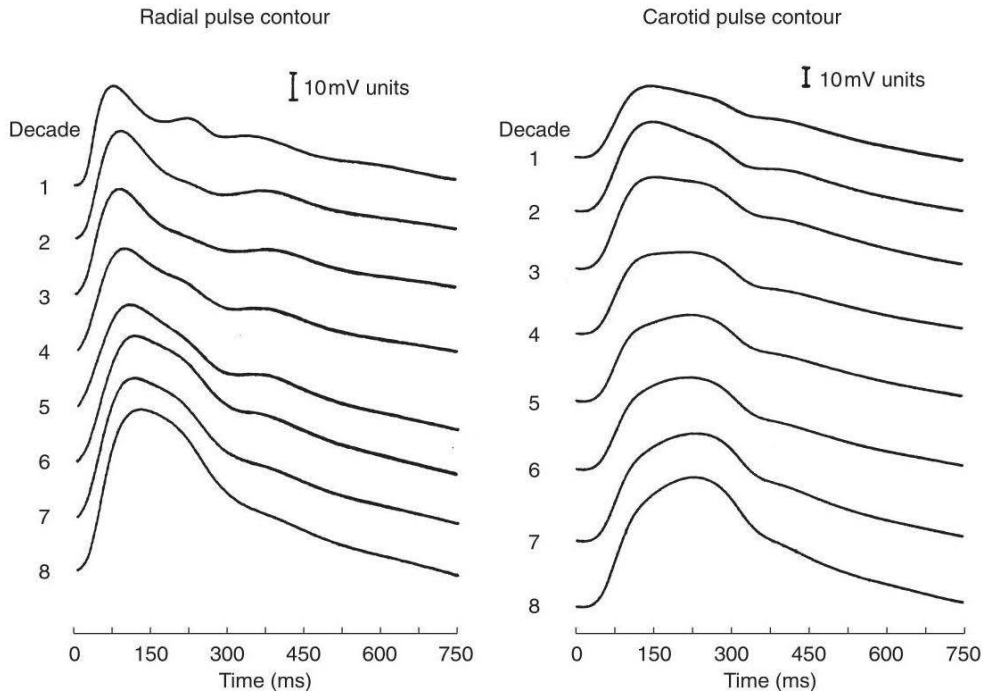

Figure 1: Effect of ageing to the shape of arterial pulse waves [11].

In addition to the analysis of individual pulse wave features, seven different supervised machine learning methods are tested to study if the combination of different data sources can provide information to the cardiovascular risk stratification compared with conventional clinical parameters or if the pulse wave parameters could be used to evaluate the condition of the arteries. These supervised machine learning methods are linear and quadratic discriminant analysis, support vector machine, naïve Bayes, AdaBoost, Random Forest and *k*-nearest neighbor. Complex deep learning methods are left outside the scope of this thesis.

Section 2 provides the physiological background of this thesis and Section 3 provides the discussion about the data science background. In Section 4, the data set and how it is analyzed is presented. The results and discussion are presented in Section 5. Lastly, the conclusions are presented in Section 6.

2. PHYSIOLOGICAL BACKGROUND

2.1 Cardiovascular system

The cardiovascular system consists of heart and blood vessels, blood, arteries and veins. Its function is to maintain the homeostasis of the body. The main functionalities of the circulating blood are to transport oxygen and nutrients into the cells, hormones from cell to cell and metabolites such as carbon dioxide from the cells. [12, p. 361]

2.1.1 Heart

The heart is a muscle that works as pump whose main function is to pump blood to the different target organs of the human body by contracting and relaxing. The size of the heart is approximately the size of the owner's fist and its weight is 250–390 g in men and 200–275 g in women. [12, p. 362], [13, p. 224]

The heart consists of four chambers: two atria and two ventricles that can be divided into right and left side. The right atrium receives deoxygenated blood from the body and then the blood is transferred into the right ventricle. The right ventricle pumps the deoxygenated blood to the pulmonary circulation and lungs for oxygenation and removal of carbon dioxide. After the oxygenation, the blood moves from the lungs to the left atrium. The left ventricle receives the blood from the left atrium and pumps it into the peripheral circulation that delivers blood to peripheral target organs. [13, pp. 227–229]

The heart has an electrical conduction system that ensures that the heart contracts in right rhythm. Firstly, the atria contract simultaneously and after that the ventricles contract simultaneously [13, p. 234]. The sinoatrial (SA) node (see Figure 2) is the place where the electrical activation cycle of the heart starts. The activation travels quickly through anterior internodal tract, middle internodal tract and posterior internodal tract around the right atrium and Bachmann's bundle to the left atrium and finally to the atrioventricular (AV) node (see Figure 2), causing the atria to contract. The electrical activation delays normally in AV node for 120 ms–210 ms, which depends on the heart rate and the tone of autonomic nervous system. During this delay, the electrical activation travels to the left atrium and in the meantime, ventricles are filled with blood. From the AV node, the electrical activation travels through bundle of His to Purkinje fibers in a few milliseconds. This causes the ventricles to contract. [14, pp. 21–24]

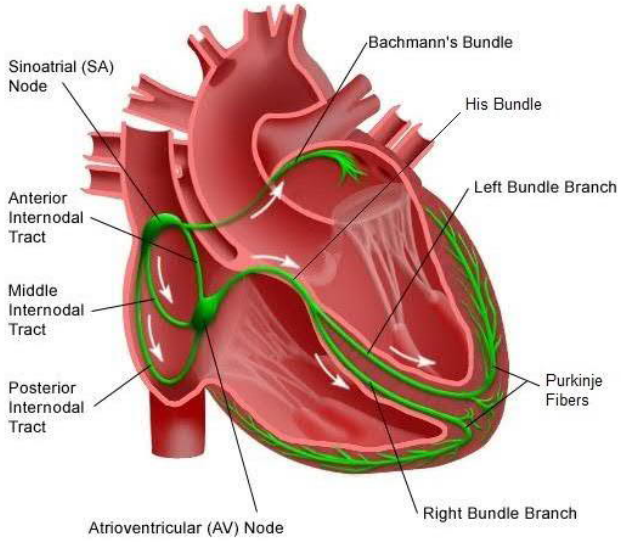

Figure 2: The electrical conduction system of the heart. Adapted from [15].

2.1.2 Depolarization, repolarization and action potential

When a cardiac cell is in resting phase, its intra-cellular potential is negative with respect to the extra-cellular potential. The electrochemical potential over the cell membrane is caused by different ion concentrations inside and outside the cell membrane. The charge balance of the cell rapidly changes because of depolarization. As a result, the intra-cellular potential becomes positive with respect to the extra-cellular potential [16, p. 4]. Repolarization refers that the cell returns into its resting state. Repolarization and depolarization propagate as wave fronts, so that every cell repolarizes and depolarizes in appropriate time [16, p. 4]. In the cell, there are slow sodium-calcium channels that let the sodium and calcium to flow into the cell. In addition, there are potassium channels which let potassium ions to flow into the cell and out from the cell.

The heart has two types of action potentials: fast and slow, which cause the electrical activation of the heart. The slow action potential is generated in the SA node and the fast action potential happens in the cells of internodal tracts, in the cells of atriums, Purkinje fibers, and in the cells of ventricular muscle. The action potential of the cardiac cell can be divided into 5 phases (0–4), see Figure 3. The phase 0 is the depolarization. During the depolarization, the sodium ion (Na^+) permeability of cardiac cell increases. When the membrane potential reaches –75 mV, all sodium channels open and the membrane potential raises to +30 mV. In the meantime, the permeability to potassium (K^+) decreases. After the membrane, potential has reached +30 mV, all sodium channels close. [14, pp. 21–24], [17, pp. 369–370], [18, pp. 365–366]

The phase 1 is the early repolarization. During this phase, the K^+ ions flow out from the cell. Through slow channels Na^+ and calcium ions (Ca^{2+}) flow into the cell and K^+ ions continue to flow out from the cell and causing the phase 2 – the plateau. The plateau ends

when slow channels are inactivated. The repolarization (phase 3) starts after the slow channel are closed and the outflow of K^+ ions increases, which causes the membrane potential to decrease. Finally, the membrane potential reaches the phase 4, which is the dynamic resting potential during the diastole. During this resting state, the Na^+ ions flow out and K^+ ions flow in. [14, pp. 21–24], [17, pp. 369–370], [18, pp. 365–366]

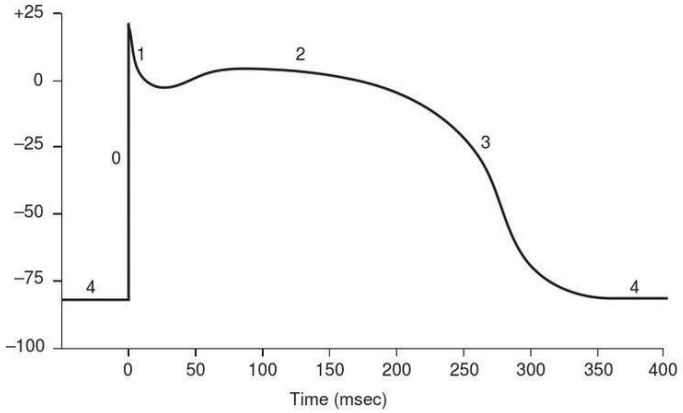

Figure 3: Phases of the action potential of the cardiac muscle: 0) depolarization, 1) early repolarization, 2) plateau, 3) repolarization and 4) resting state [17, p. 370].

2.1.3 Blood vessels

The blood vessels form a vascular system, which consists of five different type of blood vessels called arteries, arterioles, capillaries, venules and veins (see Figure 4) [12, p. 374].

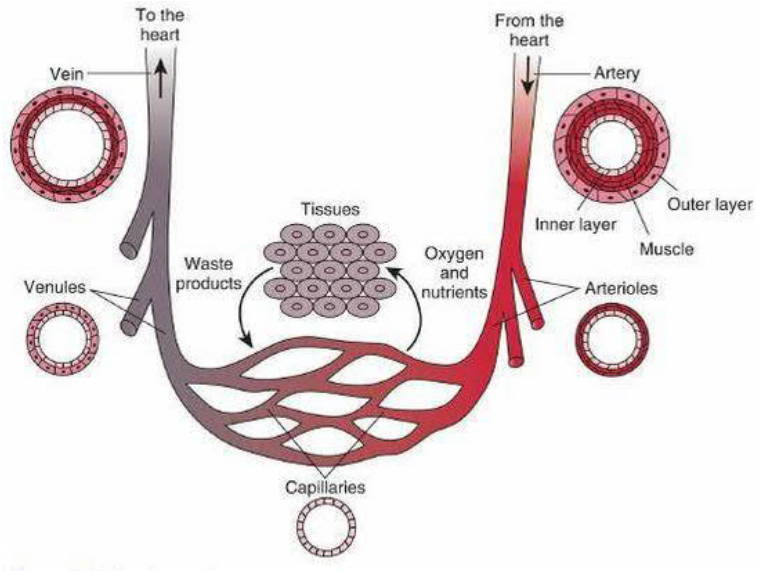

Figure 4: A schematic illustration of blood vessels [13, p. 204].

The heart pumps blood to the arteries, and the arteries transport blood away from the heart all the way to the arterioles. The arterioles are blood vessels that extend from the arteries. In the end, the arterioles become capillaries, where the exchange of oxygen, nutrients and

waste products takes place. The venules are small blood vessels where the blood flows from the capillaries. The venules are drained by the veins, which transport blood back to the heart. Almost all arteries carry oxygenated blood and all veins carry deoxygenated blood. However, the pulmonary and umbilical arteries carry deoxygenated blood whereas pulmonary and umbilical veins carry oxygenated blood. Moreover, the pressure is clearly lower in the veins than in the arteries. [12, p. 374], [13, pp. 203–204]

The walls of the arteries and veins are composed of three different layers: 1) tunica intima 2) tunica media and 3) tunica adventitia [12, pp. 374–376]. The walls of capillaries are made of endothelium that is a single layer of cells and a basement membrane [13, p. 207]. Figure 5 presents the layers of the artery, vein and capillary.

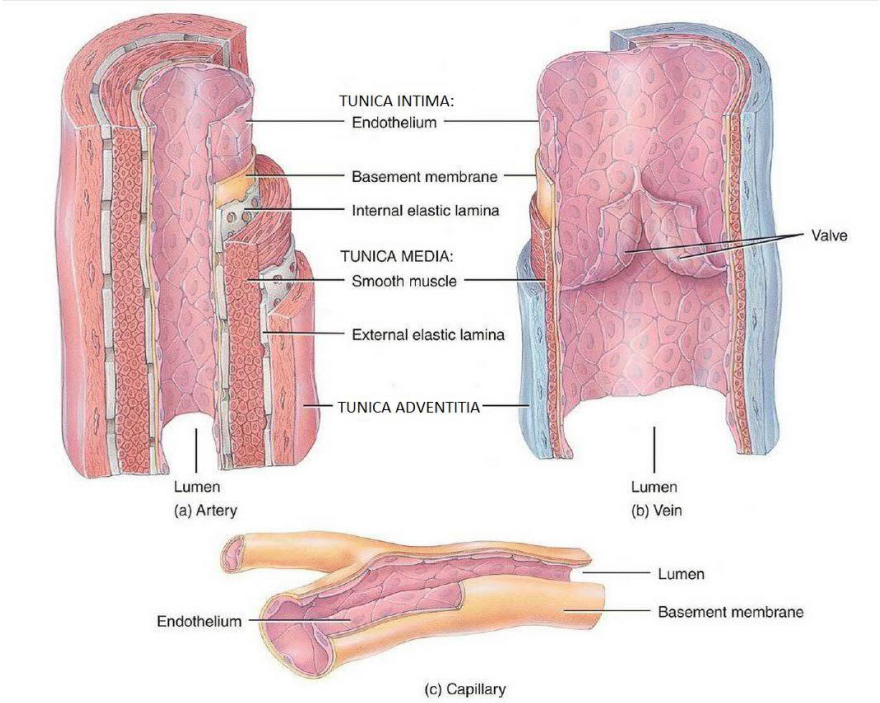

Figure 5: The structure of an artery, a vein and a capillary. Adapted from [13, p. 205].

The tunica intima has a thickness of only one endothelial cell layer. It forms a slick surface, so that the blood can flow faster. Under tunica intima is a layer called basement membrane, which can be found from the arteries and veins. However, the artery consists also of an internal elastic lamina, which is under the basement membrane. Tunica media is made of elastic fibers and smooth muscle cells that are able to constrict or dilate. This functionality plays an important role in controlling the blood pressure and blood flow. The sympathetic branch of the autonomic nervous system controls the smooth muscles. When the sympathetic branch is stimulated, the smooth muscle contracts. Moreover, the smooth muscle is much thicker in artery than in vein. External elastic lamina layer is under the tunica media, but it is not found from the vein. The outermost layer, the tunica adventitia, consists of collagen fibers. The purpose of the tunica adventitia is to support and protect the vessel. Moreover, it anchors the blood vessel to the nearby organs. The

tunica adventitia is similar for both the arteries and the veins. [12, pp. 374–376], [13, pp. 204–206]

2.1.4 Blood pressure

Blood pressure refers to the pressure that pushes the innermost wall of the blood vessels and keeps the blood flowing also between the heartbeats. The blood pressure is highest in the arteries and lowest in the right atrium during the diastole. The contraction and relaxation of the heart causes the blood pressure to rise and fall. Extremities of periodically varying blood pressure are called as systolic blood pressure (SBP) and diastolic blood pressure (DBP), which are measured conventionally from brachial artery [12, pp. 387–388], [18, pp. 381–383]. The diastolic pressure in the left ventricle is called an inflation pressure [19]. The SBP is defined as arterial pressure when the ventricles contract and the DBP is defined as arterial pressure when the ventricles relax. The average blood pressure of pulsatile varying blood pressure is called mean arterial pressure (MAP). The MAP can be approximated with the following formula

$$MAP = DBP + \frac{1}{3}(SBP - DBP). \tag{2.1}$$

[12, pp. 387–388], [18, pp. 381–383]

In Finland, the blood pressure in adults is considered normal if the SBP is under 130 mmHg and DBP is under 85 mmHg. If the SBP is between 130 mmHg–139 mmHg and DBP is between 85 mmHg–89 mmHg, the blood pressure is considered satisfactory. If the systolic blood pressure is equal or higher than 140 mmHg or the diastolic blood pressure is equal or higher than 90 mmHg, the person suffers from hypertension (high blood pressure). [20] Hypertension causes problems with the arteries and the heart, because the heart must pump against increased arterial pressure, which causes changes in the organization and properties of myocardial cells. This illness is known as left ventricular hypertrophy. In the end, the contractile function diminishes and heart failure occurs [18, p. 417], [20]. There are two types of hypertension – primary hypertension and secondary hypertension. In the primary hypertension, the blood pressure rises itself. The secondary hypertension is caused by some other mechanism, like a renal artery stenosis resulting in release of vasoactive substances followed by increased blood pressure [21].

2.1.5 Intima-media thickness

Intima-media thickness (IMT) is defined as the distance from lumen-intima (yellow line in Figure 6) interface to media-adventitia (pink line in Figure 6) interface. It is considered to be a marker of subclinical atherosclerosis, because increase of IMT is a part of developing atherosclerosis taking decades as a whole, and one of its earliest signs is the thickening of arterial wall. [22]–[25] According to the European Society of Cardiology (ESC),

IMT is abnormal if it is greater than 0.9 mm [22]. However, the American Society of Echocardiography (ASE) states that IMT is abnormal if it is over 75th percentile of the population data [22]. The growth of IMT correlates positively with the increasing number of cardiovascular risk factors [23] such as physical inactivity, smoking, hypertension and obesity [26].

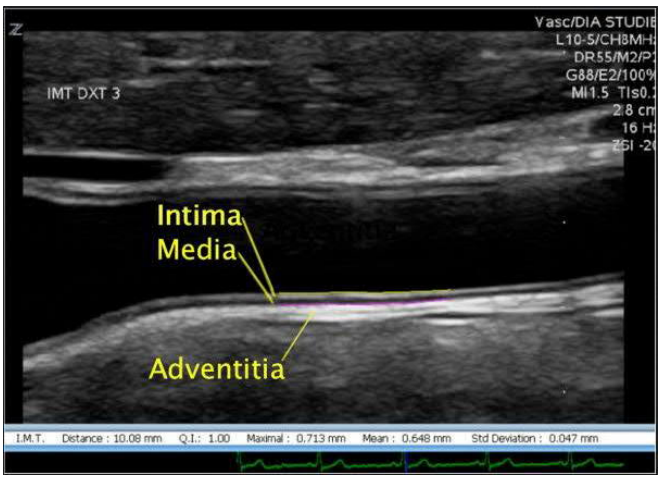

Figure 6: Ultrasonography image of IMT, where yellow line shows lumen-intima and pink line is media-adventitia interface [27].

Carotid intima-media thickness (CIMT) is IMT measured from the carotid artery. Usually, CIMT is measured from the common carotid artery rather than from internal carotid artery, because the former is found to provide more accurate information than the latter one [28]. As a superficial, rather stationary and parallel to the neck, the carotid artery is optimal for the ultrasonographic measurement [29]. The current conception is that CIMT correlates with a number of cardiovascular risk factors [25], [30]. However, the prediction of atherosclerosis by CIMT has been questioned since common carotid artery thickens with age or with hypertension [28]. The carotid artery wall starts to thicken when the person is approximately 18 years old. Therefore, the age should be considered, when CIMT is analyzed. With younger persons, the carotid wall thickness is not dependent on age nor the sex of the person. [29]

IMT or CIMT is measured with ultrasonography, because it is noninvasive and safe for the patient. In the ultrasound image, the IMT shows as a double-line pattern on both walls of common carotid artery, as shown in Figure 6. When the structure of the arterial wall is evaluated, one of the best methods is ultrasound [22]. According to ASE, CIMT should be measured on left and right carotid artery using three different measurement angles [25]. The imaging of CIMT can be done with segments that take an image during R-wave of the electrocardiogram, because the R-wave represents an end-diastolic moment of a previous heartbeat, which means that CIMT is thickest during that time. CIMT is thinnest during the peak systole [29]. Moreover, CIMT may vary from 5% to 10% during one heartbeat [31]. However, there is not any standardized method for IMT imaging, and

therefore the results may vary depending on how and where the measurements have been done [22], [25].

The intima-media thickens when low-density lipoprotein (LDL) cholesterol tries to pass through the intima. The LDL cholesterol produces cholesterol into intima and high-density lipoprotein (HDL) cholesterol collects the cholesterol and clear it away from the intima. When the endothelial malfunctions, the LDL starts to flock in the vessel, because the HDL cannot collect all the LDL away. The malfunction also causes monocytes to come into the intima to phagocytize the lipids. These monocytes prefer oxidized LDL cholesterol. After they have eaten some of the oxidized LDL, they turn into foam cell. Foam cells can be found from an artery that suffers from atherosclerosis. [32, pp. 305–309]

2.1.6 Flow-mediated dilation

Flow-mediated dilation (FMD) is a phenomenon, which refers to a dilation of arteries when the blood flow increases. FMD describes the vascular endothelial function and it is measured from a brachial artery. The first step of the FMD measurement is to measure the diameter of the artery, the baseline diameter, with ultrasound. Then a pressure cuff is placed around the proximal forearm. The cuff is filled to block the blood flow, thus causing ischemia in distal part of the upper limb. When the cuff pressure is released, the brachial artery dilates so that the blood flow can increase in the distal part of the upper limb. Then the diameter of the artery, the peak diameter, is measured again with ultrasound and the difference between first measurement and second measurement is calculated and presented in percentage as

$$\text{FMD}\ (\%) = \frac{\text{peak diameter} - \text{baseline diameter}}{\text{baseline diameter}}. \tag{2.2}$$

[33], [34]

When FMD is measured from the subject, multiple factors affect the result. For instance, the antioxidants, like vitamin C, should not be digested before FMD study because they increase FMD. However, a diet that contains many antioxidants is harder to control. Digested food can affect FMD considerably and therefore products with high-fat or high-carbohydrate levels and caffeine should be avoided before FMD measurement. Smoking is a known risk factor of cardiovascular diseases and it attenuates FMD, therefore the subject should refrain from smoking and avoid smoke exposure before FMD measurement. In contrast with smoking and caffeine, exercise improves FMD. Thus, the test subject should refrain from exercising. Moreover, the menstrual phase affects FMD; hence, the measurement should be taken during menstruation to avoid level of female sex hormones affecting the obtained value. [34]

FMD predicts cardiovascular events, such as a myocardial infraction or angina pectoris, in patients who have a diagnosed cardiovascular disease [35], [36]. Moreover, the values of FMD are lower in those people who have had a cardiovascular event after a surgery [37], [38]. However, FMD is independent of previous cardiovascular events and can therefore be used as a predictor of cardiovascular diseases. In addition, long-term cardiovascular events can be predicted also from healthy subjects with FMD and it succeeds to predict cardiovascular events even though the disease is somewhere else in the body, for instance in the lower limbs [39], [40]. Thus, FMD has long-term prognostic value. Even if FMD is shown to have connection with cardiovascular events, it is not used in clinical practice, but its use is limited in research. Therefore, FMD should be studied more and standardized, so that it could be clinically used [40].

