MASTER THESIS

Opportunities and possibilities of outcome prediction in patients who will undergo a transcatheter aortic valve procedure

A data-driven approach

Eline V. Schaft

Faculty of Science and Technology
Mater program Technical Medicine
Specialization Medical Sensing and Stimulation

Faculty of Electrical Engineering, Mathematics and Computer Science
Master program Applied Mathematics
Specialization Applied Analysis and Computational Science

EXAMINATION COMMITTEE
Prof. dr. S.A van Gils
Dr. M.M. Vis
Dr. C. Brune
Dr. M.S. van Mourik
Dr. G. Meinsma
Drs. N.S. Cramer Bornemann

December 11th, 2019

Amsterdam UMC
University Medical Centers

UNIVERSITY OF TWENTE.

Preface

This graduation project is a combined project to obtain my MSc Technical Medicine and MSc Applied Mathematics at the University of Twente. The research is conducted at the Cardiology department in the Amsterdam UMC, location AMC. This study focuses on the prediction of TAVI outcome using deep learning techniques on electrocardiograms.

This work is divided into three parts. The first part focuses on the extraction of features from a convolutional neural network. The second part describes a manner to follow the changes of these features over time (in particular before and after TAVI). In the final part the clinical implementation, especially the risks and possibilities, are given.

I would like to thank my supervisors, Martijn, Christoph, Marije and Stephan. Thank you for all your help and sharing your knowledge with me. I learned a lot from all of you and this work wouldn't be here without your help. I also want to thank you for helping me in times when I had a hard time. It really felt like all of you were always there for me. Gjerrit, thank you for being the independent member of my graduation committee. I would like to thank my colleagues at the department of intervention cardiology. Jimmy, thank you for the fun moments at the catheterization lab. Carlijn, thank you for always listening to me and the discussions we had. Ricardo, thank you for your help with the technical details of my work. It was really nice to collaborate with you and helping each other out with our knowledge. Friends and family, you're the best. Finally, I want to thank Victor, for his patience and endless support.

Amsterdam, 2019
E.V. Schaft

Abstract

Patients with an aortic stenosis who have a high risk for undergoing surgery can get a TAVI. The outcomes of the TAVI procedure are the same or even better than the outcomes of the surgical aortic valve replacement. However, there are still patients who undergo a TAVI, but do not show improvement or die within one year. In this project machine learning techniques are used to evaluate what the opportunities and possibilities are to predict the outcome of patients who will undergo a TAVI with the use of ECG. This prediction model can help a cardiologist in the decision making.

First, one ECG is used to train a convolutional neural network to classify atrial fibrillation in an ECG. This network is used with transfer learning to predict two outcomes of TAVI patients: mortality within one year and the improvement of dyspnea after the TAVI procedure. The mortality can be predicted with an accuracy of 72.3% and the improvement of dyspnea can be predicted with an accuracy of 90.1%. These results show the potential of the use of ECG in the prediction of the outcome of patient who underwent a TAVI.

To further investigate the potential of ECG in the prediction of the outcome, the course of the ECG is used: two ECGs, one before and one after TAVI. For this, a feature extraction algorithm is used based on signal analysis. The features that are extracted from the ECG are: HR, PR interval, RT interval and the presence of an irregular heartbeat. These features are used to predict the same outcomes as in the first part, mortality and the improvement of dyspnea. These four features are extracted from ECGs before and after TAVI, resulting in eight features. These features are used in two different ways: stacked into one feature vector of eight features or the distance is calculated by the Kullback-Leibler divergence, resulting in a feature vector of four features. A neural network with one hidden layer is trained with the stacked features, resulting in an accuracy of 81.1% for the prediction of mortality and an accuracy of 75.8% for the prediction of improvement of dyspnea. A support vector machine with the stacked features resulted in accuracies of 78.9% and 68.2%, respectively. A support vector machine with the distance features has an accuracy of 71.4% and 69.7% respectively.

The analysis shows that ECG and machine learning are able to predict the improvement of dyspnea and mortality. The best results in mortality prediction are obtained with two ECGs in contrast to the best results in prediction of improvement of dyspnea with one ECG. To finalize this work, the steps to implement a data-driven model in the daily clinical practice are elaborated.