2.2 Bioimpedance measurements

The bioimpedance measurements are based on the electrical conductivity of the biological fluids and the biological tissues. The electrical conductivity is a reciprocal of resistivity and it describes how well the material conducts the electrical current. The conductivity of the biological tissue is frequency dependent and this property makes it easier to measure the impedance changes of the body [41]. The measurement of the bioimpedance can be implemented with electrodes that feed current to the measured volume. Then the voltage drop is measured in order to the calculate impedance change, and thus the volume change. The conductivities vary between the tissues as can be seen from the examples shown in Table 1. The only tissue, which conductivity increases with the frequency, is the muscle. The blood has the highest conductivity, because it contains charge-carrying ions. The fat has lowest conductivity, because the fat is a nonpolar compound [42, p. 103].

Table 1: Conductivities of the blood, fat, and muscle. Adapted from [42, p. 103].

Tissue	**Conductivity σ (S/m) at 1 Hz–10 kHz**	**Conductivity σ (S/m) at 1 MHz**
Whole blood	0.7	0.7
Fat	0.02–0.05	0.02–0.05
Muscle	0.05–0.4	0.6

Bioimpedance can be measured with impedance plethysmograph (IPG) or impedance cardiograph (ICG), from which the impedance plethysmograph is older [43]. Both impedance plethysmography and impedance cardiography are based on the volume changes of the arteries. In addition, the measurement of bioimpedance is used in the impedance pneu-

mography, which measures the changes in respiratory impedance signals [44]. The impedance plethysmography is presented in 2.2.1 and impedance cardiography is presented in 2.2.2.

2.2.1 Impedance plethysmography

The IPG was introduced in 1940 by Jan Nyboer [41]. The IPG measures volume changes of body tissues by measuring the electrical impedance of the body tissue on the body surface and one of its the most relevant clinical applications is the measurement of blood flow [45]. The impedance of the tissue is modulated by the pulsatile flow of high-conductive blood (see Table 1) when the heart pumps blood into arteries, the volume and cross-sectional area of the arteries increases, and the impedance decreases. The blood plasma contains approximately 90 % water containing ions, which enables the current to flow easily with blood [46, p. 12]. An example of the IPG signal waveform recorded from the shin is shown in Figure 7.

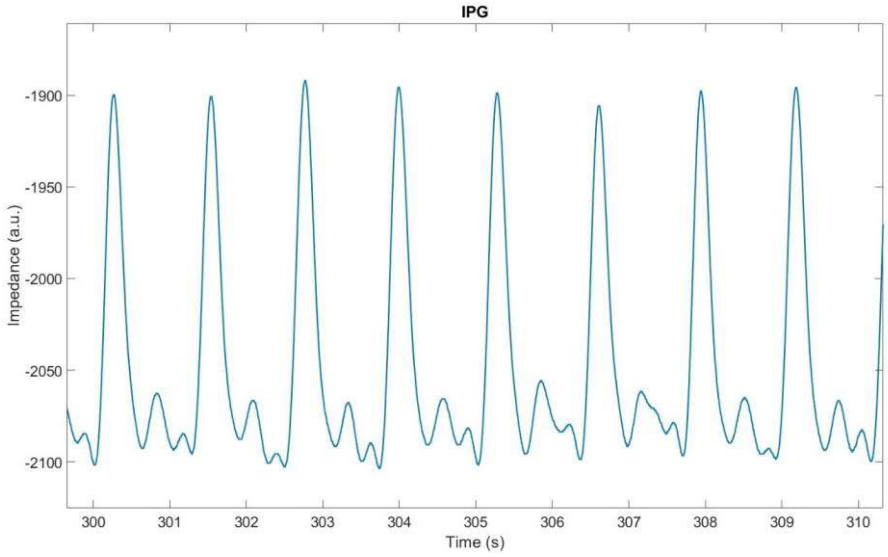

Figure 7: The IPG signal measured from lower limb.

An advantage of the IPG is that it is non-invasive and easy to use. Its disadvantage is that it might not be accurate enough for attempted application and if the impedance changes, its cause is not always clear [6, p. 372]. When the impedance is measured from the body, the frequency of the current fed to the human body is typically 30 kHz–75 kHz, because physiological effects on nerves and muscles decrease when frequency increases [47, p. 242]. However, according to Webster [6, p. 367], the best frequency is approximately 100 kHz, because with higher frequencies the impedance of skin-electrode interface and the changes in impedance caused by motion artefacts, decrease. However, the frequency should not exceed 100 kHz, because the instrument design becomes more challenging and the skin-electrode impedance is adequate. Moreover, the amplitude of the current fed

to the body should be approximately 1 mA, so that the signal-to-noise ratio (SNR) is adequate and the current is safe for the human. [6, pp. 366–367]

In the simplest model, explaining the changes in bioimpedance caused by the pulsatile blood perfusion, the human body is modeled as a cylindrical object, as illustrated in Figure 8. However, the model has following assumptions:

1. All the arteries expand equally.
2. The resistivity of the blood ρ_b is constant.
3. The arteries are parallel to the lines of current.

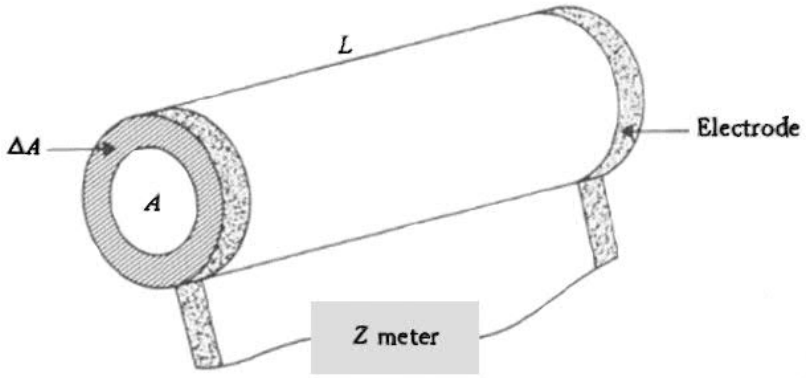

Figure 8: Model of cylindrical limb. Adapted from [6, p. 367].

The first assumption is valid only for people who have healthy arteries. The second assumption is not true, because resistivity of the blood decreases when velocity of the blood decreases. The third assumption is not valid for a knee. [6, pp. 367–368]

If ΔA (Figure 8) is the change of the cross-sectional area of a cylindrical limb then the shunting impedance appearing during systole in parallel with diastolic (baseline) impedance Z is defined as

$$Z_b = \frac{\rho_b L}{\Delta A}, \tag{2.4}$$

where Z_b is the impedance of blood, ρ_b is the frequency-dependent resistivity of the blood, L is the length of the limb and ΔA is the change of the cross-sectional area of the limb. As the change in cross-sectional area is pulsatile due pulsatile dilation of arteries, the total impedance of a cylindrical limb is pulsatile as function of time. The change of the volume ΔV of the blood is calculated using Z_b as

$$\Delta V = L\Delta A = \frac{\rho_b L^2}{Z_b}. \tag{2.5}$$

However, the change of total impedance $\Delta Z = [(Z_b \parallel Z) - Z]$ (see Figure 9) is normally measured instead of Z_b. Mathematically ΔZ is defined as

$$\Delta Z = \frac{ZZ_b}{Z + Z_b} - Z = \frac{-Z^2}{Z + Z_b}, \tag{2.6}$$

where Z is the baseline impedance. Because $Z \ll Z_b$, the Z_b can be approximated as

$$\frac{1}{Z_b} \cong \frac{-\Delta Z}{Z^2}. \tag{2.7}$$

When Eq. (2.7) is substituted into Eq. (2.5), the result is

$$\Delta V = \frac{-\rho_b L^2 \Delta Z}{Z^2} \tag{2.8}$$

The minus sign in the equation indicates that the impedance increases when the volume decreases.

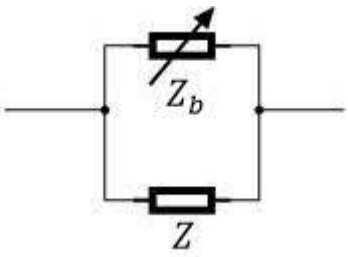

Figure 9: Electrical equivalent for baseline impedance Z and shunting impedance Z_b.

2.2.2 Impedance cardiography

ICG is closely related to IPG, because ICG is IPG measured from the thoracic region [48], [49, p. 353]. Unfortunately, sometimes IPG and ICG are used as synonyms to each other [50]. Similarly, to IPG, the ICG measures the changing impedance [51]. The ICG is dependent on the resistivity of the blood and the used frequency, which is typically between 20 kHz–100 kHz. However, the used frequency should not exceed 100 kHz as it was stated in the case of IPG, because the skin-electrode impedance is small, and the instrumentation is simple when the frequency is 100 kHz. ICG is calculated with Eq. (2.8) as is IPG. An example on ICG signal is shown in Figure 10. [52]

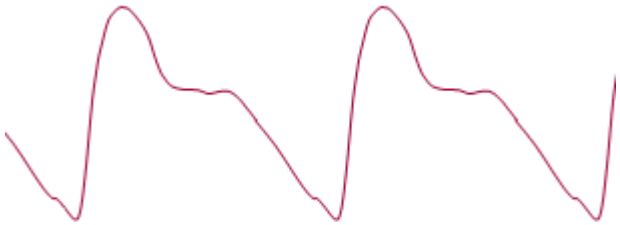

Figure 10: ICG signal measured from thoracic region. Adapted from [53, p. 409].

Usually, the ICG is used to assess stroke volume and cardiac output by measuring the impedance changes with electrodes. An advantage of the ICG-based evaluation is that the result of the ICG is not dependent on the person who is interpreting the signal as is in case of ultrasound.

2.2.3 Four-wire measurement

When the object, whose resistance is measured, is located at a significant distance from the measurement equipment, the impedances of the wires and connectors, such as electrodes, will affect the measurement [54, pp. 282–284]. Therefore, a four-wire measurement is employed. In the bioimpedance measurements, the main benefit of four-wire measurement is that it decreases the effect of skin-electrode impedance. When the effects of skin-electrode resistance decrease, the shape of the wave is not corrupted. Figure 11 presents the principle of the four-wire measurement.

The impedance is calculated with Ohm's law as

$$U = I \cdot Z \tag{2.3}$$

where the U is voltage, I is current and Z is impedance. In the four-wire measurement, the current is larger in the wires of the ammeter than in the wires of voltmeter. Therefore, the impedance of the wires or electrodes does not have a significant effect on the measured impedance.

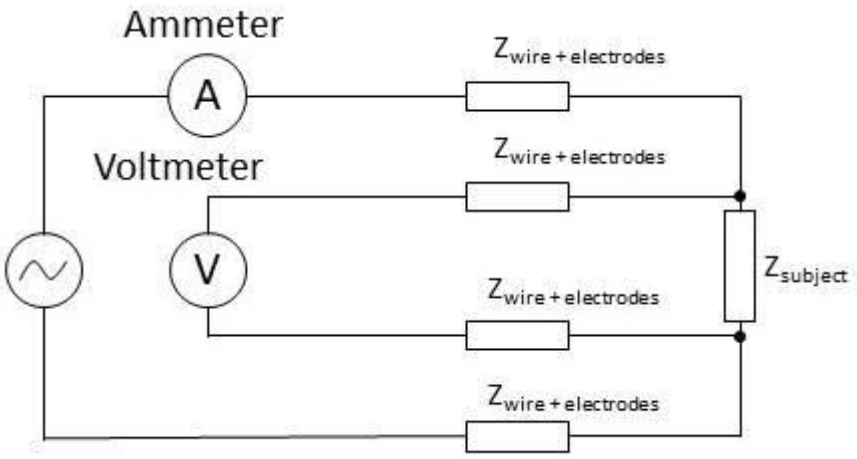

Figure 11: Four-wire measurement circuit. Adapted from [54, p. 283].

2.3 Electrocardiography

Electrocardiograph (ECG) measures the potential changes generated by the heart, therefore providing information on the electrical activity of the heart. ECG consists of P, Q, R, S and T waves, and P-R segment and S-T segment (see Figure 12). These waves and segments are produced by the depolarization or repolarization of the myocardial cells.

The electrical activation from the SA node to the AV node generates the P wave (see Figure 12) when the electrical activation travels through internodal tracts and Bachmann's bundle to the left atrium. [14, p. 25] The P-R segment is the time delay between atrial and ventricular activation, and it is used as a baseline of ECG. The QRS complex, which is composed of Q, R and S waves, provides information on the duration of ventricular depolarization. The QRS complex has the highest amplitude, because the mass of the ventricular muscle is greater than the mass of the atrial muscle. During the S-T segment, the atria are relaxed, but the ventricles are contracted. However, the electrical activity of the heart does not change. Usually, the duration of S-T segment shortens when the heart rate increases. The T wave represents the repolarization of the ventricles and its amplitude is lower than the R wave, because the ventricular repolarization is not as synchronous as the depolarization. The ECG might show also a U wave especially if the test subject has a metabolic disturbance like hypokalemia i.e. a condition, in which blood does not contain enough potassium. [55, pp. 73–84]

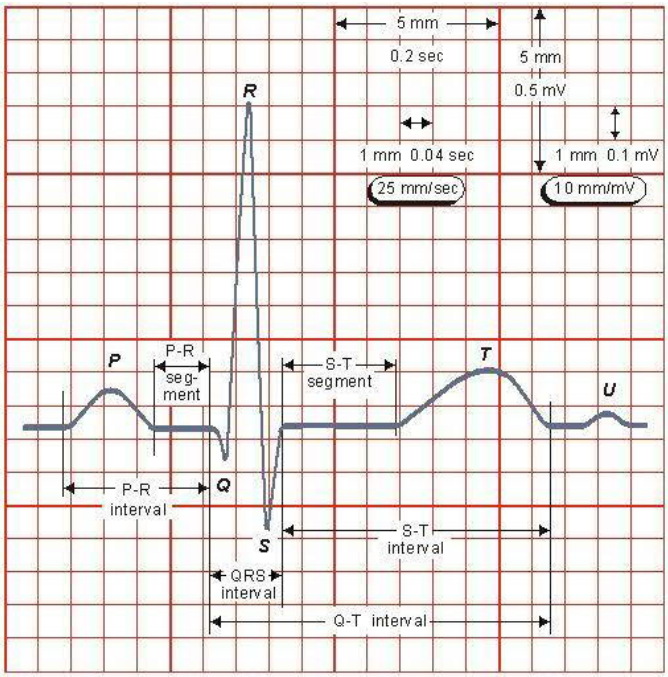

Figure 12: Waves and segments of the ECG [53, p. 284].

2.4 Pulse wave analysis

Pulse wave analysis (PWA) is a technique to get information about endothelial function and mechanical properties of the arterial tree [56]. In the PWA, numerical features are extracted from the pulse wave curves and utilized in the characterization of the condition of the arterial tree. Pulse waves that have similar kind of features can be measured at least in three different techniques: 1) oscillometric or tonometric sensors [57], 2) PPG that measures the changes in blood volume, and 3) IPG and ICG, which were described earlier. In this thesis, the methods that have been proposed for PPG and pressure waves are also employed, because similar kind of features can be found from the pulse waves, independently on the data collection technique.

The shape of the pulse waves changes with age and pathological changes, like atherosclerosis, which cause arteries to stiffen. An example on the changes in central pulse pressure wave morphology with ageing is presented in Figure 13. [58, p. 16, 75]

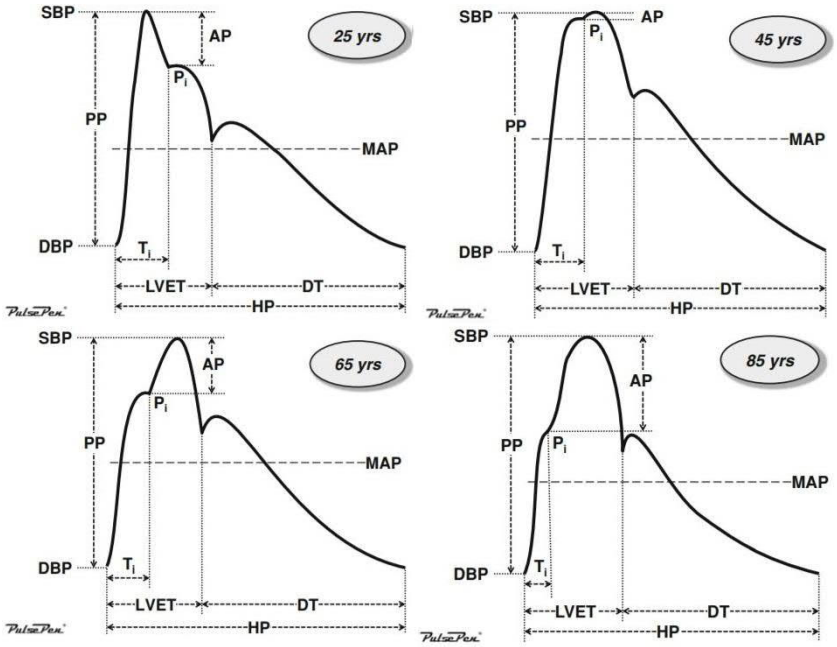

Figure 13: Changes in the central pulse pressure waves caused by ageing. Adapted from [58, pp. 76–79].

In Figure 12, T_i stands for time delay of backward wave, P_i describes inflection point where forward and backward waves meet, *AP* is augmented pressure, *PP* describes pulse pressure (PP = SBP – DBP), *LVET* describes left ventricular ejection time, *DT* describes diastolic time and *HP* is heart period that corresponds to RR-interval.

As can be seen in Figure 13, the ageing shifts the inflection point before the peak systolic pressure. This is due to increased PWV caused by arterial stiffening. The reflected wave is caused by the places where arteries, that have low acoustic impedance, change to arterioles that have high acoustic impedance. The PWV increases even more if the person has a cardiovascular disease. The ageing itself causes the PWV to double between the ages 20 and 100. However, physical fitness decelerates the effects of ageing. Exercising increases the endothelial function and therefore, decelerates PWV. [11]

Different parameters are calculated from the pulse waves. One of the most typical parameters is aortic augmentation index (AIx). AIx describes the incidence of the reflected waves and it is calculated as

$$AIx = \frac{AP}{PP}. \tag{2.9}$$

The values of AIx depend on the inflection point. If the inflection point is after the peak systolic pressure, like in the case of 25-year old person (see Figure 12), the values of AIx are negative. On the other hand, the values of AIx are positive, if the inflection point is before the peak systolic pressure. If the inflection point is not clear, it can be found from the first zero-crossing of the 4th derivative after the first positive peak. [58, pp. 73–75]

In addition to AIx, it is possible to distinguish four different types of pulse wave forms from the pulse waves. Different types of pulse waves are presented in Table 2. In Figure 13, the type A is left lower corner, type B is right upper corner, type C is left upper corner and type D is right lower corner.

Table 2: Classification of pulse wave by its shape [58, p. 80].

	AIx (%)	**Timing of reflected waves**	**Age (years)**	**Diastolic wave-form**
Type A	>12	Protosystole	>40, <65	Concave
Type B	>0, <12	Mesosystole	>30, <40	Convex
Type C	<0	End-systole	<30	Convex
Type D	>>12	Early protosystole	>65	Concave

3. DATA SCIENCE BACKGROUND

3.1 Basic terms of data science

The aim in the implementation of the machine learning methods is to determine the class of the data point by using multiple features. The class is often called label and measurements are often called predictors or features. In training phase, the classifier takes in both feature matrix and label vector. The columns of the feature matrix contain features and rows containing information about different data points i.e. observations. Classification refers to the implementation of a supervised learning method and its aim is to classify the data point according to one of the pre-defined classes, such as "Cat" or "Dog", or "Male" or "Female". A supervised learning method generates a function that maps the input into desired outputs. The mapping function should be approximated so that when a new input data is introduced to the mapping function, the mapping function should be able to predict the correct label for the output data. The outcome, also called a target, is the predicted class label. Overlearning, also known as overfitting, describes that the model does not generalize i.e. it memorizes the training data and then fails with the testing data.

Supervised machine learning methods are based on the statistical properties of their input data. Statistical properties such as variance and statistical independence of multidimensional data are described as a covariance matrix $\boldsymbol{\Sigma}$. Covariance describes the amount of dependency between two variables. Naïve Bayes and discriminant analyses use covariance matrices. Theoretically covariance matrix is defined as

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \left[\mathrm{Cov}\left(X_i, X_j\right)\right] = \left[\mathrm{E}\left[\left(X_i - \mathrm{E}[X_i]\right)\left(X_j - \mathrm{E}\left[X_\mathrm{j}\right]\right)\right]\right], \qquad (3.1)$$

where $\boldsymbol{\Sigma}$ is covariance matrix, $\boldsymbol{X}$ is p-dimensional random vector, X_i and X_j are i^{th} and j^{th} random variables, $\mathrm{E}[X_i]$ is mean, $\mathrm{E}\left[X_j\right]$ is mean and $\mathrm{E}[X]$ is expectation value [59]. The diagonal values of covariance matrix are variances. In classification, the feature matrix is considered as a collection of random variables as

$$\boldsymbol{X} = \left[X_1, X_2, \ldots, X_p\right]$$

Hence, the covariance matrix is presented in matrix notation as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \ldots & \mathrm{Var}(X_n) \end{bmatrix}.$$

3.2 Classifiers

The following seven commonly used classifiers were used in this thesis 1) Naïve Bayesian, 2) *k*-nearest neighbor, 3) linear discriminant analysis, 4) quadratic discriminant analysis, 5) support vector machine, 6) Random Forest, and 7) Adaptive Boosting. The theory behind each classifier is described in the following paragraphs.

Naïve Bayesian

Naïve Bayesian is a classifier, which assumes that its features are statistically independent meaning that all off-diagonal terms of covariance matrix are set to zero. However, the assumption is not always valid, hence the name of the classifier. The naïve Bayes calculates the probabilities for every feature and selects the one, which has the highest probability. The naïve Bayesian uses Bayes theorem that can be presented in following way

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}, \tag{3.2}$$

where $P(c|x)$ is the posterior probability of class c given by predictor x, $P(x|c)$ is likelihood that describes the probability of predictor x given class c, $P(c)$ is the prior probability of class c and $P(x)$ is the prior probability of predictor x. Naïve Bayesian classifier creates relation between the variables and the labels. It is simple and easy to implement, so the classifier can be used as a benchmark for comparing the performance of other classifiers. Moreover, sometimes the naïve Bayesian classifier might even outperform classifiers that are more complicated, because the bias of class density does not affect the posterior probabilities. In Figure 14, Fisher's iris dataset that contains three classes (types of Iris flower): setosa, versicolor and virginica [60], has been classified. [61, p. 287]–[64, pp. 210–211]

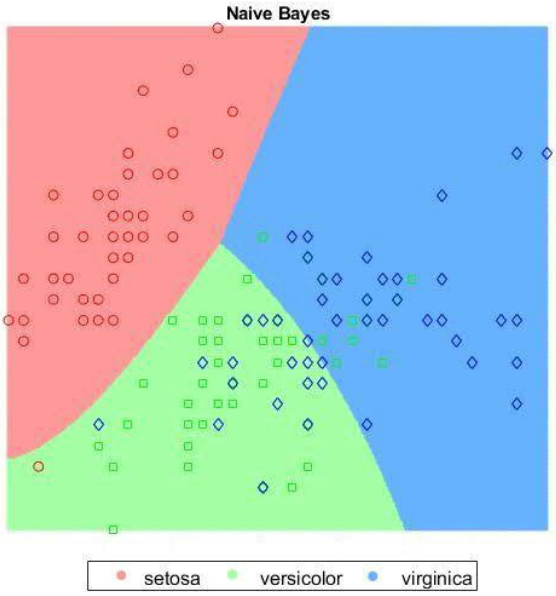

Figure 14: Classification of Fisher's iris dataset with Naïve Bayes.

k-Nearest Neighbor

k-nearest neighbor (k-NN) classifier tries to classify an unlabeled data point by memorizing the training set. It tries to label the data point by choosing the best match based on the training set. k-NN does not learn anything from the training set and is therefore, called a lazy learner. The *k* refers to the number of neighbors the classifier has to take into consideration, for instance if $k = 1$, the classifier takes the first nearest neighbor and adopts its label to the unlabeled data point. However, to avoid incorrect classifying, it is better that *k* has a greater number. [63, pp. 99–111]

One major problem with k-NN is that it can overlearn easily. Especially the 1-nearest neighbor can form isolated regions around single data points. Therefore, the bias of the estimate is low, and the variance is high [64, p. 465]. Figure 15 shows the difference between 1-NN and 10-NN. The left image is 1-NN and the right one is 10-NN.

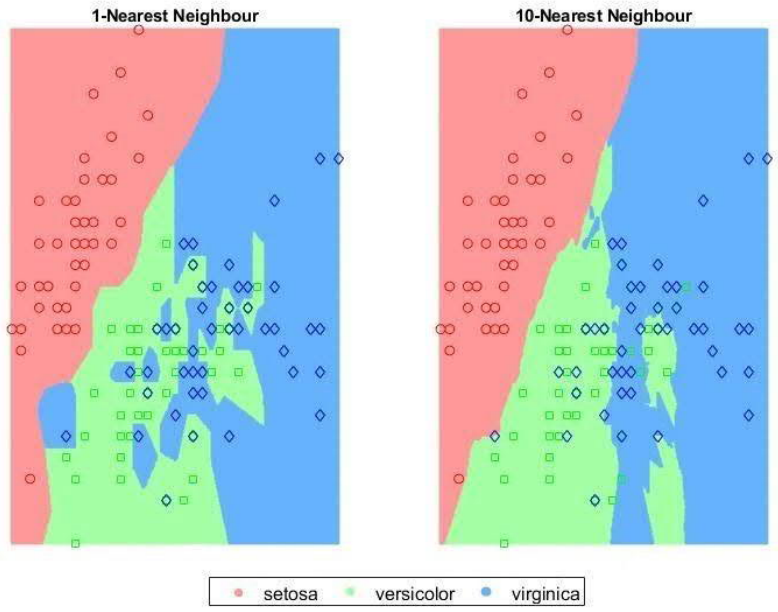

Figure 15: Difference between 1-nearest neighbor and 10-nearest neighbor.

The red, green and blue points are the actual data points, and the light red, light green and light blue areas are the areas, where the data points are classified as belonging to a corresponding class. Moreover, as can be seen in Figure 15, the 1-nearest neighbor has much more local regions compared to the 10-nearest neighbor has.

Linear discriminant analysis

Linear discriminant analysis (LDA) forms linear boundaries between the classes. When the classifier tries to classify a labeled data point, the classifier chooses the class that has the smallest misclassification cost. The linear decision surface, which is also called discriminant function, can be represented as

$$y(x) = w^T x + w_0, \tag{3.3}$$

where w is weight vector, w_0 is threshold weight and x is multidimensional data [62, p. 222], [65, pp. 181–182]. In the discriminant function, w determines the orientation of the hyperplane that divides the feature space and w_0 is the distance of the data point from the hyperplane. The weight vector w is actually a linear projection of x. In this thesis, only two classes are used, hence the LDA presented is for two class case. The weight vector w can be calculated with

$$w = (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1}(\mu_1 - \mu_2), \tag{3.4}$$

where $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are covariance matrices of the samples of class 1 and class 2, respectively, and μ_1 and μ_2 are mean vectors of the samples of class 1 and class 2. However,

linear discriminant analysis assumes that covariance matrix is identical to all classes i.e. $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}, \forall j$. [62, p. 42]

In the two-class case the threshold weight can be calculated

$$w_0 = -\log\left(\frac{p(\omega_2)}{p(\omega_1)}\right) - \frac{1}{2}(\mu_1 - \mu_2)^T w, \tag{3.5}$$

where $p(\omega_1)$ and $p(\omega_2)$ are class priors. However, if the $p(\omega_1)$ and $p(\omega_2)$ are equal, the ideal cut point is in the middle of the projected means. If the $p(\omega_1)$ and $p(\omega_2)$ are not equal, moving the cut point towards smaller class, improves the error-rate. The hyper-plane is perpendicular to the weight vector w and it goes through the threshold. Figure 16 illustrates the hyperplane. When a new data point is presented to the model, the model calculates if the data point is greater or smaller than the threshold. [62, pp. 227–228], [64, pp. 117–119]

The linear discriminant analysis assumes features to be a Gaussian distributed. The Gaussian distribution is defined as

$$f_j(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_j\right|^{1/2}} e^{-\frac{1}{2}(x-\mu_j)^T \boldsymbol{\Sigma}_j{}^T (x-\mu_j)}, \tag{3.6}$$

where x is multidimensional data, d is dimensionality of data, $\boldsymbol{\Sigma}_j$ is covariance matrix of data and μ_j is mean vector of data. [62, p. 22, 36], [64, p. 108]

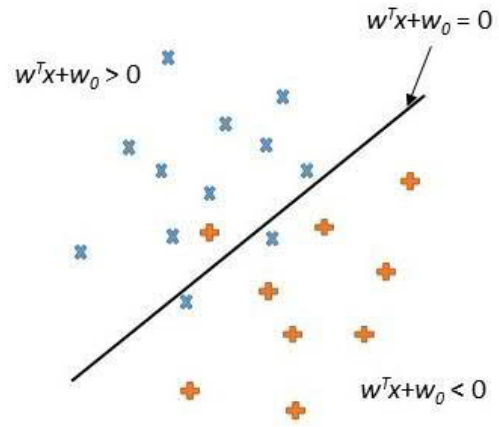

Figure 16: Hyperplane. Adapted from [66].

Quadratic discriminant analysis

The basic principle behind quadratic discriminant analysis (QDA) is similar than the principle behind LDA as they both are based on Gaussian distributions. However, the QDA function is

$$g_j(\boldsymbol{x}) = \log\left(p(\omega_j)\right) - \frac{1}{2}\log\left(|\boldsymbol{\Sigma}_j|\right) - \frac{1}{2}(x-\mu_j)^T\boldsymbol{\Sigma}_j^{-1}(x-\mu_j), \qquad (3.7)$$

where j is the number of class, $p(\omega_j)$ is probability of class, x is multidimensional data, $\boldsymbol{\Sigma}_j$ is covariance matrix of data and μ_j is mean vector of data. The QDA does not assume that covariance matrix is identical to all classes. Instead of using linear boundaries, the QDA allows the boundaries to be quadratic. Otherwise, the QDA and LDA work in a similar way. Figure 17 presents the differences between LDA and QDA on the Fisher's iris dataset. [62, p. 36], [64, p. 110]

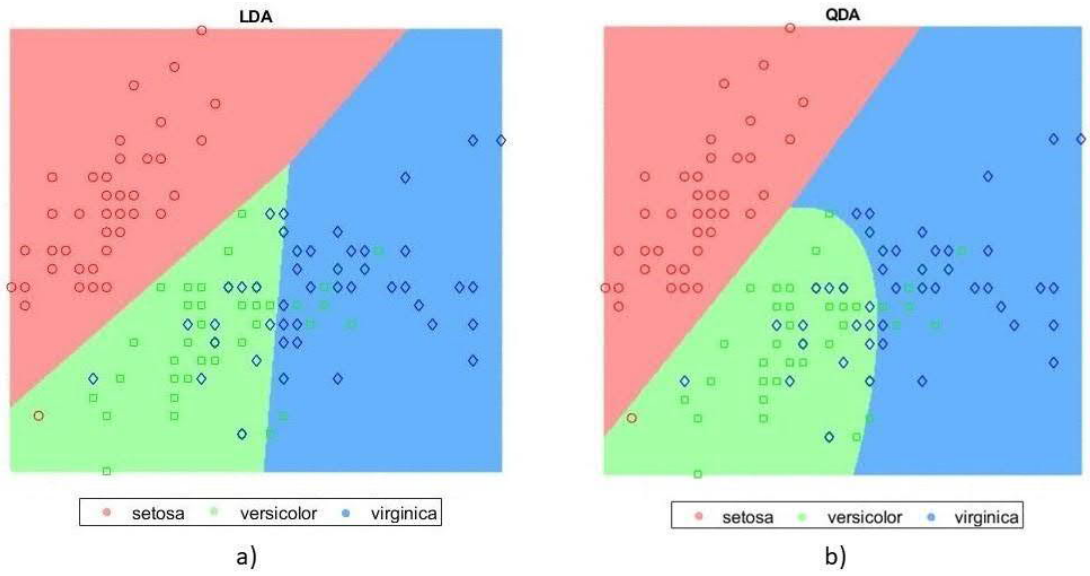

Figure 17: Difference on hyperplanes of LDA (a) and QDA (b).

Support vector machine

Support vector machine (SVM) tries to maximize the margin M – the smallest distance between hyperplane and any of the samples, between the classes, which means that the hyperplane is as far from the data points as possible. Points that define the place of the margin are called support vectors as can be seen in Figure 18 (a). If the classes are separable, points cannot be found between the margins as is in Figure 17. However, usually the classes are not separable and data points may be on the wrong side of the margins (see Figure 18 (b)). Hence, these points are called slack variables ξ_i , $i = 1,2,\dots,n$. If a slack variable equals to zero, the data point is either on the right side of the margin or on the

margin. However, if the slack variable is $0 < \xi < 1$, the data point is on the wrong side of the margin but on the right side of the hyperplane. Lastly, if the slack variable is greater than one, the data point is on the wrong side of the hyperplane. [64, pp. 417–418], [65, p. 331]

Nevertheless, finding support vectors is an optimization problem:

$$\max_{w,w_0,\|w\|=1} M$$
$$\text{Subject to } y_i(x_i^T w + w_0) \geq M(1-\xi), \text{for } i = 1,2,\dots,N \tag{3.8}$$

M is the margin, y_i is discriminant function, x_i is multidimensional data, w is weight vector, w_0 is threshold weight and ξ is slack term. The second condition can be written in other form as

$$\begin{cases} x_i^T w + w_0 \geq M, & \text{if } y_i = 1 \\ x_i^T w + w_0 < M, & \text{if } y_i = -1 \end{cases} \tag{3.9}$$

Eq. (3.8) can be expressed as

$$\min_{\boldsymbol{w},w_0} \|w^T w\|$$
$$\text{Subject to } \begin{cases} y_i(x_i^T w + w_0) \geq 1 - \xi_i, \text{for } i = 1,2,\dots,N \\ \xi_i \geq 0, \sum \xi_i \leq \text{constant} \end{cases} \tag{3.10}$$

Eq. (3.10) is can be re-expressed in other form

$$\min_{w,w_0} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i$$
$$\text{Subject to } \xi_i \geq 0, y_i(x_i^T w + w_0) \geq 1 - \xi_i, \forall i, \tag{3.11}$$

where C is a cost parameter. Eq. (3.11) can be solved with Lagrange function.

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i[y_i(x_i^T w + w_0) - (1-\xi_i)] - \sum_{i=1}^{N} r_i\xi_i, \tag{3.12}$$

where α_i and r_i are Lagrange multipliers. r_i ensures that ξ_i is positive. The minima of term $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i$ with respect to w, w_0 and ξ_i are found by differentiating with respect to L_p and setting respective terms to zero. This results in the following conditions:

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{3.13}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{3.14}$$

$$\alpha_i = C - r_i, \forall i \tag{3.15}$$

When Eq. (3.13)–Eq. (3.15) are substituted to Eq. (3.12), the result is a dual form of Lagrange function

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i\prime=1}^{N} \alpha_i \alpha_{i\prime} y_i y_{i\prime} x_i^T x_{i\prime}, \tag{3.16}$$

which is maximized respect to $0 \leq \alpha_i \leq C$ and Eq. (3.14). Karush–Kuhn–Tucker conditions (see [64, p. 420]) provide conditions that need to be fulfilled. These conditions include constraints

$$\alpha_i [y_i (x_i^T w + w_0) - (1 - \xi_i)] = 0 \tag{3.17}$$

$$\mu_i \xi_i = 0 \tag{3.18}$$

$$y_i (x_i w + w_0) - (1 - \xi_i) \geq 0 \tag{3.19}$$

Those nonzero values of α_i that met the constraints of Eq. (3.19) are called support vectors. Now, the w_0 can be calculated from Eq. (3.17) by using any of the margin points ($\xi_i = 0$). The sign of Eq. (3.3) selects the class for new data points. [64, pp. 418–420], [65, pp. 331–334]

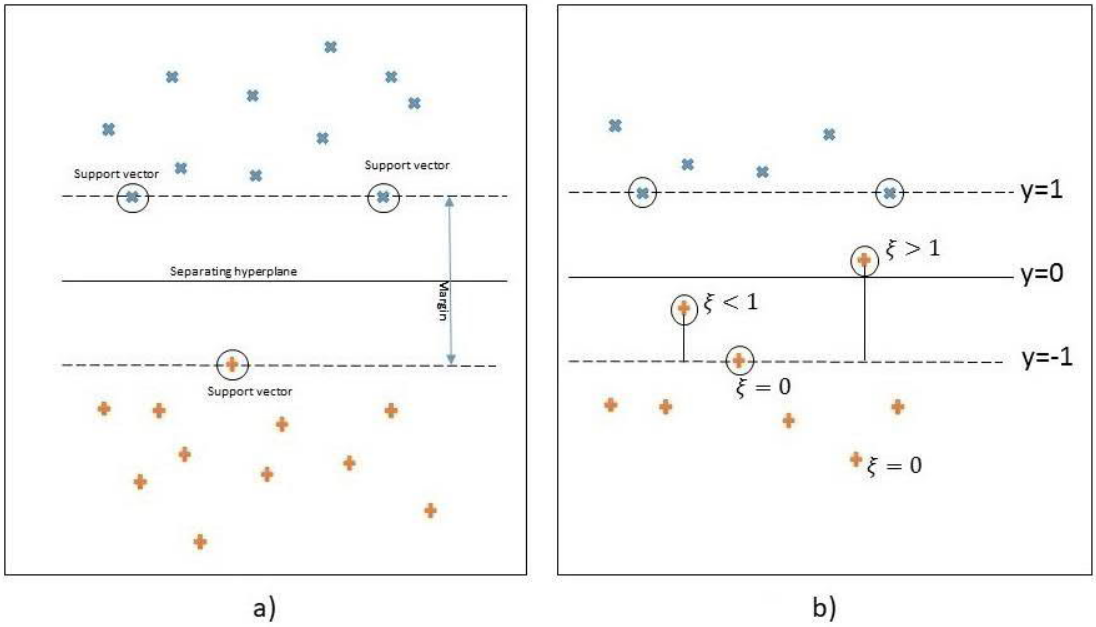

Figure 18: Separable class (a) and non-separable class (b). Adapted from [67] (a) and [65, p. 332] (b).

Eq. (3.3) can be changed into following form with Eq. (3.13)

$$f(x) = x_i^T w + w_0$$
$$= \sum_{i=1}^{N} \alpha_i y_i \langle x, x_i \rangle + w_0, \tag{3.20}$$

where $\langle x, x_i \rangle$ is kernel function $K(x, x')$. Kernel function is used to maximize the margin by mapping the feature vectors into high-dimensional space, where a hyperplane can be constructed. This is known as the kernel trick. For instance, two-dimensional samples are moved into three-dimensional space, where a hyperplane can be constructed. The classification of the Fisher's iris data set is presented in Figure 19. [64, pp. 417–429], [65, pp. 331–334]

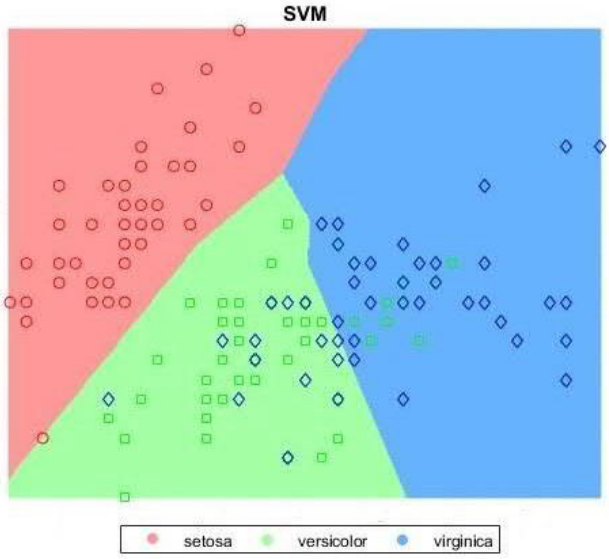

Figure 19: Classification of Fisher's iris dataset with support vector machine.

Random Forest

Random Forest consists of collection of decision trees and it was developed in 2001 by Leo Breiman [68]. Random Forest reduces correlation between the decision trees without increasing variance too much. In Random Forest, each tree is trained with a subset of samples and a subset of features. Therefore, some features and samples (rows in feature matrix) are hidden from the training. Figure 20 presents an example of a decision tree. [62, pp. 389–390], [64, pp. 587–589]

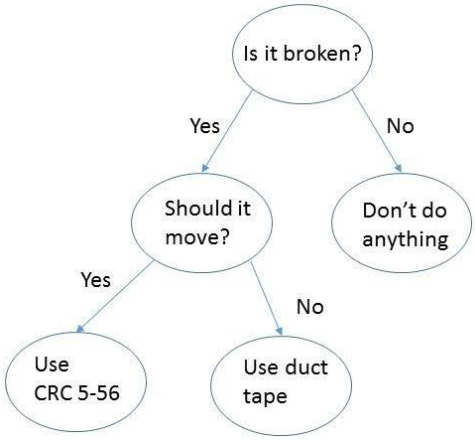

Figure 20: Example of a decision tree.

Training with subsets of data, results that some of the samples are left for testing the classifier. The label of unlabeled data point is predicted by taking majority vote from the outputs of different decision trees The Fisher's iris data set is classified in Figure 21. [62, pp. 389–390], [64, pp. 587–589]

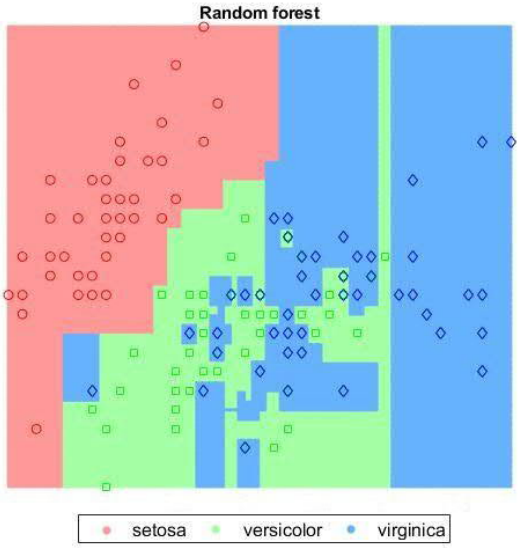

Figure 21: Classification of Fisher's iris data set with Random Forest.

Adaptive Boosting

Adaptive Boosting, more commonly known as AdaBoost, is an ensemble method that uses weaker classifiers like decision trees. AdaBoost was developed in 1997 by Yoav Freund and Robert Schapire [69]. AdaBoost gives a weight for each sample and grows a weak classifier after the classifiers are trained. The label is predicted by weighted majority vote. This is expressed mathematically as

$$G(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right), \tag{3.21}$$

where α_m is weight by AdaBoost and $G_m(x)$ is a weak classifier. The Figure 22 shows the schematic representation of AdaBoost. After first iteration (m=1 in Eq. (3.21)), all the weights are modified individually. The modification gives greater weights to those data points that were misclassified by classifier $G_{m-1}(x)$ and decreases weights for those that were classified correctly. Hence, the weights of data points that are difficult to classify increase in every iteration. [64, pp. 337–339]

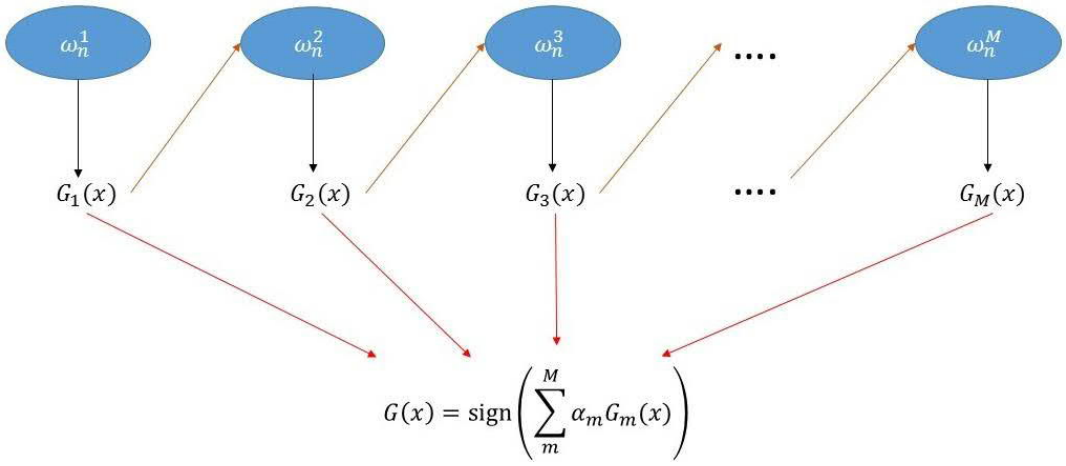

Figure 22: Schematic representation of AdaBoost. Adapted from [65] p.658.

The weights ω_n^m are dependent on the previous classifier (brown arrows) that are trained with weighted training set (black arrows). These are then combined to produce the prediction. The classification of the Fisher's iris data set is presented in Figure 23 [64, p. 339], [65, pp. 657–659]

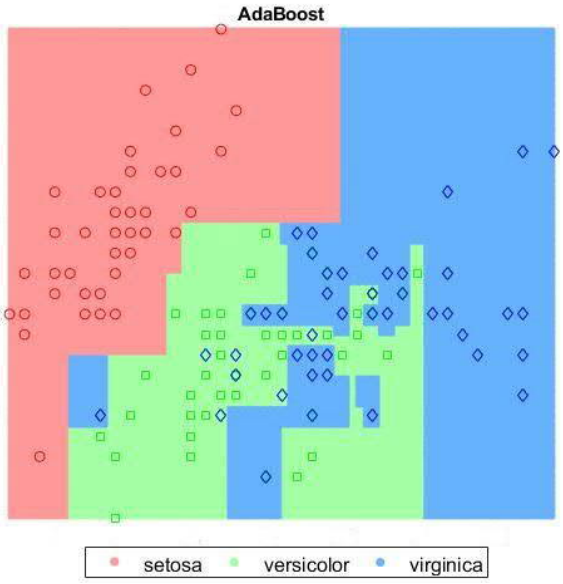

Figure 23: Classification of Fisher's iris dataset with AdaBoost.