Contents

List of abbreviations

		Introduced at page
AF	Atrial fibillation	1
ADDMoP	Amsterdam Data-Driven Model for Prediction	58
AoS	Aortic valve stenosis	4
AVA	Aortic valve area	4
BMI	Body mass index	25
CAG	Coronary angiography	5
CABG	Coronary artery bypass grafting	25
Cath lab	Catheterization laboratory	5
CNN	Convolutional neural network	1
COPD	Chronic obstructive pulmonary disease	25
CTA	Computed tomography angiography	5
DL	Deep learning	1
ECG	Electrocardiogram	1
EF	Ejection fraction	4
EuroSCORE	European System for Cardiac Operative Risk Evaluation	25
eGFR	Estimated glomerular filtration rate	25
FN	False negative	17
FP	False positive	17
FRANCE	French Aortic National CoreValve and Edwards	58
HR	Heart rate	7
HRV	Heart rate variability	52
JSON	JavaScript Object Notation	57
KL	Kullback-Leibler	16
LVF	Left ventricle function	4
ML	Machine learning	1
NLL	Negative log-likelihood	16
NN	Neural network	12
NYHA	New York Heart Association	8
PCI	Percutaneous coronary intervention	25
PPV	Positive predictive value, also precision	17
REST API	Representation State Transfer Application Programming Interface	57
RVF	Right ventricle function	25
SAVR	Surgical aortic valve replacement	1
SGD	Stochastic gradient descent	15
SPAP	Systolic pulmonary artery pressure	25
STS	Society of thoracic surgeons	25
SVM	Support vector machine	12
TAVI	Transcatheter aortic valve implantation	1
TF	Trans femoral	5
TN	True negative	17
TNR	True negative rate, also specificity	17
TP	True positive	17
TPR	True positive rate, also sensitivity	17
TTE	Transthoracic echocardiography	1
WCE	Weighted cross-entropy	16

Chapter 1

Introduction and motivation

The development of transcatheter aortic valve implantation (TAVI) has increased significantly over the past decade. TAVI is possible in patients with aortic valve stenosis (AoS). Where this procedure used to be chosen for the high-risk patients, who were denied for surgical treatment of the stenosis, the procedure is now also used for patients with an intermediate or low risk. Different studies have shown that outcomes of the TAVI procedure are the same or even better than the outcomes of the surgical aortic valve replacement (SAVR). [1, 2, 3, 4, 5] However, there are still patients who undergo a TAVI, but do not show improvement or die within one year. [6]

A lot of data is collected from patients during screening, procedure and follow up from TAVI, including clinical and laboratory data, electrocardiograms (ECGs), transthoracic echocardiography (TTE) and CT scans. An overview of this data collection is shown in Figure 1.1. This data might be useful to help the clinician in their decision making. Ideally, all the data is used for this purpose. As a first start a project within the Amsterdam UMC focused on the prediction of outcome based on clinical and laboratory data.[7] This study showed a slightly better performance of machine learning (ML) techniques for predicting mortality after TAVI in comparison to traditional methods. A next step will be to include other types of data in a model to predict outcome after TAVI, such as ECGs. A hypothesis is that the more data is included in the model, the more accurate the predictions will be. It is known that rhythm disorders, for example atrial fibrillation (AF), are related to bad outcomes in TAVI. [8] Conduction disorder can be seen in an ECG. The assumption is that the ECG contains information about the electrical activity of the heart and has therefore a predictive value for the outcome of TAVI patients. An addition to this hypothesis is that the course of the ECG over time generates additional information.

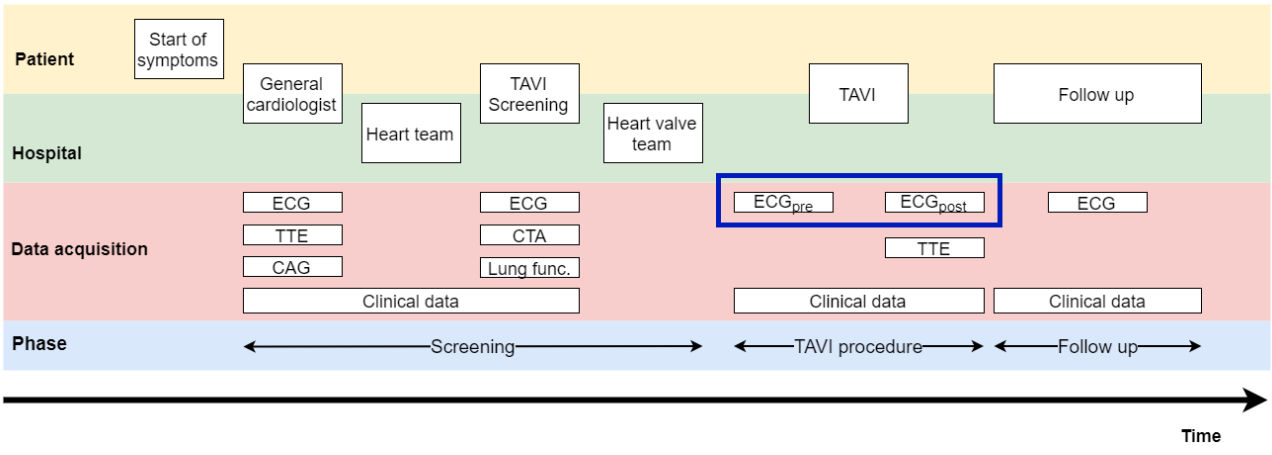


Figure 1.1: An overview of the data collected from a patient who will undergo a TAVI. In this project the focus is on two ECGs: ECG_{pre} and ECG_{post}. These ECGs (shown in the blue frame) are used to predict the outcome of patients who will undergo a TAVI.