3.3 Cross-validation

Cross-validation is a method that is used to estimate prediction error. In other words, the generalization of the classifier is tested. Usually, the amount of data is not great enough for the dataset to be divided into training and testing sets and therefore, the cross-validation is implemented. Leave-one-out (LOOCV), hold-out, and *k*-fold cross-validation are the most commonly used iterative cross-validation methods. LOOCV leaves out only one

data point from training and uses that specific data point for testing. Hold-out cross-validation divides the dataset into two datasets: training and testing.

K-fold cross-validation divides the data in *k* parts and uses one of the *k*th parts for testing and the rest *k*-1 parts for training the classifier. In each *k* iteration, *k*-1 parts of the dataset are utilized as training data and the remaining part of the dataset is utilized in cross-validation. Figure 24 presents how testing and training data sets can be formed for the whole available dataset.

Figure 24: 5-fold cross validation.

The *k*-fold cross validation is useful when the amount of data is not great enough to divide it into training and testing sets [64, p. 241]. The classifier should be able to generalize and therefore the *k*-fold cross-validation is utilized. The most common values for *k* are 5 or 10. The advantage of k-fold cross validation is that every set is one time the test set and the training set *k*-1 times. However, the disadvantage of the *k*-fold cross validation is that the training must be done every time all over again, which takes computation time.

3.4 Forward selection

When the feature matrix contains many features, the feature selection becomes suitable [64, p. 58]. Feature selection seeks those features that are the most relevant for the model making a subset of features from the whole set of features. However, it is not same as reducing dimensions, because dimension reduction may create new features. One of the feature selection methods is forward selection, which will be discussed here. [70, p. 271]

Forward selection is a procedure, where the features that give the best improvement for the model are selected. Forward selection stops when adding the variables into model does not improve the model. The improvement is determined with criterion that could be, for instance, a misclassification rate. In such case, the forward selection tries to decrease the misclassification rate. Forward selection tests all the variables and calculates the criterion value, selecting then the one with the best criterion value. Forward selection starts with zero variables in the model. The first variable is the one, which has the strongest

correlation with the outcome. The main reason for using forward selection for the feature selection is to improve the classification performance. [70, pp. 13–14]

3.5 ROC curve and AUC

Receiver operating characteristic (ROC) curve and area under curve (AUC) are used in this thesis to evaluate the classification results. Moreover, both ROC curve and AUC can be used with either binary classification or multi-class classification. ROC curve shows the relationship between true positive rate (TPR) and false positive rate (FPR). The horizontal axis is FPR and the vertical axis is TPR. The values of TPR and FPR can be calculated from a confusion matrix (see Table 3), which contains information about actual and predicted classes. If the classifier classifies all classes correctly, the non-zero values are only in the diagonal from the upper left to the lower right corner. [62, p. 415], [71, p. 515], [72, p. 239]

Table 3: Confusion matrix.

		Actual value	
		Positive	Negative
Predicted value	Positive	True positive	False positive
	Negative	False negative	True negative

TPR is defined as

$$TPR = \frac{True\ positive}{True\ positive + false\ negative}. \tag{3.22}$$

FPR can be defined as

$$FPR = 1 - \frac{True\ negative}{True\ negative + false\ positive}. \tag{3.23}$$

or FPR can be defined as

$$FPR = \frac{False\ positive}{True\ negative + False\ positive}. \tag{3.24}$$

Figure 25 illustrates how the values in Eqs. (3.22–3.24) are chosen from the probability distributions when the threshold is known. The blue curve represents positive class, for instance, people with cancer and orange curve represents a negative class, for instance, people without cancer. Therefore, the false negative means that a person who has cancer is diagnosed as healthy and false positive means that a person who does not have cancer is diagnosed as cancer patient. In Figure 25, the threshold value is 0.5. Figure 26 shows the ROC curve, where the grey line is the expected value of a random guess and the violet line is ROC curve that is above the random guess. The ROC curve can be drawn when the threshold is moved and then TPR and FPR are checked with each threshold value.

AUC is calculated from the area that falls under the ROC curve. AUC provides single value that describes the performance of the classifier. The values of AUC are from 0 to 1, where AUC of 1 is the ideal situation. The AUC value of 0.5 means random guess.

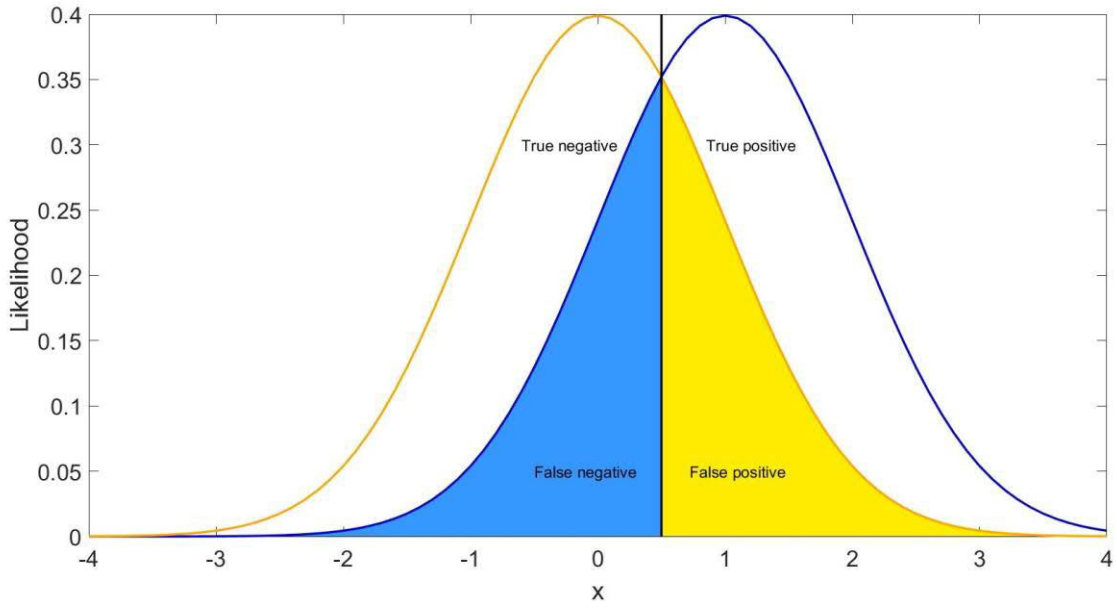

Figure 25: Overlapping probability density functions of a classifying parameter for positive (blue curve) and negative (yellow curve) classes.

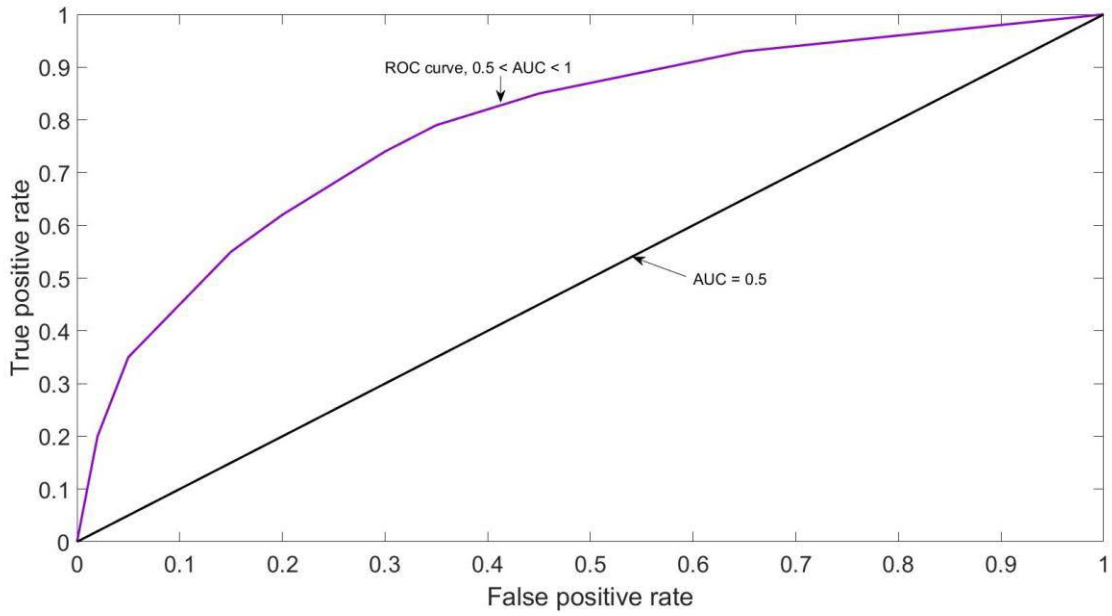

Figure 26: Examples on ROC curve.

Let us assume that we have two classes: class 0 and class 1. The class 0 is *Negatives* and class 1 is *Positives* in Table 3 *Actual value*. With these assumptions, TPR is calculated from the class 1 and FPR is calculated from class 0. In MATLAB, parameter called `score` is used to describe the probability of the sample being in class 0 and class 1. The left column of the `score` is the probability of the sample to be in class 0 and the right column is the probability of the sample to be class 1. Let us choose that we aim to classify according to class 0, i.e. we take probabilities from left column of `score`. Firstly, we choose a threshold value. If the threshold value is lower than the probabilities of the samples, the predicted label is 1. If the threshold is over the probability, the predicted label is 0. Secondly, TPR and FPR are calculated with Eqs. (3.22–3.24). These TPR and FPR values present one point in ROC space. Then the threshold is increased, the labels are predicted again and TRP and FPR are calculated. In order to draw a ROC curve, threshold can be increased as many times as it is necessary.

4. MATERIALS AND METHODS

4.1 Cardiovascular Risk in Young Finns Study dataset

Cardiovascular Risk in Young Finns Study

Dataset for this thesis has been collected as a part of The Cardiovascular Risk in Young Finns Study also known as Young Finns Study (YFS), which started in the 1980s and is still ongoing [2]. The data analyzed in this thesis has been collected in 2007 at Tampere University Hospital, Turku University Hospital, Helsinki University Hospital, Kuopio University Hospital and Oulu University Hospital.

In 2007, the YFS-participants had a 27-year follow-up [2]. 1853 participants, aged 30 to 45, took part in the clinical examinations. To ensure that the participants remain anonymous, they were divided into six age groups based on their age. The age groups were 30, 33, 36, 39, 42 and 45 years old. As a part of the study protocol ECG, IPG and ICG were measured from all participants. However, in this thesis, 13 participants were excluded from the dataset, because the SNR of the IPG signals was poor, or the measurement had somehow failed. In addition to the measurements of ECG, IPG and ICG, there was clinical data from the participants.

Clinical data

The dataset included clinical data from the participants. This data was used to label and classify people who have low and high risk for cardiovascular diseases. The characteristics for the clinical data are presented in Table 4 to illustrate the clinical values of the subject population. The data is presented for each research center and for all 1840 participants. The data contains two different blood pressure measurements, one being measured during the recording of the physiological signals and the other at some other time. The FMD% is defined as a relative dilation of the carotid artery from the initial point to the point where the brachial artery is dilated as much as possible.

Table 4: Clinical data used in this thesis. Values are mean ± standard deviation, or median (25^{th} to 75^{th} percentile), or proportions. Letter c before the parameter refers that the parameter was measured during the physiological signal collection.

Variable	Values
Gender (males/females)	832/1008
Age (years)	37.7 ± 5.0
Height (cm)	172.0 ± 9.2
Mass (kg)	77.2 ± 16.9
BMI (kg/m^2)	26.0 ± 4.8
cSystolic BP (mmHg)	125.5 ± 13.8
cDiastolic BP (mmHg)	76.8 ± 9.2
cMAP (mmHg)	93.0 ± 10.1
Systolic BP (mmHg)	120.5 ± 14.3
Diastolic BP (mmHg)	75.5 ± 11.3
MAP (mmHg)	90.5 ± 11.6
Total cholesterol (mmol/l)	5.0 ± 0.9
LDL (mmol/l)	3.1 ± 0.8
HDL (mmol/l)	1.3 ± 0.3
Triglycerides (mmol/l)	1.2 (0.9–1.7)
Glucose (mmol/l)	5.3 (4.9–5.6)
Insulin (mmol/l)	6.9 (4.2–10.9)
PWV (m/s)	5.2 ± 1.5
IMT (mm)	0.6 ± 0.1
FMD% (%)	9.0 ± 4.6
Smoking (%)	18.9
Plaque (%)	2.2
Hypertension (%)	5.9
Antihypertensive medication (%)	6.8

Bioimpedance measurements

The bioimpedance measurements were conducted by using a non-invasive cardiovascular monitoring device CircMon (JR Medical Ltd, Tallinn, Estonia) [7]. CircMon measures ECG, whole-body ICG and IPG. There were 12 electrodes in total placed on the test subject's body (see Figure 27).

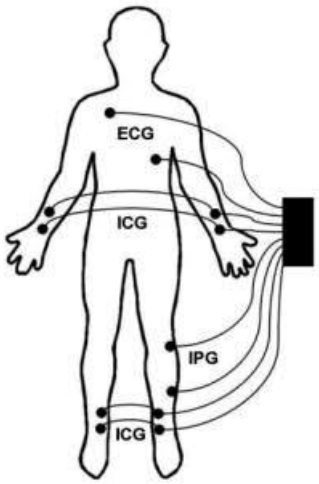

Figure 27: Placement of electrodes. Adapted from [73].

ECG was measured with two electrodes placed on the patient's thorax. ICG was measured from the upper and lower limbs. There were two electrodes on each upper limb and each lower limb: in each limb, one electrode was voltage-sensing electrode and the other electrode was current-feeding electrode. The voltage-sensing electrodes on hands were connected to each other and these electrodes were connected to CircMon device (see Figure 27). The current-feeding electrodes were connected to each other in a similar way and to the CircMon device. On the lower limbs, the voltage and the current electrodes were connected in similarly as on the hands. The electrodes of IPG were placed on the popliteal artery and on the shin, see Figure 27.

4.2 Processing of IPG and ICG signals

Processing of the data

The IPG and ICG signals were interpolated to 400 Hz for improving the time resolution and for enabling the usage of already implemented signal analysis algorithms used in [3], [10] assuming integer-based sampling rates. Firstly, the R-peaks were detected from the ECG signal. Because it is known that an ICG or IPG pulse wave appears after an R-peak, the starting point of the pulse wave is determined to be a local minimum that is right after the R-peak i.e. the starting point of the next pulse wave. The ending point of the pulse wave is the local minimum, also called a foot, after the next R-peak. The maximum of the pulse wave was determined to be the local maximum of the pulse wave that appears after the R-peak.

In most of the subjects, the IPG signal resembles the first derivative of volume pulse signal that is utilized in pulse wave analysis and for which the exploited pulse wave analysis algorithms were originally developed. Therefore, to produce pulse waves resembling conventional volume pulse waves, the IPG signals were transformed by using trapezoidal

numerical integration method. An example of a transformed IPG signal is shown in Figure 28. The pulse waves are detected from the signal as explained earlier.

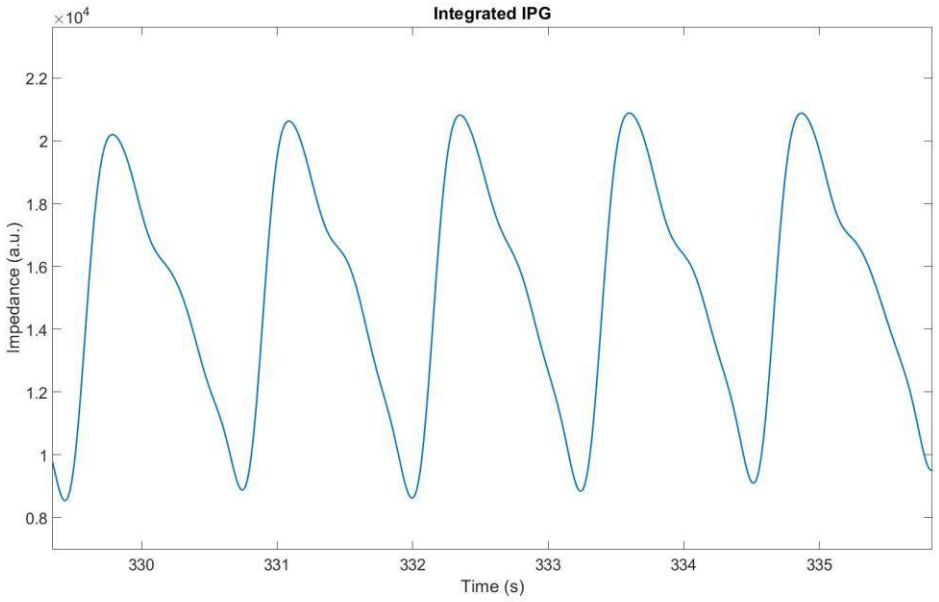

Figure 28: Transformed IPG signal.

When the foots and peaks of the pulse waves are recognized, the parameters can be calculated. From these three different (ICG, IPG and transformed IPG) signals 80, parameters were calculated. 27 parameters were calculated from transformed IPG signal, 27 calculated from ICG signal and 26 from IPG signal.

Pulse wave parameters

Fast Fourier Transform (FFT) was computed in order to calculate the ratio between the amplitudes of first two harmonic peaks. An example on the two harmonic peaks is presented in Figure 29.

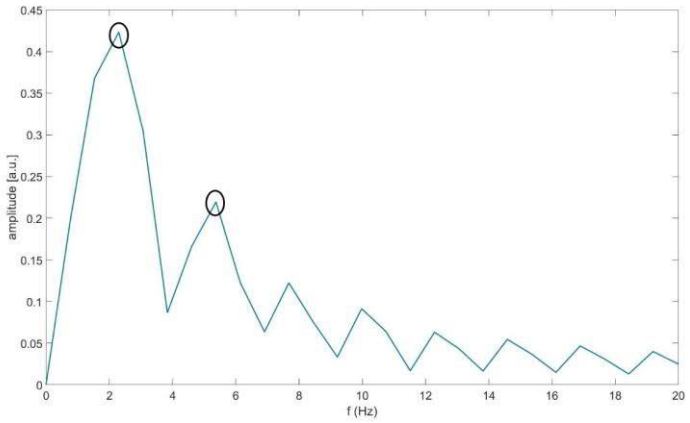

Figure 29: Harmonic peaks of FFT.

In addition to harmonic peaks of FFT, the ratio between the area under amplitude normalized transformed IPG pulse wave curve and the area under the amplitude normalized ICG pulse wave curve as well as the area under amplitude normalized IPG pulse wave curve and the area under amplitude normalized ICG pulse wave curve were computed. Rise time was defined as a time difference between foot point *a* and peak point *b* as illustrated in Figure 30. Besides rise times, height-normalized and unnormalized transmission times were computed from all three pulse wave signals as a time difference between foot points *a* and *c* in Figure 30. Decay time was calculated from all pulse wave signals defined as the time difference between points *b* and *c* in Figure 30.

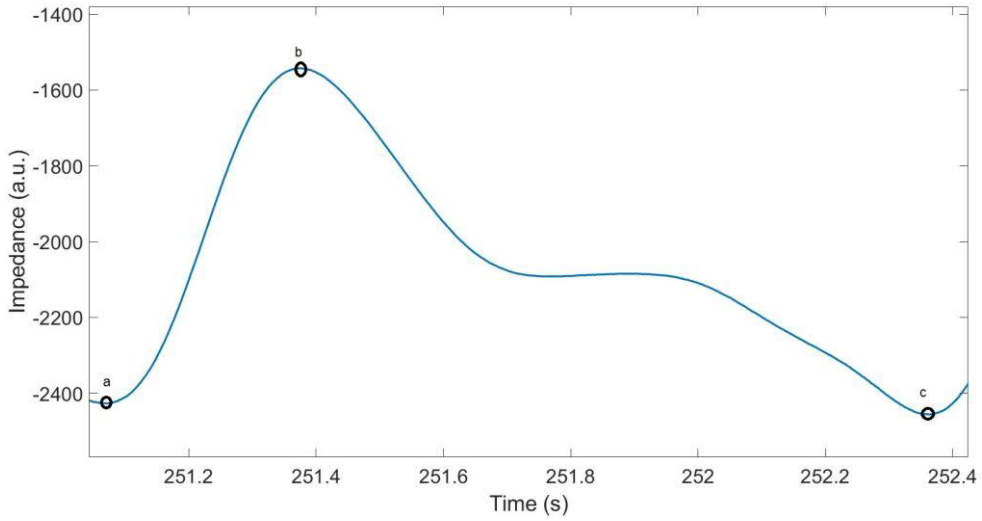

Figure 30: Fiducial points of a pulse wave.

The length of amplitude-normalized discrete-time curve was approximated with Pythagorean Theorem. In order to calculate common parameter describing the shape a pulse waves, three fiducial points P_1, P_2, and B are extracted from the individual pulse waves (see Figure 31 (a)) and their definition of P_1, P_2, and B can be found in [3]. The defined shape-describing fiducial points are utilized in computing amplitude ratios called R_1, R_2, R_3, and R_4 defined as:

$$R_1 = \frac{B}{\max(P_1, P_2)} \tag{4.1}$$

$$R_2 = \frac{B}{P_1} \tag{4.2}$$

$$R_3 = \frac{B}{P_2} \tag{4.3}$$

$$R_4 = \frac{P_2}{P_1} \tag{4.4}$$

Parameter R_1 is originally defined for PPG signals and it is called as Reflection index [5]. Parameter R_4 is originally defined for brachial systolic and diastolic blood pressure and it is called as peripheral augmentation index [74]. Besides the amplitude ratios, time delays between fiducial points P_1, P_2, and B are computed as follows: T_1 is the time delay between systolic maximum and the peak of diastolic wave, T_2 is the time delay between P_1 and B and T_3 is the time delay between P_2 and the B. AGI [3] is defined as the 2nd derivative of the pulse wave and it is calculated as

$$\mathrm{AGI} = \frac{b - c - d - e}{a}. \tag{4.8}$$

The points for AGI are presented on the Figure 31 (b).

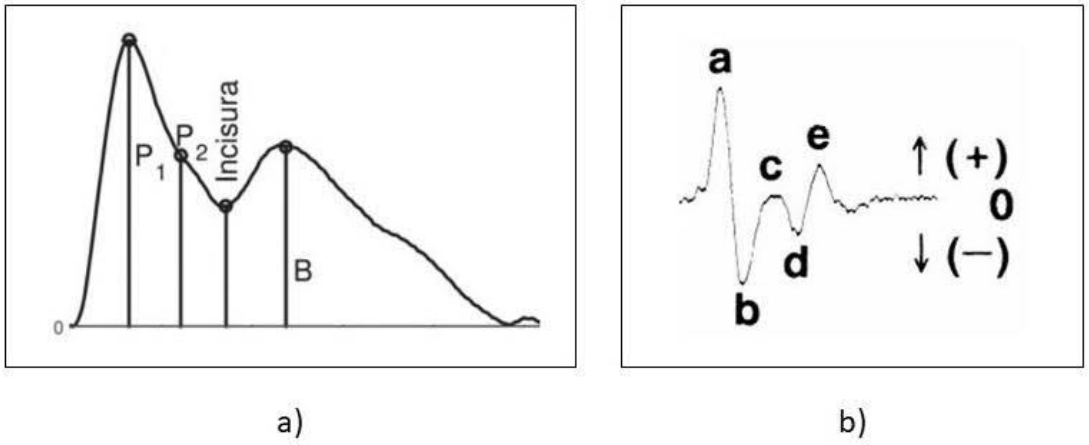

Figure 31: Fiducial points for *R*, *T* (a) and AGI (b). Adapted from [3], [4].

Lastly, parameters were calculated from decompositions of the pulse waves. The two decomposition methods were developed in [10] and they use Levenberg-Marquardt method. One of the decomposition methods uses 5 log-normal curves as basis functions, called *l5* and the other decomposition method uses 1 log-normal curve and 4 Gaussian curves as basis functions, called *gln4*. The parameters that were calculated from these decompositions are:

- the ratio of amplitudes of the highest and second highest decomposition wave
- the ratio of amplitudes of the highest and third highest decomposition wave
- the time difference between the highest and second highest decomposition wave
- the time difference between the highest and third highest decomposition wave
- the ratio of areas of the highest and second highest decomposition wave
- the ratio of areas of the highest and third highest decomposition wave.

The fiducial points of the pulse wave decomposition components are shown in Figure 32.

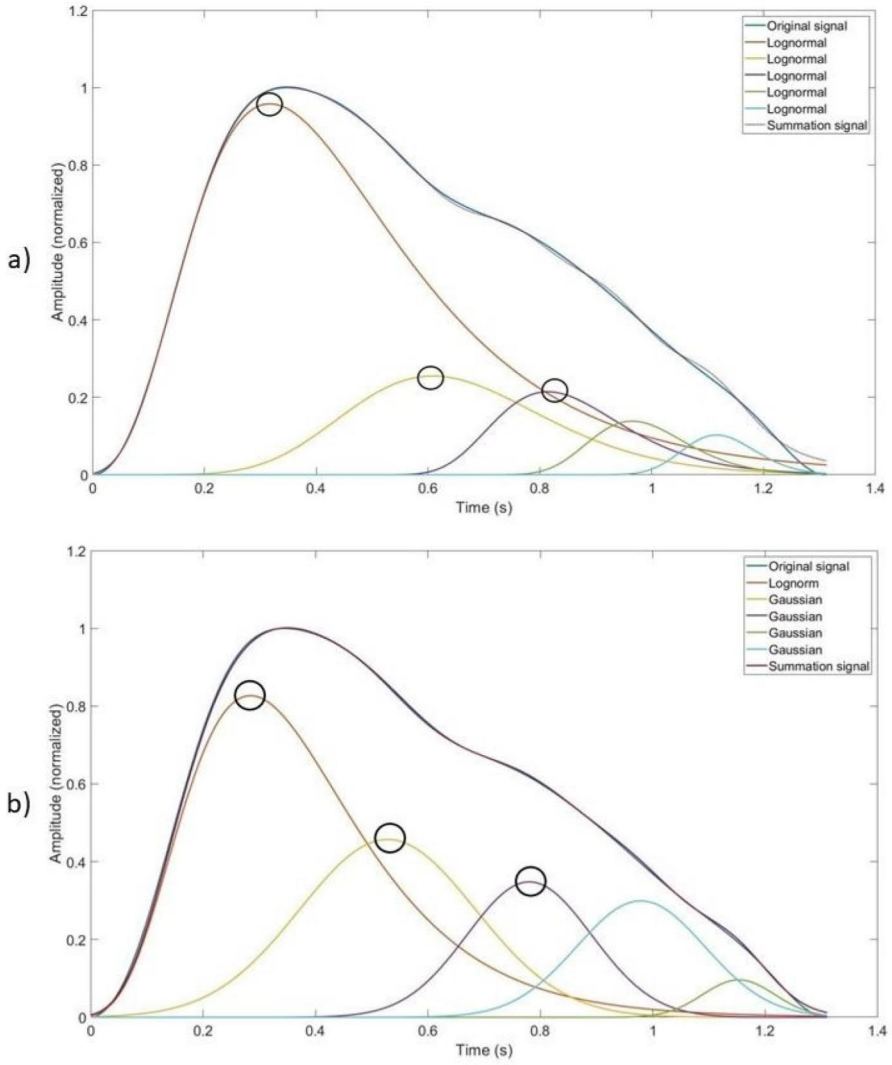

Figure 32: Fiducial points of pulse wave decompositions l5 (a) and gln4 (b).

After the parameters were calculated, a time series was formed for all parameters for all test subjects. An average of time series of each parameter for each test subject was calculated and utilized in the further analysis.

As was mentioned earlier, there were 80 pulse wave parameters calculated in total. However, with some pulse wave parameters, more than 2% of the subjects had extremely deviating parameter values indicating that they are possible outliers and hence, these pulse wave parameters were excluded from the analysis. In the end, 56 pulse wave parameters were accepted into the feature matrix, which then consisted of 20 pulse wave parameters from ICG signal, 14 pulse wave parameters from IPG signal and 20 pulse wave parameters from the transformed IPG signal and 2 parameters that are calculated from the area ratios of 2 signals. The names of these parameters are presented in Appendix B.

4.3 Previous research with the dataset

Koivistoinen *et al*. studied association between PWV, which is a pulse wave parameter, and IMT as well as PWV and FMD in [7]. They used unadjusted model, sex and age adjusted model and multivariable adjusted model to find associations between parameters. Adjusting seeks to take multiple predictors into consideration rather than two predictors when the outcome variable is considered. When the model is unadjusted, it means in this thesis that a model consists of only two predictors. When it is adjusted with some variables, then the model contains multiple predictor variables. The unadjusted model did not include any parameters other than PWV and IMT or PWV and FMD. The key results of the earlier study [7] by Koivistoinen *et al*. are presented in Table 5.

Table 5: Results by Koivistoinen *et al*. Adapted from [7].

	IMT (mm) vs. PWV (m/s)		**FMD (%) vs. PWV (m/s)**	
	$\beta \pm$ SE	***p***	**$\beta \pm$ SE**	***p***
Age 30–36 years (*n* = 825)				
Unadjusted	2.274 ± 0.519	<0.001	−0.015 ± 0.009	0.111
Age and sex-adjusted	1.255 ± 0.507	0.014	0.009 ± 0.009	0.334
Multivariable-adjusted*	0.007 ± 0.497	0.989	−0.001 ± 0.008	0.936
Age 39–45 years (*n* = 929)				
Unadjusted	2.859 ± 0.502	<0.001	−0.051 ± 0.012	<0.001
Age and sex-adjusted	1.398 ± 0.482	0.004	−0.014 ± 0.011	0.202
Multivariable-adjusted*	0.235 ± 0.459	0.609	−0.014 ± 0.010	0.167

* contains: age, sex, systolic blood pressure, HDL, LDL, BMI, triglycerides (log), glucose (log), insulin (log), C-reactive protein

In Table 5, the β refers to the estimate of the regression coefficients, *SE* refers to standard error and *p* is *p*-value. In [7], *p*-values smaller than 0.05 are considered to be statistically significant. After adjusting, IMT and FMD have worse association with PWV than the unadjusted model, as can be seen in Table 5. In this thesis, all calculated pulse wave parameters were tested by means of age- and sex-adjusted, as well as multivariable adjusted models to find out if they have stronger association with FMD and IMT than PWV. The results of the calculated pulse wave parameters are presented in Results.

4.4 Labeling the test subjects

Labeling refers to defining a reference class for the samples. In order to train and characterize the constructed classifiers, the test subjects must be labeled as test subjects who

have a high or low risk for cardiovascular diseases. Three different ways to label the test subjects were tested in this thesis.

First labeling method (LM1) was to use cardiovascular risk factors as proposed in [7]. If the test subject had at least one risk factor, the test subject was labeled as a test subject at high risk for cardiovascular diseases. The cardiovascular risk factors were defined as SBP, HDL cholesterol, LDL cholesterol, BMI, triglycerides, fasting glucose and fasting insulin. People who had value or values equal or greater than sex- and age-specific 80^{th} percentile for BMI, LDL, SBP, triglycerides, fasting glucose or fasting insulin, equal or less than 20^{th} percentile for HDL or who were smokers, were defined as test subjects at high risk for cardiovascular diseases. This labeling method was chosen because this thesis analyzes almost an identical dataset that was used in [7]. The only difference between the risk factors in [7] and this study is that the available data does not contain any information about C-reactive protein, which is one of the clinical reference parameters in [7].

The second labeling method (LM2) was to use Finnish reference values for total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides, fasting glucose and BMI. Reference values for both male and female are presented in Table 6. A person was labeled as a test subject at high risk for cardiovascular disease if at least one of the conditions in Table 6 was fulfilled.

Table 6*:* Conditions of being labeled as test subject at high risk for cardiovascular diseases.

Reference value	Women	Men
Triglycerides	> 1.7 mmol/l	> 1.7 mmol/l
HDL cholesterol	< 1.2 mmol/l	< 1.0 mmol/l
LDL cholesterol	> 3.0 mmol/l	> 3.0 mmol/l
Total cholesterol	> 5.0 mmol/l	> 5.0 mmol/l
Fasting glucose	> 7.0 mmol/l	> 7.0 mmol/l
BMI	> 25 kg/m^2	> 25 kg/m^2

This labeling method was selected, because the reference values are unambiguous, and these reference values are risk factors of type 2 diabetes and atherosclerosis. Threshold value for fasting glucose was selected to be over 7.0 mmol/l, because higher values indicate that the person has either type 1 or type 2 diabetes [75].

In the third labeling method (LM3), the test subjects were labeled based on the existence of plaque in their carotid artery, antihypertensive medication or diagnosed hypertension. This labeling method was chosen, because hypertension is a clear sign of cardiovascular problems and antihypertensive medication indicates that the person has hypertension. Moreover, plaque obstructs the normal blood flow in the arteries, which means that the arteries do not function normally compared to the arteries without plaque and thus the

blood flow into the brains has decreased. The distributions of test subjects at high and low risk for cardiovascular diseases according to different labeling methods are shown in Table 7.

Table 7: Labeling methods and number of persons in them.

	Labeling method, LM1	**Labeling method, LM2**	**Labeling method, LM3**
Number of test subject at high risk for cardio-vascular diseases	1119	165	1357
Number of test subject at low risk for cardio-vascular diseases	619	1573	381

4.5 Feature matrices

Supervised machine learning methods were implemented to classify test subjects to those at high risk for cardiovascular diseases and those at low risk, thus this thesis utilizes 18 different feature matrices but only one label vector per labeling method. The 18 different feature matrices were used, because one of the research questions is, whether it is the clinical data, pulse wave parameters or combination of these two that provides the best results. These 18 feature matrices can be divided into three subgroups: 1) feature matrix contains a combination of the clinical data and the pulse parameters (PWC), 2) feature matrix contains only the clinical data (CLIN), and 3) feature matrix contains the pulse parameters (PWP). However, the sizes of these subgroups vary.

The sizes of the feature matrices are different depending on the label vector, which contains information on who is at high risk for cardiovascular diseases and who is at low risk. The features that were utilized in the construction of the label vectors were removed from the feature matrices to ensure the independence of input variables and labels of the classifiers. In the first labeling method (LM1) the parameters of BMI, height, weight, fasting glucose, systolic blood pressure, MAP, HDL and LDL cholesterol, triglycerides and fasting insulin were removed from the feature matrices. The feature matrices of the second labeling method (LM2) did not include information about BMI, height, weight, HDL and LDL cholesterols, triglycerides, total cholesterol and fasting glucose. For the third labeling method (LM3), the plaque, antihypertensive medication and hypertension parameters were left out. Height and weight were removed from the first and second labeling methods LM1 and LM2, because BMI is calculated with them and therefore, they are closely correlated with labels. All the sizes of resulting feature matrices are presented in Table 8.

Table 8: Sizes of feature matrices in all labeling methods.

Labeling method	Clinical data + pulse parameters (PWC)	Clinical data (CLIN)	Pulse parameters (PWP)
LM1	1738×71	1738×15	1738×56
LM2	1738×74	1738×18	$1738 \times 56.$
LM3	1738×79	1738×23	1738×56

The values of continuous variables in feature matrix were z-score normalized in order to ensure that all features are equally important. Z-score normalization scales the mean to be 0 and the variance 1. All the normalization coefficients are presented in Appendix A. In order to see if the classification results differ in different groups, three different comparisons were made: 1) all data versus the data where the results of the blood tests are removed, 2) male versus female, and 3) old people versus young people.

All data versus the data where the results of the blood tests are removed

It is known that abnormal lipid levels of blood and diabetes are risk factors for cardiovascular diseases [26]. To find out if a cardiovascular risk factor estimation could be done with a non-invasive examination and still get the similar results, the results of the tests requiring blood sampling were removed from the feature matrices. These removed features are fasting insulin, fasting glucose, triglycerides, LDL and HDL cholesterols and total cholesterol. The complete data containing feature matrices for those, which are mentioned in Table 8 and the sizes of feature matrices for those, which do not contain blood tests are presented in Table 9. The pulse wave parameters in Table 8 and Table 9 are identical.

Table 9: Sizes of feature matrices where blood tests are removed.

Labeling method	Clinical data + pulse parameters (PWC)	Clinical data (CLIN)	Pulse parameters (PWP)
LM1	1738×70	1738×14	1738×56
LM2	1738×73	1738×17	1738×56
LM3	1738×73	1738×17	1738×56

Male versus female

Men have a higher risk for cardiovascular diseases than pre-menopausal women [26]. Therefore, the results for men and women were analyzed separately. It is also interesting to investigate if both men and women are classified by the same parameters to those who are at high risk for cardiovascular diseases. The only parameter, which was removed from the feature matrices, was gender, so that it will not be a classifying parameter. The size of feature matrices for men and for women are presented in Table 10.

Table 10: Sizes of feature matrices for men and women.

	Men			Women		
Labeling method	**Clinical data + pulse parameters (PWC)**	**Clinical data (CLIN)**	**Pulse wave parameters (PWP)**	**Clinical data + pulse parameters (PWC)**	**Clinical data (CLIN)**	**Pulse wave parameters (PWP)**
LM1	776×70	776×14	776×56	962×70	962×14	962×56
LM2	776×73	776×17	776×56	962×73	962×17	962×56
LM3	776×78	776×22	776×56	962×78	962×22	962×56

30–36 years old versus 39–45 years old

Ageing increases the risk for having a cardiovascular disease [26]. Therefore, it is intriguing to see if the classifying parameters change, when the age group is different. The number of columns of the different feature matrices for younger age group and for the older age group are exactly the same as is for men and women (see Table 7), but the numbers of rows are 838 and 900 for younger and older test subjects, respectively.

4.6 Linear regression analysis

Regression analysis is a statistical method for modelling the relationship between response variable and one or more regressor variables. The case of one regressor variable is called simple linear regression analysis. For more regressor variables, the method is called multiple linear regression analysis. Multiple linear regression analysis is often used in empirical models. Simple linear regression model is defined as

$$y = \beta_0 + \beta_1 d + \varepsilon, \tag{4.9}$$

where y is the response, d is the regressor variable, β_0 and β_1 are regression coefficients and ε is random error component. [76, p. 12]

Multiple linear regression analysis is defined as

$$y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \cdots + \beta_k d_k + \varepsilon, \tag{4.10}$$

where d_i are the regressor variables, $i = 1, 2, \ldots, k$, and β_j are the corresponding regression coefficients, $j = 0, 1, 2, \ldots, k$. Eq. (4.10) can be expressed in matrix notation as

$$\boldsymbol{y} = \boldsymbol{D\beta} + \boldsymbol{\varepsilon}, \tag{4.11}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{D} = \begin{matrix} 1 & d_{11} & d_{12} & \cdots & d_{1k} \\ 1 & d_{21} & d_{22} & \cdots & d_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & d_{n1} & d_{n2} & \cdots & d_{nk} \end{matrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

In the above, $\boldsymbol{y}$ is observations, $\boldsymbol{D}$ is regressors for different observations, $\boldsymbol{\beta}$ is regression coefficients and $\boldsymbol{\varepsilon}$ is random errors. The multiple linear regression analysis describes the hyperplane in the *k*-dimensional space of the regressor variables d_i. The β_j describes the expected change in the response y per unit change in d_i when all the remaining regressor variables are held constant. However, the shape of the hyperplane might not be linear but the β coefficients are linear. To find optimal parameters β, least-squares fit is applied. The least-squares estimator of $\boldsymbol{\beta}'$ is defined as

$$\boldsymbol{\beta}' = (\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{y}, \tag{4.12}$$

where and $\boldsymbol{\beta}'$ is least-square estimator for different observations. Thus, the linear regression can be presented as

$$\boldsymbol{y}' = \boldsymbol{D\beta}'. \tag{4.13}$$

Usually, the *p*-values based on the Student's *t*-test and standard errors for estimated coefficients are calculated from the linear regression model. [76, pp. 67–73]

4.7 Classifying in MATLAB

MATLAB (MathWorks, Inc.) was used in this thesis to find relevant parameters and for the classification. Each of the 18 feature matrices were used to find the relevant parameters for the classification. Before the classification, the forward selection was applied to

find the most relevant parameters. Forward selection was implemented by using `sequentialfs`-function.

`Sequentialfs`-function takes in the feature matrix, labels and a function handle of the classifier. This classifier takes in a part of the dataset and their labels for training and rest of the dataset and their labels for testing. This partitioning of the dataset is done with 10-fold cross-validation, which is one of the inputs for the `sequentialfs`-function, which predicts the labels. These predicted labels are compared to the testing labels. The `sequentialfs`-function needs a criterion that describes if the classification improves or not at each iteration. In this thesis, the criterion is the number of wrongly predicted class labels and therefore, the number of wrongly predicted class labels should decrease with every iteration.

The results are analyzed with AUC and ROC. However, to calculate AUC and ROC, the classification of all feature matrices has to be done again with the parameters that were chosen in forward selection. Therefore, classifying is implemented with the feature matrix from which the relevant parameters are chosen. Also in this phase, 10-fold cross validation is applied during the fitting. A `kfoldPredict`-function is used to predict the class labels. The output of `kfoldPredict`-function is `score`.

5. RESULTS AND DISCUSSION

The results obtained in this thesis are presented in two parts. Firstly, the association of IMT and FMD to different parameters is presented, for enabling a comparison between results shown here and Table 5 in Section 4.3. Secondly, the ROC curves and AUC values are shown. A description of classifying parameters is presented in Appendix C.

5.1 Association of the calculated pulse wave parameters with FMD and IMT

56 parameters in total were derived from the IPG and ICG signals, but only the five pulse wave derived parameters that have the strongest association with IMT are presented in Table 11 and the five parameters that have the strongest association with FMD are presented in Table 12. If same parameter appears also in sex- and age-adjusted part and multivariable-adjusted part, the parameter performs well. The potential of the parameter is higher if it shows significant association with the output variable also in multivariable-adjusted mode. The test subjects are divided into two age groups similarly as in [7], so that the results are comparable with the results that Koivistoinen *et al*. obtained.

Table 11: Five pulse wave parameters that have the strongest association with IMT and their p-values. Number is reference to number of the parameter in Appendix B.

	IMT	
	Age 30–36 years (*n* = 868)	**Age 39–45 years (*n* = 938)**
Parameter	*p*-value	*p*-value
Sex and age-adjusted		
Gl4 of ICG with time differences between 1st and 2nd highest peaks (67)	0.00798	0.0668
FFT of transformed IPG (29)	0.00803	0.198
R_3 of ICG (55)	0.00863	0.00881
Unnormalized transmission time of ICG (38)	0.0119	0.243
R_4 of ICG (48)	0.0829	0.0391
Multivariable-adjusted†		
FFT of IPG (28)	0.00195	0.718
FFT of transformed IPG (29)	0.00258	0.726
Rise time of IPG (33)	0.00259	0.225
AGI of ICG (63)	0.0666	0.286
Gl4 of transformed IPG with time differences between 1st and 2nd highest peaks (80)	0.0850	0.902

† contains: age, sex, systolic blood pressure, HDL, LDL, BMI, triglycerides (log), glucose (log), insulin (log)

Table 12: Five pulse wave parameters that have strongest the association with FMD and their p-values. Number is reference to number of the parameter in Appendix B.

	FMD	
	Age 30–36 years (*n* = 868)	**Age 39–45 years (*n* = 938)**
Parameter	*p*-value	*p*-value
Sex and age-adjusted		
Decay time of ICG (41)	2.82e-06	0.00102
L5 of IPG with time differences between 1st and 3rd highest peaks (75)	2.26e-05	0.0130
T_1 of ICG (57)	0.00218	0.00499
Gln4 of ICG with time differences between 1st and 2nd highest peaks (67)	0.00716	0.0162
T_2 of ICG (59)	0.00774	0.00387
Multivariable-adjusted†		
Decay time of ICG (41)	7.38e-06	0.00197
L5 of IPG with time differences between 1st and 3rd highest peaks (75)	0.000136	0.0248
Gln4 of ICG with time differences between 1st and 2nd highest peaks (67)	0.007235	0.0289
T_1 of ICG (57)	0.0100	0.0120
T_2 of ICG (59)	0.0306	0.0120

† contains: age, sex, systolic blood pressure, HDL, LDL, BMI, triglycerides (log), glucose (log), insulin (log)

In the analyses presented in Table 5, Table 11 and Table 12 the triglycerides, glucose and insulin and C-reactive protein (only in Table 5) are used as common logarithms (10-base logarithm), because the distributions of these are skewed. As can be seen in Table 11, the

FFT of transformed IPG performs well in both sex- and age-adjusted mode and multivariable-adjusted model when the age group is 30–36 years old. However, the age group 39–45 does not show any statistically significant association with any parameter in the multivariable adjusted model. The associations with FMD are generally stronger than with IMT. Moreover, the parameters reoccur in the sex- and age-adjusted model and multivariable-adjusted model. Because the parameters seem to have shown by smaller *p*-value than presented in [7] with IMT and FMD than PWV, it is possible that these parameters provide information about the condition of the arteries compared with PWV, which currently used in clinical practice.

When these parameters are compared to the classifying parameters in Appendix C, it is noticeable that all these parameters present themselves also as classifying parameters. In addition to having stronger connection with IMT and FMD, these parameters might give information about the condition of the arteries. The parameters, in Table 11 and Table 12, which are found most often as classifying parameters are *R_3 of ICG* and *FFT of transformed IPG*. The two parameters, in Table 11 and Table 12, which are most seldom as classifying parameters are *Rise time of IPG* and *L5 of IPG with time differences between 1st and 3rd highest peaks* (see Table 19–Table 24).

5.2 Analysis of AUC and ROC curve

The ROC curves are presented only for the three labeling methods (LM1, LM2 and LM3) that contain all the parameters excluding those, which were used in labeling. Figure 33–Figure 41 present the ROC curves of Table 13. Table 13–Table 18 present the AUC values of experiments with all data, the results of blood test were removed, old age group, young age group, as well as men and women, respectively. The best AUC value is written in green and the worst is written in red in each classifying method. If a hyphen (–) is presented in some table, the classifier did not find any significant parameters. Figure 33 presents the ROC curves of different classifiers for the labeling method LM1 and feature matrix that contains pulse wave parameters and clinical data.

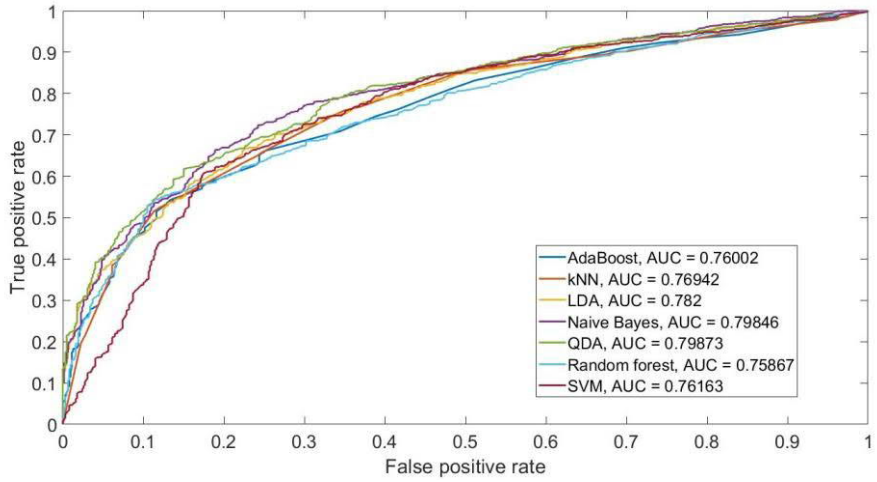

Figure 33: ROC-curves for labeling method LM1 and a feature matrix containing both clinical data and pulse wave parameters (PWC).

It can be seen in Figure 33 that QDA performs the best and Random Forest performs the worst but the difference in their AUC-values is only approximately 0.04. Different features were selected for different classifier by the forward selection. Resulting from the implemented forward selections, the Random Forest used only clinical data whereas QDA has both clinical and calculated pulse wave parameters (see Table 19 in Appendix C). AdaBoost employs also only clinical data and it performs second worst. To see how the ROC curves and AUC values change if the feature matrix contains only the clinical data, the ROC curves for the labeling method LM1 with the feature matrix that contains only the clinical data for each classifier are presented in Figure 34.

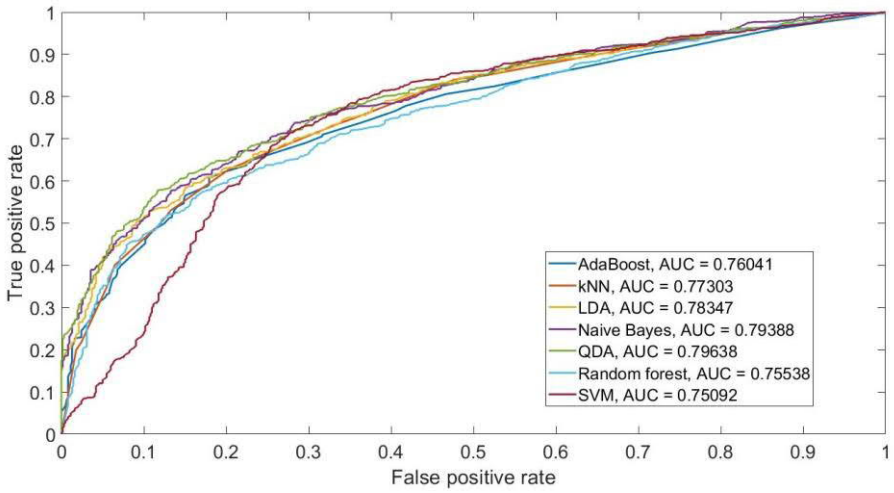

Figure 34: ROC-curves for labeling method LM1 and the feature matrix containing only clinical data (CLIN).

The AUC values do not change significantly even though only the clinical parameters are utilized as can be seen when Figure 34 is compared with Figure 33 (LM1+PWC). As can be seen in Figure 33 QDA performs better than in Figure 34. The difference between the best and the worst AUC value is approximately 0.05. To compare how the ROC curves

and the AUC values of the pulse wave parameters with those in Figure 33 and Figure 34, the ROC curves for labeling method LM1 with the feature matrix that contains only pulse wave parameters is presented in Figure 35.

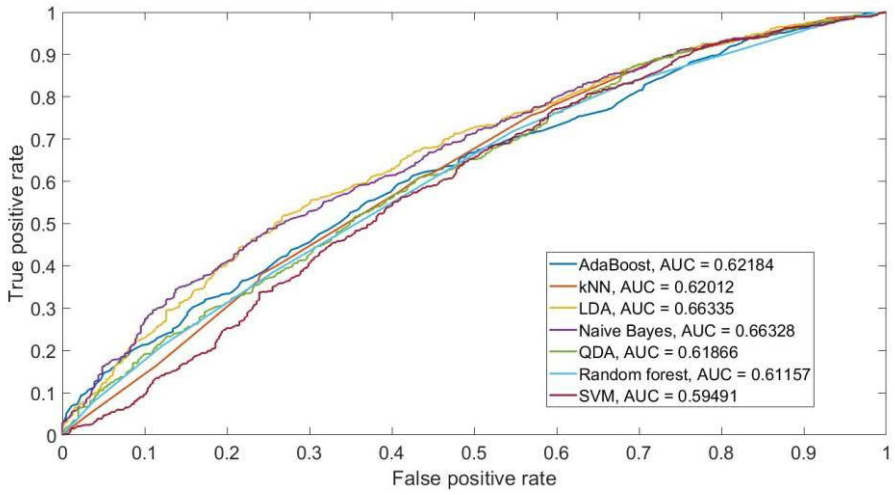

Figure 35: ROC-curves for labeling method LM1 and a feature matrix containing only pulse wave parameters (PWP).

The LDA and Naïve Bayes provide the best AUC values as can be seen in Figure 35 (LM1+PWP). In Figure 35, the classification is not as good as in Figure 33 (LM1+PWC) and Figure 34 (LM1+CLIN): the maximum difference is approximately 0.15 between AUC values of clinical data and AUC values of calculated pulse wave parameters.

Unhealthy lifestyle changes the clinical parameters and unhealthy lifestyle is connected with cardiovascular diseases. Therefore, the clinical parameters that were used in labeling are associated with the rest of the clinical data thus improving the AUC values of the clinical data. For instance, high blood pressure correlates with hypertension. Given that none of the pulse wave parameters are utilized in labeling, they do not necessarily correlate with the parameters of the clinical data that were used in labeling.

As it was stated earlier, different labeling methods were tested. Hence, Figure 36–Figure 38 present labeling method LM2 with different feature matrices. Figure 36 present the ROC curves for each classifier of the labeling method LM2 with the feature matrix that contains the clinical data and the pulse wave parameters.

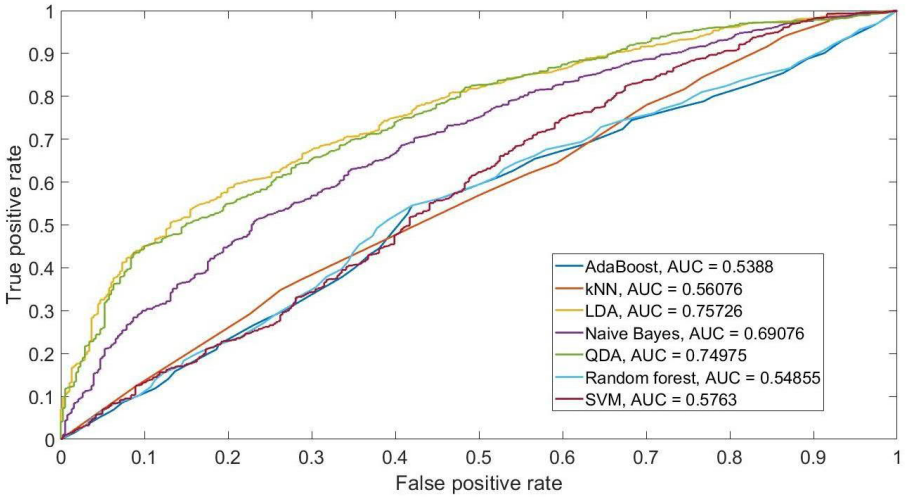

Figure 36: ROC-curves for labeling method LM2 and a feature matrix containing both clinical data and pulse wave parameters (PWC).

Labeling method LM2 results in larger variation between different classifying methods than the labeling method LM1+PWC (see Figure 33 and Figure 36). The best AUC value is provided by the LDA. AdaBoost and Random Forest provide the worst AUC values as can be seen in Figure 36. In case of the AdaBoost and Random Forest, the feature selection found only one parameter and thus, the classification is made based on only one feature. Other classifiers found more than one parameter (see Table 19), for instance LDA found six relevant parameters. Therefore, it is possible that the forward selection does not function correctly in every case. Hence, the results might not be completely reliable. In general, the more parameters forward selection found, the higher the AUC values are. In Figure 37, the ROC curves of each classifier utilizing only the clinical data for the labeling method LM2 are presented so that the comparison can be made with Figure 36.

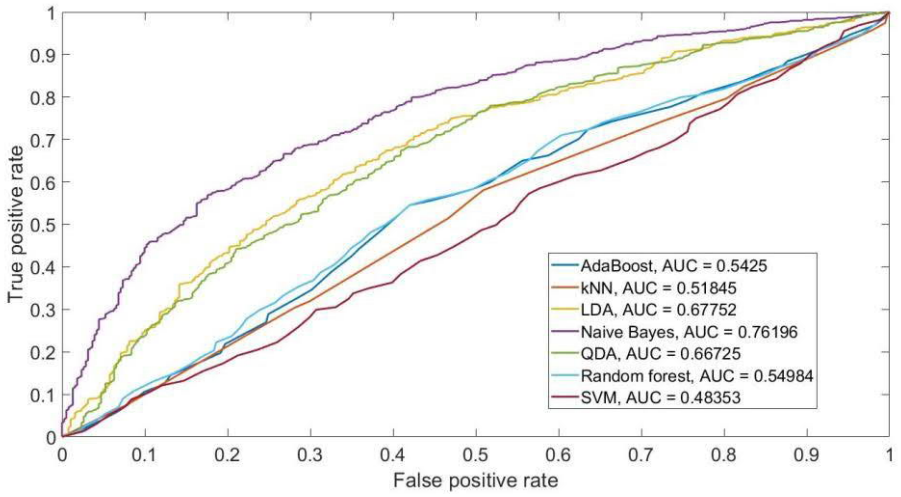

Figure 37: ROC-curves for labeling method LM2 and a feature matrix containing only clinical data (CLIN).

It is seen in Figure 37 that the best classifying result is provided by the naïve Bayes. The SVM provided the worst result and it found only one relevant parameter (see Table 19). The naïve Bayes found five relevant parameters and it performed the best. However, the LDA found only one parameter (MAP during CircMon-measurement) and its AUC is second best. Figure 38 shows the ROC curves for the labeling method LM2 with the feature matrix that utilizes only the pulse wave parameters so that comparison can be made with Figure 35.

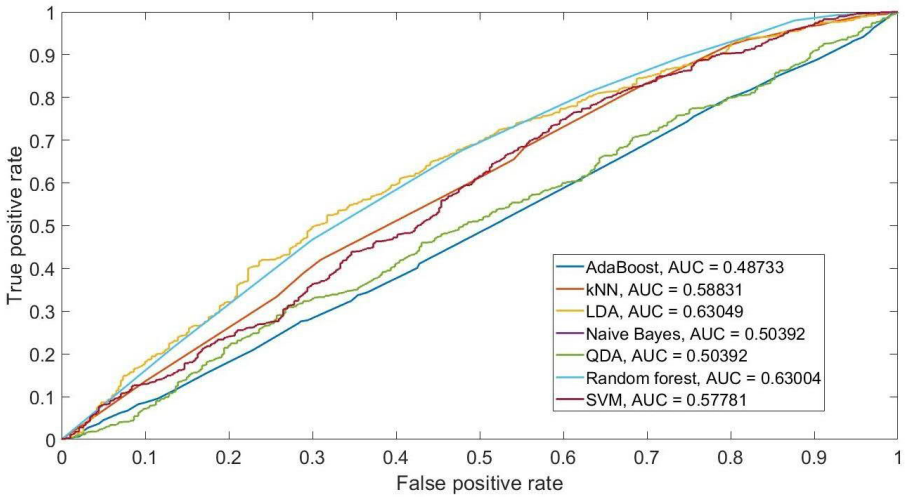

Figure 38: ROC-curves for labeling method LM2 and a feature matrix containing only pulse wave parameters (PWP).

In Figure 38, the AUC values are worse than in Figure 35 (LM1+PWP). The worst AUC values are with AdaBoost, Naïve Bayes and QDA might be caused by the fact that each of these classifiers have found only one relevant parameter. Further, Random Forest performed best, and it found six relevant parameters. LDA, SVM and k-NN found three relevant parameters and LDA has the best AUC value from these three classifiers.

Lastly, the ROC curves for the third labeling method LM3 with different feature matrices are presented in Figure 39–Figure 41. Figure 39 presents the ROC curves of the labeling method LM3. The feature matrix consists of both clinical data and pulse wave parameters.

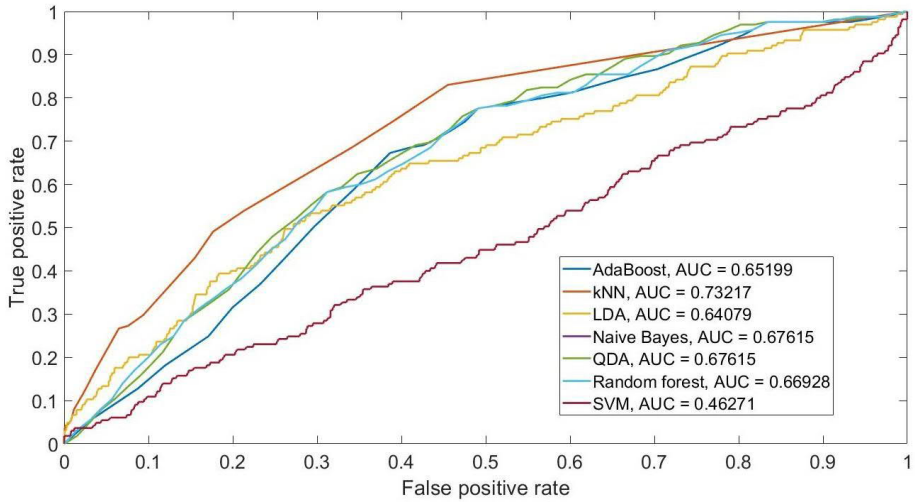

Figure 39: ROC-curves for labeling method LM3 and a feature matrix containing both clinical data and pulse wave parameters (PWC).

It is seen in Figure 39 that the k-NN performs best and it utilizes both clinical and calculated pulse wave parameters to classify test subjects. It is also seen that the SVM performs worst, which could be a result of overlearning. However, AdaBoost, Naïve Bayes, QDA and Random Forest have found only the age parameter to be relevant (see Table 19) but it classifies the test subject well. It is known that age is a risk factor for cardiovascular diseases and LM3 method seems to prove it. To compare the ROC curves with ROC curves in Figure 34 (LM1+CLIN) and Figure 37 (LM2+CLIN), the ROC curves of labeling method LM3 with the feature matrix, which contains only the clinical data, are presented in Figure 40.

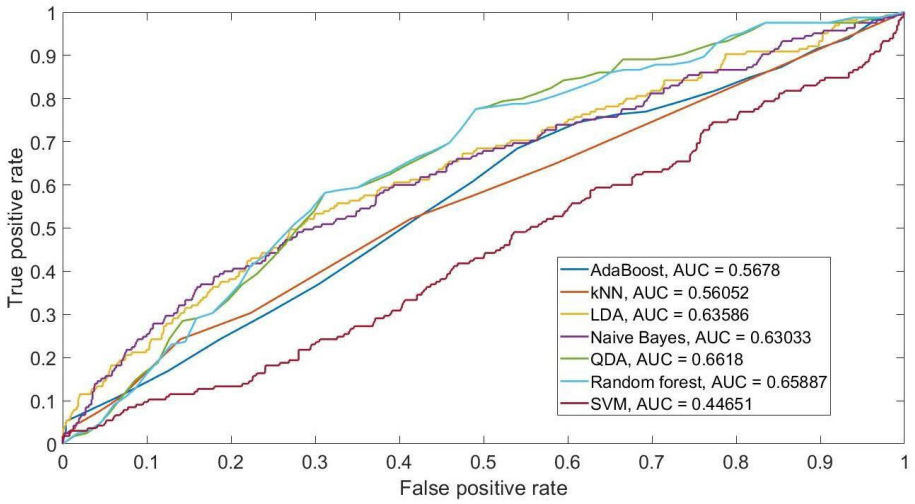

Figure 40: ROC-curves for labeling method LM3 and a feature matrix containing only clinical data (CLIN).

The AUC values in Figure 40 are not as good as in Figure 34 (LM1+CLIN) and Figure 37 (LM2+CLIN). Especially the SVM seems not to be able to generalize. However, the

AUC values have the same order of magnitude as in Figure 39 (LM3+PWC), which indicates that the AUC values in Figure 39 consist mostly of the clinical data (see Table 19). Moreover, the number of the test subjects who have been labeled as being high risk for cardiovascular diseases is smaller than the number of test subjects who have been labeled as being low risk for cardiovascular diseases. Therefore, the classifying result might be decreased. To see how the ROC curves change when the feature matrix utilizes only pulse wave parameters, the ROC curves of the labeling method LM3 presented in Figure 41.

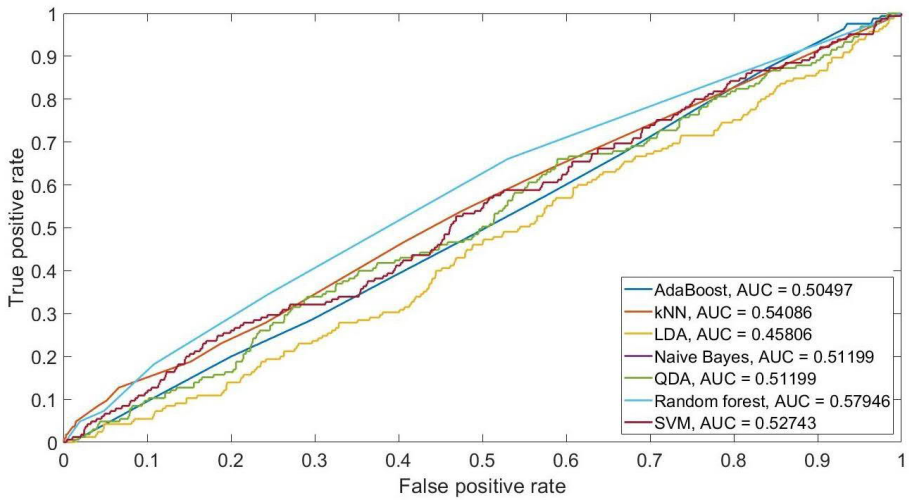

Figure 41: ROC-curves for labeling method LM3 and a feature matrix containing only pulse wave parameters (PWP).

In Figure 41, the AUC values have the same order of magnitude as in Figure 38 (LM2+PWP). The LDA seems to have overlearned the model and therefore, performs badly. k-NN and Random Forest are the only classifiers that did not find the FFT of pulse wave signals (see Table 19) as a relevant parameter and they have the best AUC values. Random Forest found four parameters to be relevant and it has the highest AUC value. Other classifiers found only one relevant parameter, thus their AUC values might not be reliable.

All data versus the data where the results of the blood tests are removed

Table 13 presents all the AUC values which are presented in Figure 33–Figure 41. This part compares all the data, excluding those that were used in labeling with the data where the results of the blood tests were removed.

Table 13: AUC values for the whole dataset.

Labeling method		Ada-Boost	K-nearest neighbor	LDA	Naive Bayes	QDA	Random forest	SVM
LM1	Clinical + pulse parameters	0.760	0.769	0.782	0.799	0.799	0.759	0.762
	Clinical	0.760	0.773	0.784	0.794	0.796	0.755	0.751
	Pulse parameters	0.622	0.620	0.663	0.663	0.619	0.612	0.595
LM2	Clinical + pulse parameters	0.539	0.561	0.757	0.691	0.750	0.549	0.576
	Clinical	0.543	0.519	0.678	0.762	0.667	0.550	0.484
	Pulse parameters	0.487	0.588	0.631	0.504	0.504	0.630	0.578
LM3	Clinical + pulse parameters	0.652	0.732	0.641	0.676	0.676	0.669	0.463
	Clinical	0.568	0.561	0.636	0.630	0.662	0.659	0.447
	Pulse parameters	0.505	0.541	0.458	0.512	0.512	0.580	0.527

As can be seen in Table 13, the SVM performs worse than any other classifier. The AUC values that are below 0.5 might mean that the classifier has overlearned. Therefore, it seems that the calculated pulse wave parameters hold relevant information and improve the classification if the classifier employs the calculated pulse wave parameters in the case of clinical data and pulse wave parameters. To see if the AUC values differ when the blood tests are removed, the AUC values for the feature matrices that do not contain the results of the blood tests are presented in Table 14.

Table 14: AUC values for the dataset where the results of blood tests are removed.

Labeling method		Ada-Boost	K-nearest neighbor	LDA	Naive Bayes	QDA	Random forest	SVM
LM1	Clinical + pulse parameters	0.708	0.686	0.740	0.746	0.749	0.706	0.686
	Clinical	0.714	0.703	0.734	0.719	0.726	0.710	0.692
	Pulse parameters	0.638	0.625	0.687	0.652	0.666	0.585	0.628
LM2	Clinical + pulse parameters	0.548	0.560	0.680	0.673	0.566	0.550	0.602
	Clinical	0.545	0.519	0.667	0.564	0.564	0.552	0.494
	Pulse parameters	0.478	0.539	0.625	0.519	0.621	0.608	0.550
LM3	Clinical + pulse parameters	0.467	0.730	0.671	0.667	0.667	0.663	0.480
	Clinical	0.640	0.684	0.615	0.664	0.664	0.662	0.473
	Pulse parameters	0.509	0.519	0.439	0.499	0.499	0.528	0.513

The results obtained with calculated pulse wave parameters are quite close to the ones obtained with the clinical data in the labeling methods LM1 and LM2. Therefore, it can be concluded that also the pulse wave parameters describe the condition of the arteries. Moreover, the best AUC values in each labeling method are produced when, both clinical data and calculated pulse wave parameters are used as predictors.

When Table 13 and Table 14 are compared with each other, removing the results of the blood tests from the feature matrix worsened especially the classification result of the clinical data. However, the combination of the pulse wave parameters and the clinical data have clearly better classification results than the clinical data itself. Therefore, the pulse wave parameters contain some independent information about the condition of the arteries, which might indicate that some information of the blood tests could be replaceable with pulse wave parameters. Furthermore, the number of classifying parameters in each labeling method and in each classifier does not indicate better or worse classification result (see Table 19 and Table 20 in Appendix C).

Young versus old test subjects

In this part, the old test subjects (39–45 years old) are compared to the young test subjects (30–36 years). Table 15 shows the AUC values for the feature matrix that contained only people who are 39–45 years old.

Table 15: AUC values of the old age group (39–45 years old).

Labeling method		Ada-Boost	K-nearest neighbor	LDA	Naive Bayes	QDA	Random forest	SVM
LM1	Clinical + pulse parameters	0.775	0.789	0.782	0.820	0.802	0.572	0.756
	Clinical	0.759	0.766	0.795	0.804	0.794	0.748	0.760
	Pulse parameters	0.664	0.661	0.666	0.666	0.690	0.646	0.647
LM2	Clinical + pulse parameters	0.514	0.717	0.772	0.773	0.675	0.685	0.543
	Clinical	0.599	0.714	0.771	0.774	0.767	0.672	0.649
	Pulse parameters	0.450	0.509	0.587	0.640	0.518	0.547	0.550
LM3	Clinical + pulse parameters	0.495	0.488	0.582	0.484	0.627	0.519	0.498
	Clinical	0.487	0.565	0.644	0.630	0.630	0.504	0.500
	Pulse parameters	0.476	0.431	0.478	0.494	0.494	0.488	0.469

In Table 15, the calculated pulse wave parameters might have some additional information when the LM1 labeling method is examined, except in the results of the LDA, random forest and SVM where the performance is decreased when also the pulse wave parameters are considered. The random forest in clinical parameters and pulse wave parameters of LM1 found only one relevant parameter (see Table 21 in Appendix C), which would explain the worst AUC value in LM1. The calculated pulse wave parameters seem to have a worsening effect in LM2 labeling method, because the AUC values are greater when the feature matrix consists only the clinical parameters compared with the case where feature matrix consists of both clinical and pulse wave parameters. The clinical data might give more information about the condition of the arteries in older people, which affects the classification result. As it was stated earlier the old test subjects are compared to the young test subjects, and thus the AUC values for the young test subjects are presented in Table 16.

Table 16: AUC values of the young age group (30–36 years old).

Labeling method		Ada-Boost	K-nearest neighbor	LDA	Naive Bayes	QDA	Random forest	SVM
LM1	Clinical + pulse parameters	0.766	0.791	0.796	0.803	0.800	0.459	0.758
	Clinical	0.664	0.778	0.788	0.801	0.795	0.758	0.771
	Pulse parameters	0.555	0.592	0.659	0.629	0.673	0.530	0.631
LM2	Clinical + pulse parameters	0.493	0.667	0.731	–	0.549	0.552	0.680
	Clinical	0.518	0.551	0.728	–	0.731	0.561	0.492
	Pulse parameters	0.491	0.467	0.550	0.626	0.618	0.627	0.583
LM3	Clinical + pulse parameters	0.429	0.500	0.550	0.477	0.477	0.478	0.472
	Clinical	0.442	0.500	0.566	0.442	0.442	0.443	0.518
	Pulse parameters	0.504	0.493	0.513	0.486	0.486	0.580	0.531

The AUC values, in Table 16, are almost identical with the AUC values in Table 15. All the best AUC values in each labeling method employ calculated pulse wave parameters. Moreover, the AUC values of calculated pulse wave parameters are close to those of the clinical data. Therefore, the calculated pulse wave parameters hold information about the arteries even when the tests subjects are 36 years old or younger. It is noteworthy, that the best AUC value in LM3 labeling method is in the calculated pulse wave parameters.

When Table 15 and Table 16 are compared, the difference between them is small. Moreover, the differences between Table 21 and Table 22 in Appendix C are equally small. Thus, the test age gap between the old and young group is not so great that it would show significant differences in the classifying parameters or in the AUC values. In addition, the changes in the arteries might be small and therefore, they are not visible in the classifying result of the test subjects who are 39–45 years old.

Male versus female test subjects

This part compares the AUC values between men and women. The AUC values of the men are presented first in Table 17.

Table 17: AUC values of men.

Labeling method		Ada-Boost	K-nearest neighbor	LDA	Naive Bayes	QDA	Random forest	SVM
LM1	Clinical + pulse parameters	0.784	0.760	0.791	0.803	0.814	0.513	0.667
	Clinical	0.679	0.627	0.779	0.790	0.787	0.524	0.730
	Pulse parameters	0.720	0.620	0.723	0.631	0.670	0.641	0.692
LM2	Clinical + pulse parameters	0.522	0.557	0.683	–	0.547	0.615	0.519
	Clinical	0.665	0.586	0.680	–	0.682	0.608	0.555
	Pulse parameters	0.530	0.518	0.614	0.541	0.606	0.643	0.530
LM3	Clinical + pulse parameters	0.652	0.596	0.674	0.674	0.674	0.667	0.520
	Clinical	0.670	0.694	0.546	0.700	0.678	0.675	0.527
	Pulse parameters	0.537	0.440	0.502	0.525	0.525	0.504	0.497

The AUC values of the men in Table 17 does not differ from other the groups presented earlier. The best AUC values in LM1 and LM2 labeling method both employ the calculated pulse wave parameters (see Table 23 in Appendix C). However, the labeling method LM3 did not employ any pulse wave parameters when the clinical data and pulse wave parameters are combined. To see if the AUC values of women are different from men, the AUC values of women are presented in Table 18.

Table 18: AUC values of women.

Labeling method		Ada-Boost	K-nearest neighbor	LDA	Naive Bayes	QDA	Random forest	SVM
LM1	Clinical + pulse parameters	0.779	0.768	0.789	0.809	0.820	0.755	0.769
	Clinical	0.768	0.775	0.783	0.801	0.796	0.758	0.770
	Pulse parameters	0.595	0.644	0.646	0.658	0.658	0.628	0.541
LM2	Clinical + pulse parameters	0.565	0.680	0.620	–	0.586	0.515	0.572
	Clinical	0.684	0.715	0.727	–	0.708	0.533	0.681
	Pulse parameters	0.617	0.567	0.613	0.602	0.589	0.603	0.569
LM3	Clinical + pulse parameters	0.558	0.474	0.635	0.665	0.665	0.561	0.459
	Clinical	0.645	0.481	0.753	0.655	0.692	0.646	0.426
	Pulse parameters	0.485	0.498	0.454	0.461	0.461	0.589	0.505

The best AUC value in labeling method LM1 uses both clinical data and calculated pulse wave parameters (see Table 24 in Appendix C). The labeling method LM3 did not utilize the pulse wave parameters when the calculated pulse wave parameters and clinical data were combined. The worst AUC value in labeling method LM2 did not utilize any pulse wave parameters whereas other methods in the combination of the clinical data and pulse wave parameters did. The AUC values in Table 17 and Table 18 are similar with each other. Hence, there is not differences between sexes even though the men have greater risk of heart diseases. In both cases, if the best AUC value was in the clinical data and pulse parameters section, the classifier used both clinical data and calculated pulse wave parameters.

Three different labeling methods were tested because the test subjects were young and therefore relatively coherent group from the cardiovascular status point of view. Thus, it is difficult to make a definite distinction between high and low risk for cardiovascular diseases. Moreover, the pulse wave parameters do not have any reference values to whom they could be compared to. The labeling method LM1 is based on this dataset and therefore, it might be biased. Hence, the AUC values are high. Nevertheless, the LM1 is most

evenly distributed: 60-% is being labeled as being high risk for cardiovascular diseases. The labeling method LM2 is based on the national reference values and thus, is independent on the dataset. However, the labeling method LM2 labels 90-% of the test subjects as being at high risk for cardiovascular diseases. Therefore, the classification result might be worsened. The labeling method LM3 has the lowest AUC values and the test subjects who are labeled as being at high risk for cardiovascular diseases consists of only 22-% of the whole dataset, thus the labeling is not evenly distributed.

The labeling methods LM2 and LM3 might cause the 10-fold cross-validation to have some testing groups that contain only few low risk or high risk test subjects. Therefore, the classification results between each classifier might vary and the number of relevant parameters might be decreased. As stated earlier, it is difficult to distinguish test subjects who are at high risk for cardiovascular from those who are low risk. Thus, the classes are not separable, which decreases the AUC values and causes different results between classifiers. Moreover, the classifiers could provide different weights for the outlier values, which could explain the different AUC values of different classifiers.

The antihypertensive medication might cause the poorer classifying result in labeling method LM3. If the person has an antihypertensive medication, the blood pressure values and the lifestyle of that test subject might be better than for someone who does not suffer from hypertension. Therefore, a test subject who is labeled as being at high risk for cardiovascular diseases might have better values in feature matrix than a test subject marked at low risk for cardiovascular diseases.

When the pulse wave parameters were calculated for each test subject, the median of the time series might have been more robust, because the outlier values would have not biased it as much as in the case of the average of the time series. Now, there may be higher or lower values when using mean than there would have been with median. In addition, some parameters might still have contained some outlier values that affect the classifying and therefore, worsen it. However, a manual removing all outliers from the data is impossible in practice. The removal of the outliers has based on the statistical characteristics i.e. histogram where the clearly outlying values were removed. Thus, the dataset might still contain them.

The test subjects are still young, and the atherosclerosis takes decades to develop. Therefore, the majority of the processes related with the disease remain subclinical without any indication in the clinical data. Hence, the classifying is not as accurate as it would be with old people who have, for instance, atherosclerosis, which has made damage on the arteries and objectively observed clinical symptoms, findings, and changes in FMT and CIMT. Furthermore, those who have a higher risk for cardiovascular diseases might be exposed to intensive healthcare interventions and therefore, monitored more accurately, be in-

formed about lifestyle change, and be administrated pharmacological treatment if the lifestyle change fails to improve the risk profile. Therefore, they might be even healthier than those who have been labelled as low risk for cardiovascular diseases.

Most of the pulse wave analysis methods and parameters are originally developed for other types of sensors, for instance PPG. Therefore, these parameters developed for other types of sensors might not work as well for the impedance-based measurements. Thus, the parameters might not have been calculated correctly in every case, which causes the outliers to the data. Hence, these analysis algorithms for the parameters should be optimized for the impedance-based measurements.

It is also possible that the reduced AUC values are affected by incorrect operation or unoptimally set inputs of the forward selection because sometimes the forward selection found only one parameter to be relevant. Most of the times this one relevant parameter was the first feature of the used feature matrix. Thus, the implementation of the forward selection might contain errors. Especially in the cases where only one or two classifiers found only one relevant parameter and others found two or more relevant parameters. The optimization of the classifiers might reduce this problem and thus improve the classification results and the changing the optimization criteria for the forward selection.

6. CONCLUSIONS

In this thesis, linear regression analysis was utilized to investigate the association between pulse wave parameters and IMT and FMD. In addition, the performance of pulse wave parameters calculated from the IPG and ICG signals was investigated by feeding them to the different classifiers used to classify test subjects who have either high or low risk for cardiovascular diseases. Forward selection was used to select the most relevant features and 10-fold cross-validation was implemented to investigate the generalization of the results.

The linear regression analysis showed that the pulse wave parameters have stronger association with IMT and FMD than the PWV had. The association between pulse wave parameters and FMD is strong in regardless of the age of the test subject. However, the association with pulse wave parameters and IMT is weaker than with FMD but stronger than the association between IMT and PWV. Therefore, there is possibility that the pulse wave parameters provide information about the condition of the arteries compared with the PWV, which is used in clinical practice.

The calculated pulse wave parameters provide some independent information about the condition of the arteries because the best AUC values were found with a combination of clinical data and pulse wave parameters in each labeling method in most of the cases. However, sometimes the forward selection found only clinical data to be relevant even though the pulse wave parameters were also presented. Therefore, the clinical data seems to hold more information about the condition of the arteries than the pulse wave parameters. However, it is possible that pulse wave parameters could replace some of the clinical data because the classification without blood tests worsened the classification result of clinical data, but when using of clinical data in combination with pulse wave parameters this did not affect as much. The calculated pulse wave parameters alone do not provide equal information of the condition of the arteries as the clinical data but the use of right inputs of the classifiers might improve the classification result and thus the calculated parameters might provide equally good classification results as the clinical data.

The risk stratification with only calculated pulse wave parameters is not as good as the combination of the clinical data and calculated pulse wave parameters or only the clinical data. However, the calculated parameters may still contain outliers that decrease the classification result. Therefore, if these outliers could be removed from the feature matrices, the classification could improve. Further, the impedance measurement might provide a non-invasive way to classify people and help to find those people who are at risk for cardiovascular diseases.

In future work, the deep neural networks could be utilized to learn some hidden features of the pulse waves. Moreover, the classifiers could be improved by testing different input parameters such as different kernel functions to find the best combination. With longitudinal studies, it would be more reliable to find out how early the pulse wave parameters predict the cardiovascular diseases. Even though the utilized data set is part of longitudinal study, the test subjects were only 45-years old or younger at the time of recordings. Therefore, the examination and the analysis should be repeated with the same test subjects but decades between the measurements so that it would be possible to see if the pulse wave parameters provide information about the condition of the arteries. Moreover, the analysis algorithms should be optimized for the impedance-based measurements, which could improve the classification. In addition, forward selection should be studied more carefully, and its optimization criteria could be changed.

As a conclusion, the calculated pulse wave parameters describe the condition of the arteries at some level. However, the clinical data still holds more information about the condition of the arteries but in the most cases, the AUC values of the clinical data were worse than the AUC values of the combination of the clinical data and pulse wave parameters. Therefore, the pulse wave parameters hold some information about the condition of the arteries that is independent from the clinical data. In order to improve the classification, labeling of the test subjects should be considered so that approximately half of the whole number of test subjects are people at higher or lower risk for cardiovascular diseases. Furthermore, the test subjects who are labeled as being at high risk for cardiovascular diseases should have distinct clinical values from the test subjects who are labeled as being at low risk for cardiovascular diseases.

REFERENCES

[1] Sydän- ja verisuonitautien yleisyys. Available: https://thl.fi/fi/web/kansantaudit/sydan-ja-verisuonitaudit/sydan-ja-verisuonitautien-yleisyys. [Accessed 1.2.2019]

[2] The Cardiovascular Risk in Young Finns Study. Available: http://youngfinnsstudy.utu.fi/index.html.

[3] M. Peltokangas, A. A. Telembeci, J. Verho, V. M. Mattila, P. Romsi, A. Vehkaoja, J. Lekkala and N. Oksala, "Parameters Extracted From Arterial Pulse Waves as Markers of Atherosclerotic Changes: Performance and Repeatability," *IEEE Journal of Biomedical and Health Informatics,* vol. 22, no. 3, pp. 750–757, 2018.

[4] K. Takazawa, N. Tanaka, M. Fujita, O. Masami, T. Saiki, M. Aikawa, S. Tamura and C. Ibukiyama, "Assessment of Vasoactive Agents and Vascular Aging by the Second Derivative of Photoplethysmogram Waveform," *Hypertension,* vol. 32, no. 2, pp. 365–370, 1998.

[5] S. C. Millasseau, R. P. Kelly, J. M. Ritter and P. J. Chowienczyk, "Determination of age-related increases in large artery stiffness by digital pulse contour analysis," *Clinical Science*, vol. 103, no. 4, pp. 371–377, 2002.

[6] J. G. Webster and J. W. Clark, *Medical Instrumentation: Application and Design*, 4.th ed., Hoboken, NJ: John Wiley & Sons, 2010.

[7] T. Koivistoinen, M. Virtanen, N. Hutri-Kähönen, T. Lehtimäki, A. Jula, M. Juonala, L. Moilanen, H. Aatola, J. Hyttinen, J. S. A. Viikari, O. T. Raitakari and M. Kähönen, "Arterial pulse wave velocity in relation to carotid intima-media thickness, brachial flow-mediated dilation and carotid artery distensibility: The Cardiovascular Risk in Young Finns Study and the Health 2000 Survey," *Atherosclerosis,* vol. 220, no. 2, pp. 387–393, 2012.

[8] T. Koivistoinen, L. P. Lyytikäinen, H. Aatola, T. Luukkaala, M. Juonala, J. Viikari, T. Lehtimäki, O. Raitakari, M. Kähönen and N. Hutri-Kähönen, "Pulse Wave Velocity Predicts the Progression of Blood Pressure and Development of Hypertension in Young Adults," *Hypertension,* vol. 71, no. 3, pp. 451–456, 2018.

[9] T. Koivistoinen , T. Kööbi, A. Jula, N. Hutri-Kähönen, O. T. Raitakari, S. Majahalme, K. Kukkonen-Harjula, T. Lehtimäki, A. Reunanen, J. Viikari, V. Turjanmaa and M. Kähönen, "Pulse wave velocity reference values in healthy adults aged 26–75 years," *Clinical Physiology and Functional Imaging,* vol. 27, no. 3, pp. 191–196, 2007.

[10] M. Peltokangas, "Analysis methods for arterial pulse wave signals recorded with body sensor network", Master of Science Thesis, Tampere University of Technology 2013.

[11] M. O'Rourke F., A. Pauca and X. Jiang, "Pulse wave analysis," *British Journal of Clinical Pharmacology.,* vol. 51, no. 6, pp. 507–522, 2001.

[12] E. N. Marieb, *Essentials of Human Anatomy & Physiology.* (9th - 10th International 2012. ed.) San Francisco: Pearson/Benjamin Cummings, 2009.

[13] I. Peate and M. Nair, *Fundamentals of Anatomy and Physiology : For Nursing and Healthcare Students.* New York: John Wiley & Sons, Incorporated, 2016.

[14] J. Heikkilä, M. Mäkijärvi and A. Hedman, *Ekg.* Helsinki: Duodecim, 2003.

[15] Anatomy and Function of the Heart's Electrical System. Available: http://www.gwheartandvascular.org/education/cardiovascular-diseases/anatomy-and-function-of-the-hearts-electrical-system/. [Accessed 23.10.2018]

[16] J. Goy, J. Staufer and J. Schlaepfer, *Electrocardiography (ECG)*, Bentham Science Publishers, 2013.

[17] A. M. Katz, *Physiology of the Heart*, 5th ed., Wolters Kluwer Health, 2011.

[18] A. J. Vander, J. H. Sherman and D. S. Luciano, *Human Physiology: The Mechanisms of Body Function*, 5th ed., New York (NY): McGraw-Hill, 1990.

[19] M. Kupari and J. Lommi, "Sydämen vajaatoiminta," *Kapseli,* vol. 34, pp. 11–68, 2004.

[20] P. Mustajoki, *Kohonnut verenpaine (verenpainetauti)*. Available: http://www.terveyskirjasto.fi/terveyskirjasto/tk.koti?p_artikkeli=dlk00034. [Accessed 25.10.2018]

[21] Suomalaisen Lääkäriseuran Duodecimin ja Suomen Verenpaineyhdistys ry:n asettama työryhmä, *Kohonnut verenpaine. Käypä hoito -suositus*, Helsinki: Suomalainen Lääkäriseura Duodecim, 2014. Available: http://www.kaypahoito.fi/web/kh/suositukset/suositus?id=hoi04010. [Accessed 26.10.2018]

[22] I. Simova, "Intima-media thickness: Appropriate evaluation and proper measurement described," *E-Journal of Cardiology Practice,* vol. 13, no. 21, 2015. Available: https://www.escardio.org/Journals/E-Journal-of-Cardiology-Practice/Volume-13/Intima-media-thickness-Appropriate-evaluation-and-proper-measurement-described. [Accessed 14.10.2018]

[23] E. de Groot, S. I. van Leuven, R. Duivenvoorden, M. C. Meuwese, F. Akdim, M. L. Bots and J. J. Kastelein, "Measurement of carotid intima-media thickness to assess progression and regression of atherosclerosis," *Nature Clinical Practice Cardiovascular Medicine,* vol. 5, no. 5, pp. 280–288, 2008.

[24] J. F. Polak, M. J. Pencina, M. K. Pencina, J. C. O'Donnell and A. P Wolf, "Carotid-Wall Intima–Media Thickness and Cardiovascular Events," *New England Journal of Medicine*, vol. 365, no. 3, pp. 213–221, 2011.

[25] S. C. H. van den Oord, E. J. G. Sijbrands, G. L. ten Kate, D. van Klaveren, R. T. van Domburg, A. F. W. van der Steen and A. F. L. Schinkel, "Carotid intima-media thickness for cardiovascular risk assessment: Systematic review and meta-analysis," *Atherosclerosis,* vol. 228, no. 1, pp. 1–11, 2013.

[26] Risk factors. Available: https://www.world-heart-federation.org/resources/risk-factors/. [Accessed 25.10.2018]

[27] M. Gaarder and T. Seierstad, "Measurements of carotid intima media thickness in non-invasive high-frequency ultrasound images: the effect of dynamic range setting," *Cardiovascular Ultrasound,* vol. 13, no. 1, pp. 5, 2015.

[28] Y. Inaba, J. A. Chen and S. R Bergmann, "Carotid plaque, compared with carotid intima-media thickness, more accurately predicts coronary artery disease events: A meta-analysis," *Atherosclerosis,* vol. 220, no.1, pp. 128–133, 2012.

[29] D. H. O'Leary and M. L. Bots, "Imaging of atherosclerosis: carotid intima–media thickness," *European Heart Journal*, vol. 31, no. 14, pp. 1682–1689, 2010.

[30] T. Nezu, N. Hosomi, S. Aoki and M. Matsumoto, "Carotid Intima-Media Thickness for Atherosclerosis," *Journal of Atherosclerosis and Thrombosis*, vol. 23, no. 1, pp. 18–31, 2016.

[31] M. Bauer, , S. Caviezel, A. Teynor, R. Erbel, A. A. Mahabadi and A. Schmidt-Trucksäss, "Carotid intima-media thickness as a biomarker of subclinical atherosclerosis," *Swiss Medical Weekly,* vol. 142, pp. w13705, 2012.

[32] J. Heikkilä, M. Kupari and J. Airaksinen, *Kardiologia,* 2nd new. ed. Helsinki: Duodecim, 2008.

[33] M. Kelm, "Flow-mediated dilatation in human circulation: Diagnostic and therapeutic aspects," *American Journal of Physiology - Heart and Circulatory Physiology,* vol. 282, no. 1, pp. H5, 2002.

[34] R. A. Harris, S. K. Nishiyama, W. D Wray and R. S. Richardson, "Ultrasound Assessment of Flow-Mediated Dilation," *Hypertension,* vol. 55, no. 5, pp. 1075–1085, 2010.

[35] D. Montero, J. Padilla, C. Diaz-Canestro, M. J. D. Muris, K. E. Pyke, P. Obert and G. Walther, "Flow-Mediated Dilation in Athletes: Influence of Aging," *Medicine and Science in Sports and Exercise,* vol. 46, no. 11, pp. 2148–2158, 2014.

[36] L. Ghiadoni, F. Faita, M. Salvetti, C. Cordiano, A. Biggi, M. Puato, A. Di Monaco, L. De Siati, M. Volpe, G. Ambrosio, V. Gemignani, M. L. Muiesan, S. Taddei, G. A. Lanza and F. Cosentino, "Assessment of flow-mediated dilation reproducibility: a nationwide multicenter study," *Journal of Hypertensio,* vol. 30, no. 7, pp. 1399–1405, 2012.

[37] N. Gokce, J. F. Keaney, L. M. Hunter, M. T. Watkins, Z. S. Nedeljkovic, J. O. Menzoian and J. A. Vita, "Predictive value of noninvasivelydetermined endothelial

dysfunction for long-term cardiovascular events inpatients with peripheral vascular disease," *Journal of the American College of Cardiology,* vol. 41, no. 10, pp. 1769–1775, 2003.

[38] S. Y. Chan, G. B. J. Mancini, L. Kuramoto, M. Schulzer, J. Frohlich and A. Ignaszewski, "The prognostic importance of endothelial dysfunction and carotid atheroma burden in patients with coronary artery disease," *Journal of the American College of Cardiology*, vol. 42, no. 6, pp. 1037–1043, 2003.

[39] M. Shechter, A. Issachar, I. Marai, N. Koren-Morag, D. Freinark, Y. Shahar, A. Shechter and M. S. Feinberg, "Long-term association of brachial artery flow-mediated vasodilation and cardiovascular events in middle-aged subjects with no apparent heart disease," *International Journal of Cardiology,* vol. 134, no. 1, pp. 52–58, 2009.

[40] G. Brevetti, A. Silvestro, V. Schiano and M. Chiariello, "Endothelial Dysfunction and Cardiovascular Risk Prediction in Peripheral Arterial Disease: Additive Value of Flow-Mediated Dilation to Ankle-Brachial Pressure Index," *Circulation,* vol. 108, no. 17, pp. 2093–2098, 2003.

[41] P. C. Choudhari and M. S. Panse, "Intelligent system based on impedance cardiography for non-invasive measurement and diagnosis," in *Intelligent Systems Technologies and Applications,* 2016.

[42] S. Grimnes and Ø G. Martinsen, "Chapter 4 - PASSIVE TISSUE ELECTRICAL PROPERTIES," in *Bioimpedance and Bioelectricity Basics (Second Edition)*, S. Grimnes and Ø G. Martinsen, Eds. 2008.

[43] E. Niemi and J. Malmivuo, *Improving the Accuracy of Impedance Cardiography.* 1990.

[44] V. Seppä, *Development and Clinical Application of Impedance Pneumography Technique.* Tampere University of Technology, 2014.

[45] H. Hutten, "Impedance plethysmography," in *Encyclopedia of Medical Devices and Instrumentation*, J. G. Webster, Ed. 2006.

[46] J. Schaller, S. Gerber, U. Kaempfer, S. Lejon and C. Trachsel, *Human Blood Plasma Proteins: Structure and Function*, 1st ed., Wiley & Sons, Incorporated, 2008

[47] R. B. Northrop, *Noninvasive Instrumentation and Measurement in Medical Diagnosis.* Boca Raton (FL): CRC Press, 2002.

[48] A. K. Deshpande, G. D. Jindal, P.M. Jagasia, K. V. Murali, P. A. Bharadwaj, K. I. Tahilkar and G. B Parulkar, "Impedance plethysmography of thoracic region: impedance cardiography," *Journal of Postgraduate Medicine,* vol. 36, no. 4, pp. 207–212, 1990.

[49] S. Grimnes and Ø G. Martinsen, "Chapter 9 - CLINICAL APPLICATIONS," in *Bioimpedance and Bioelectricity Basics (Second Edition)*, S. Grimnes and Ø G. Martinsen, Eds. 2008.

[50] J. M. Van De Water, T. W. Miller, R. L. Vogel, B. E. Mount and M. L. Dalton," Impedance Cardiography," *Chest,* vol. 123, no. 6, pp. 2028, 2003.

[51] G. Cybulski, A. Strasz, W. Niewiadomski and A. Gasiorowska, "Impedance cardiography: Recent advancements," *Cardiology Journal,* vol. 19, no. 5, pp. 550–556, 2012.

[52] G. Cybulski, "Impedance cardiography," in *Ambulatory Impedance Cardiography: The Systems and their Applications*, G. Cybulski, Ed. 2011.

[53] J. Malmivuo and R. Plonsey, *Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields.* New York: Oxford University Press, 1995.

[54] T. R. Kuphaldt, "Volume I - DC," in *Lessons in Electric* Circuits, 2006.

[55] R. X. Stroobandt, S. S. Barold and A. F. Sinnaeve, *ECG from Basics to Essentials: Step by Step*, Hoboken: John Wiley & Sons, 2015.

[56] L. Stoner, J. M. Young and S. Fryer, "Assessments of arterial stiffness and endothelial function using pulse wave analysis," *International Journal of Vascular Medicine,* vol. 2012, pp. 1–9, 2012.

[57] S. Wassertheurer, J. Kropf, T. Weber, M. Van Der Giet, J. Baulmann, M. Ammer, B. Hametner, C. C. Mayer, B. Eber and D. Magometschnigg, "A new oscillometric method for pulse wave analysis: comparison with a common tonometric method," *Journal of Human Hypertension,* vol. 24, no. 8, pp. 498–504, 2010.

[58] P. Salvi, "Pulse wave analysis," in *Pulse Waves: How Vascular Hemodynamics Affects Blood Pressure*, P. Salvi, Ed. 2012.

[59] I. Mellin, *Monimuuttujamenetelmät: Moniulotteiset jakaumat ja havaintoaineistot*. Available: salserver.org.aalto.fi/vanhat_sivut/Opinnot/Mat-2.3128/.../MONIULJAKAIN.pdf. [Accessed 20.1.2019]

[60] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics,* vol. 7, no. 2, pp. 179–188, 1936.

[61] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, , Academic Press, 2015.

[62] A. R. Webb, K. D. Copsey and G. Cawley, *Statistical Pattern Recognition*, Hoboken: John Wiley & Sons, 2011.

[63] V. Kotu and B. Deshpande, *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*, San Francisco: Elsevier Science & Technology, 2014.

[64] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., New York: Springer, 2009.

[65] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York: Springer 2006.

[66] R. Spataro, A. Chella, B. Allison, M. Giardina, R. Sorbello, S. Tramonte, C. Guger and V. La Bella, "Reaching and grasping a glass of water by locked-In ALS patients through a BCI-controlled humanoid robot," *Frontiers in Human Neuroscience,* vol. 11, pp. 68, 2017.

[67] Support Vector Machines for Binary Classification. Available: https://se.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html. [Accessed 7.1.2019]

[68] L. Breiman, "Random Forests," *Mach. Learning,* vol. 45, no. 1, pp. 5–32, 2001.

[69] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences,* vol. 55, no. 1, pp. 119–139, 1997.

[70] J. Kacprzyk, I. Guyon, M. Nikravesh and L. A. Zadeh, *Feature Extraction : Foundations and Applications*, Springer Berlin Heidelberg, 2006.

[71] T. J. Cleophas and A. H. Zwinderman, *Statistics Applied to Clinical Studies*, 5th ed., Heidelberg: Springer, 2015.

[72] A. Meyer-Bäese and V. J. Schmid, *Pattern Recognition and Signal Analysis in Medical Imaging*, 2nd ed., Boston: Academic Press, 2014.

[73] H. Aatola, N. Hutri-Kähönen, M. Juonala, J. Viikari, J. Hulkkonen, T. Laitinen, L. Taittonen, T. Lehtimäki, O. Raitakari and M. Kähönen, "Lifetime Risk Factors and Arterial Pulse Wave Velocity in Adulthood: The Cardiovascular Risk in Young Finns Study," *Hypertension,* vol. 55, no. 3, pp. 806–811, 2010.

[74] S. Munir, Guilcher, T. Kamalesh, B. Clapp, S. Redwood, M. Marber and P. Chowienczyk, "Peripheral Augmentation Index Defines the Relationship Between Central and Peripheral Pulse Pressure," *Hypertension,* vol. 51, no. 1, pp. 112–118, 2008.

[75] S. Eskelinen. *Glukoosi*. Available: https://www.terveyskirjasto.fi/terveyskirjasto/tk.koti?p_artikkeli=snk03091. [Accessed 2.12.2018]

[76] D. C. Montgomery, E. A. Peck, G. G. Vining, A. M. Ryan, A. G. Ryan and G. G. Vining, *Introduction to Linear Regression Analysis*, John Wiley & Sons, Incorporated, 2015.

APPENDIX A: COEFFICIENTS OF THE Z-SCORE NORMALIZATION

This appendix contains the mean (μ) and standard deviation (σ) of the z-score. The number in bold font refer the name of the parameter in Appendix B. The hyphen (–) related to parameters 2, 12 and 14 indicates that the parameter is binary and thus not z-score normalized.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
μ	37.6381	–	8.1897	0.6272	1.9062	25.8904	125.4453	3.0916	1.3284	0.1697
σ	5.0030	–	1.4873	0.0975	0.7718	4.6355	13.8299	0.8017	0.3214	0.4894

	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
μ	1.6636	–	76.7552	–	171.9448	76.8714	92.9852	120.5070	75.3917	90.4302
σ	0.1141	–	9.2767	–	9.1736	16.5555	10.1814	14.3306	11.3137	11.6363

	21	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
μ	5.0198	65.5633	–	–	8.9518	3.0069	1.5872	1.4934	1.8897	0.7777
σ	0.9076	9.6321	–	–	4.5926	1.4551	0.1784	0.3119	0.2226	0.069

	31	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
μ	1.5528	0.3466	0.2276	0.3875	0.0551	0.1630	0.2046	0.0936	0.2765	0.3471
σ	0.1454	0.0329	0.0296	0.0353	0.0140	0.0227	0.0185	0.0239	0.0371	0.0297

	41	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**
μ	0.7457	0.8334	0.7153	2.4354	3.1755	2.3454	0.9806	0.9661	0.9951	0.6088
σ	0.0268	0.0212	0.0386	0.0537	0.5417	0.0246	0.0487	0.1307	0.2651	0.0888

	51	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
μ	0.4324	0.6201	0.4710	0.6334	0.5085	0.4506	0.4698	0.4818	0.4946	0.3708
σ	0.0776	0.0915	0.0932	0.0915	0.0818	0.0949	0.0915	0.0799	0.0956	0.0698

	61	**62**	**63**	**64**	**65**	**66**	**67**	**68**	**69**	**70**
μ	0.3769	-0.6621	-1.1860	-1.9779	0.3931	3.5513	0.4558	5.8338	0.4361	0.4330
σ	0.0798	0.1876	0.3733	0.6037	0.1118	0.8717	0.1119	2.0545	0.1302	0.1381

	71	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**	**80**
μ	0.1827	3.1626	0.3899	4.3631	0.6589	0.7120	3.0583	0.3620	3.4231	0.4381
σ	0.1698	2.2890	0.2047	2.6928	0.1716	0.1607	2.6885	0.0953	2.5233	0.1059

	81	**82**
μ	0.4871	0.4997
σ	0.0804	0.1502

APPENDIX B: INDEXING OF THE CLASSIFYING PARAMETERS

This appendix contains the names of the used features in this thesis.

1) Age
2) Sex
3) PWV
4) IMT
5) Fasting insulin
6) BMI
7) SBP during Circmon-measurement
8) LDL
9) HDL
10) Triglycerides
11) Fasting glucose
12) Smoking
13) DBP during Circmon-measurement
14) Antihypertensive medication
15) Height
16) Mass
17) MAP during Circmon-measurment
18) SBP
19) DBP
20) MAP
21) Total cholesterol
22) Heart rate
23) Plaque
24) Hypertension
25) FMD-%
26) City
27) FFT of ICG
28) FFT of IPG
29) FFT of transformed IPG
30) Ratio between area under transformed IPG and area under ICG
31) Ratio between area under IPG and area under ICG
32) Rise time of ICG
33) Rise time of IPG
34) Rise time of transformed IPG
35) Normalized transmission time of ICG
36) Normalized transmission time of IPG
37) Normalized transmission time of transformed IPG
38) Unnormalized transmission time of ICG
39) Unnormalized transmission time of IPG
40) Unnormalized transmission time of transformed IPG
41) Decay time of ICG
42) Decay time of IPG
43) Decay time of transformed IPG

44) Length of curve of ICG
45) Length of curve of IPG
46) Length of curve of transformed IPG
47) R_4 of transformed IPG
48) R_4 of ICG
49) R_4 of IPG
50) R_1 of transformed IPG
51) R_1 of ICG
52) R_2 of transformed IPG
53) R_2 of ICG
54) R_3 of transformed IPG
55) R_3 of ICG
56) T_1 of transformed IPG
57) T_1 of ICG
58) T_2 of transformed IPG
59) T_2 of ICG
60) T_3 of transformed IPG
61) T_3 of ICG
62) AGI of transformed IPG
63) AGI of ICG
64) AGI of IPG
65) L5 of ICG with time difference between 1^{st} and 2^{nd} highest peaks
66) Gln4 of ICG with difference of amplitudes of 1^{st} and 2^{nd} highest peaks
67) Gln4 of ICG with time difference between 1^{st} and 2^{nd} highest peaks
68) Gln4 of ICG with ratio of areas of 1^{st} and 2^{nd} highest peaks
69) L5 of ICG with time difference between 1^{st} and 3^{rd} highest peaks
70) Gln4 of ICG with time difference between 1^{st} and 3^{rd} highest peaks
71) L5 of IPG with time difference between 1^{st} and 2^{nd} highest peaks
72) Gln4 of IPG with difference of amplitudes of 1^{st} and 2^{nd} highest peaks
73) Gln4 of IPG with time difference between 1^{st} and 2^{nd} highest peaks
74) Gln4 of IPG with ratio of areas of 1^{st} and 2^{nd} highest peaks
75) L5 of IPG with time difference between 1^{st} and 3^{rd} highest peaks
76) Gln4 of IPG with time difference between 1^{st} and 3^{rd} highest peaks
77) L5 of transformed IPG with difference of amplitudes of 1^{st} and 2^{nd} highest peaks
78) L5 of transformed IPG with time difference between 1^{st} and 2^{nd} highest peaks
79) Glne4 of transformed IPG with difference of amplitudes of 1^{st} and 2^{nd} highest peaks
80) Gln4 of transformed IPG with time difference between 1^{st} and 2^{nd} highest peaks
81) L5 of transformed IPG with time difference between 1^{st} and 3^{rd} highest peaks
82) Gln4 of transformed IPG with time difference between 1^{st} and 3^{rd} highest peak

APPENDIX C: CLASSIFYING PARAMETERS SELECTED IN THE EVALUATED SCENARIOS

This appendix consists of tables that contain the features that were found for each classifier in the feature selection phase.

Table 19: Classifying parameters in all dataset.

Labeling method		AdaBoost	K-nearest neighbor	LDA	Naive Bayes	QDA	Random forest	SVM
LM1	Clinical + pulse parameters	13, 17, 21	13, 21, 23, 28, 46	13, 17, 21, 31, 56, 60	4, 13, 21, 45, 47, 51	13, 21, 39, 49, 75	13, 14, 21, 24	13, 21, 46
	Clinical	13, 18, 21	4, 13, 19, 20, 21	1, 13, 14, 19, 21, 23, 25	13, 21, 25	13, 20, 21, 22	13, 14, 21, 23	4, 13, 14, 20, 21, 24
	Pulse parameters	32, 37, 44	32, 47, 51	40, 43, 55, 66	27, 37, 43, 55	49, 51, 68, 76	32, 44, 51, 59, 76, 77	31, 39, 41, 66
LM2	Clinical + pulse parameters	1	22, 47	5, 20, 25, 45, 55, 68	13, 25, 31, 52, 78	1, 4, 5, 31, 34	1	2, 24, 38 ,55
	Clinical	1	22	17	1, 5, 7, 12, 25	7	1	1
	Pulse parameters	60	29, 55, 63	48, 55, 71	29	29	44, 47, 48, 51, 53, 69	51, 55, 71
LM3	Clinical + pulse parameters	1	1, 13, 78	5, 11, 29	1	1	1	5
	Clinical	11	5, 11	5, 11	5, 12	1	1	5
	Pulse parameters	29	40	27	27	27	40, 66, 74, 78	27

Table 20: Classifying parameters in dataset where blood tests are removed.

Labeling method		AdaBoost	K-nearest neighbor	LDA	Naive Bayes	QDA	Random forest	SVM
LM1	Clinical + pulse parameters	13, 47	13, 70	13, 23, 43, 51, 62, 71	13, 39, 47, 51	1, 2, 4, 13, 47, 51, 62, 63	13, 14, 24	13, 43, 59
	Clinical	13, 22	13, 14, 19, 22, 23	1, 4, 13, 24	1, 13	1, 4, 13, 26	13, 14, 24	13, 19
	Pulse parameters	31, 33, 36, 42, 43, 62, 81	42, 47, 51, 77	40, 42, 51, 55, 56, 59, 62, 71, 74	27, 40, 42, 48, 50, 55, 62	29, 31, 40, 55, 70	53, 56, 81	32, 50, 55, 59, 63, 77
LM2	Clinical + pulse parameters	1	18, 23	2, 14, 24, 47, 55	1, 13, 31, 63	1	1	7, 14, 44, 68
	Clinical	1	22	7	1	1	1	1
	Pulse parameters	60	53	55, 65	29	29, 31, 37, 65	43, 49, 51, 54, 58, 61, 73, 74, 75	44, 53, 57, 59, 65, 68
LM3	Clinical + pulse parameters	82	1, 13, 78	1	1	1	1	1
	Clinical	1	1, 6, 16, 18	16	1	1	1	1
	Pulse parameters	29	27	27	27	27	45, 67, 68, 69, 73	27

Table 21: Classifying parameters in old age group.

Labeling method		AdaBoost	K-nearest neighbor	LDA	Naive Bayes	QDA	Random forest	SVM
	Clinical + pulse parameters	13, 21, 43	13, 14, 21, 23, 45, 46, 69	13, 21, 26, 28, 31, 41	13, 21, 50, 51, 62, 64, 69	13, 21, 29, 50, 70	22	13, 21, 23, 24, 42, 46
LM1	Clinical	13, 21	2, 13, 21	3, 13, 14, 21, 22	13, 21	4, 13, 21, 26	2, 13, 18, 21	13, 19, 21
	Pulse parameters	27, 36, 37, 43, 44, 55	43, 49, 51, 61, 76	43, 49, 55, 80	27, 28, 43, 51	34, 37, 43, 51, 78	43, 45, 51, 58, 71	43, 51, 69, 79
	Clinical + pulse parameters	22	2, 5, 23	5, 14, 20, 25, 34, 55, 81	2, 4, 5, 42	7	5	43, 58, 69
LM2	Clinical	2	2, 5, 23	5, 7, 19	2, 4, 5	2, 4, 5, 26	5	2, 5, 25
	Pulse parameters	49	28	41, 56, 57	30, 47, 63	29	36, 49, 51, 58, 71	58, 62, 78
	Clinical + pulse parameters	2	12, 72	11	77	5	2	2
LM3	Clinical	2	11, 17	5, 6, 25	5	5	2	2
	Pulse parameters	27	27, 71	27	27	27	34, 50, 63	27

Table 22: Classifying parameters in young age group.

Labeling method		AdaBoost	K-nearest neighbor	LDA	Naive Bayes	QDA	Random forest	SVM
LM1	Clinical + pulse parameters	13, 21	4, 13, 20, 21, 77	2, 13, 14, 21, 47, 74	13, 18, 21, 70	13, 21, 26, 32, 39, 53	2	13, 21, 31, 36, 39
	Clinical	4, 19, 22	2, 4, 13, 21, 23, 26	13, 18, 21, 22, 23	4, 13, 21, 22	4, 13, 19, 21, 25	13, 20, 21, 22	13, 18, 21, 24
	Pulse parameters	42, 46, 75, 81	32, 54	30, 40, 48, 49, 55, 57, 59, 76	31, 55, 78, 81	31, 40, 62, 80	70	31, 39, 72, 78
LM2	Clinical + pulse parameters	22	5, 74, 78	5, 23, 24, 30, 31, 47	–	38, 40, 41, 62, 65, 78	2	2, 4, 5, 14, 23, 37, 41, 55
	Clinical	22	12, 14, 17	4, 5, 14, 19, 22, 23, 26	–	5, 7	2	2
	Pulse parameters	78	73	36, 44	29, 31, 52, 62, 65, 78	31, 47, 51, 52, 67, 69	30, 41, 43	28, 55, 59, 77, 78
LM3	Clinical + pulse parameters	21	2	11	2	2	2	2
	Clinical	2	2	11, 12	2	2	2	2
	Pulse parameters	27	27	27	27	27	38, 45	27

Table 23: Classifying parameters of men.

Labeling method		AdaBoost	K-nearest neighbor	LDA	Naive Bayes	QDA	Random forest	SVM
LM1	Clinical + pulse parameters	18, 21, 31, 32, 35, 47, 56	4, 13, 21, 23, 24, 27, 43, 50, 52	13, 21, 23, 24, 42, 48, 62	13, 21, 25, 27, 33, 34, 51, 62	13, 21, 31, 62, 72	1	32, 44, 52, 66
	Clinical	3, 14, 20	4, 18	13, 21, 22, 23, 24	13, 18, 21	13, 18, 21, 22, 25	1	13, 21, 23
	Pulse parameters	31, 32, 44, 49, 52, 77, 82	61	31, 32, 42, 44, 55, 61, 62	32, 68	31, 32, 37, 74	30, 32, 43, 52, 63	31, 32, 50, 56, 63, 70
LM2	Clinical + pulse parameters	35	14, 22, 23, 62	7, 20, 59	–	80	1	7, 27
	Clinical	4, 20	1	20	–	1, 4, 22	1	1
	Pulse parameters	42	31, 36, 37, 46	44	80	32, 40, 45, 48, 65	31, 43, 60, 78	27
LM3	Clinical + pulse parameters	1	5	1	1	1	1	4, 5
	Clinical	1	1, 4	11	1, 22	1	1	1
	Pulse parameters	27	28, 32	27	27	27	32, 49, 62, 67, 73	27

Table 24: Classifying parameters of women.

Labeling method		AdaBoost	K-nearest neighbor	LDA	Naive Bayes	QDA	Random forest	SVM
LM1	Clinical + pulse parameters	13, 20, 21	14, 18, 19, 21, 24, 26	1, 13, 14, 21, 24, 36	13, 21, 38, 51	13, 18, 21, 33, 35, 39, 40, 61	13, 21, 31, 49, 67	13, 21, 24, 31
	Clinical	13, 18, 21	13, 14, 18, 21, 23	1, 13, 20, 21, 22, 23, 26	13, 21, 26	13, 19, 21, 26	13, 20, 21	13, 21
	Pulse parameters	27, 55, 57, 69, 79	47, 51, 57	27, 28, 50, 55, 61, 63, 71	35, 40, 43, 52, 55	27, 43, 51, 67	40, 43, 58, 69	55, 65
LM2	Clinical + pulse parameters	53	3, 5, 14, 60, 63	14, 47, 55, 78	–	3, 40, 55	1	14, 24, 35, 38, 43, 55
	Clinical	5	5, 12, 13, 14, 24	5, 12, 14, 19, 22	–	5, 12	1	5, 25
	Pulse parameters	38, 51, 55, 62, 68, 79	27, 50, 54, 55, 71	28, 47, 55, 39	37, 43, 55, 57	35, 38, 43, 55, 57	27, 42, 47, 55, 72	38, 43, 55
LM3	Clinical + pulse parameters	11	1	8, 11	1	1	11	1
	Clinical	1	1	1, 5, 20	1	1, 5, 22, 25	1	1
	Pulse parameters	43	27	27	27	27	28, 29, 62, 63	27