Recently, ML models have been developed on clinical data. Different medical related problems can be addressed with deep learning (DL) techniques: abnormality detection, segmentation of images, classification, diagnosis and prediction of outcome. [9] For example, prediction of sepsis in the intensive care unit or prediction of functional outcome in ischemic stroke patients. [10, 11] However, these algorithms are rarely used to help the clinician in daily practice. If specifically looking at DL for physiological signals, electromyogram, electroencephalogram, ECG and electrooculogram, the number of articles is growing each year. [12] The most used architecture in ECG is a convolutional neural network (CNN). Other architectures are recurrent neural network and auto encoder. In this work the focus is on the combination of CNN and ECG. The method of CNN is chosen, because it has shown good results in earlier studies. [12] In these studies it has reached accuracy above 90% in arrhythmia detection, heartbeat classification and diagnosis. Other advantages of CNNs are that they are easy to learn and the possibility to perform visualization techniques to investigate what the network has learned. This insight in the black box, the CNN, is very important in this application, because a cardiologist wants to understand what the network is learning. If a cardiologist understands what the network learns and if it is easy to use, the model can be used in clinical practice.

The goal of this project is to investigate how ECG can help with the prediction of outcome of TAVI patients. This question is two-folded. The first thing is that a better prediction before the TAVI can help the decision making of the clinicians. The other thing is that the workflow after TAVI can be personalized for each patient. For example, patients with high risk on complications will have a more frequent follow up than patients with low risk. The focus is on two outcomes: improvement of dyspnea, by using the NYHA classification score, and the mortality, by calculating the survival days between the date of the TAVI procedure and the mortality date. These outcomes are chosen, because dyspnea is one of the most common symptoms in patient with AoS and the mortality is chosen, because this is an important criterion in the current protocols for a patient to get approved for a TAVI. This research is unique to earlier studies by using the 8 leads ECG signal instead of using only one or two leads and by using the ECG signal for prediction instead of detection.

The two hypotheses stated are tested throughout this report. These hypotheses are summarized in Table 1.1. In Chapter 2 and Chapter 3, the necessary clinical and technical background is explained. The first hypothesis is tested in Chapter 4. In this part, the use of a DL method, the CNN, is investigated. The ECG after TAVI is used, because this ECG is closest in time to the outcome which is predicted. Since AF is a predictor for a bad outcome, the network is first trained to detect AF. With transfer learning the network is used to predict the improvement of dyspnea and the mortality. In this part also visualization techniques are applied to break open the black box, CNN, and to try to understand how the network makes the prediction. The second hypothesis time is tested in the second part Chapter 5. In this part, the ECG before TAVI and the ECG after TAVI are both used to predict the outcomes. A model must be found to translate the features from the ECG before TAVI into the features after TAVI and to make a prediction based on this. In the final part, the focus is on the implementation of such a model in the clinical practice. This can be found in Chapter 6. To make it easy for the clinician to use such a model, the technical details must be right. In this part, a pathway is described where the technical challenges are named and how it can be used in the clinical practice.

Table 1.1: Overview of the hypotheses tested in this report

Number	**Hypothesis**	**Chapter**
1	One ECG contains information about the electrical activity of the heart and has a predictive value for the outcome of TAVI patients.	4
2	Mulitple ECGs have additional value in predicting the outcome of TAVI patients.	5

Chapter 2

Clinical Background

In this chapter, the clinical background is described. First the disease, AoS, is introduced. Then the treatment, TAVI, is described. In the next two sections the mechanism of the ECG and its clinical features are presented. To complete this chapter, a section about all the data which is collected from patients is added.

2.1 Aortic valve stenosis

The heart consists of four chambers, two atria and two ventricles. The heart can be divided in the right and the left side. The right atrium and the right ventricle form the right side of the heart. The left side consists of the left atrium and the left ventricle. Blood flows from the superior and inferior vena cava into the right atrium, where it is pumped to the right ventricle during the atrial systole. This is shown in Figure 2.1. During systole, the ventricular wall contracts so the blood flows into the pulmonary artery. Then the blood is oxygenated in the lungs and flows back to the heart trough the pulmonary vein and the left atrium. During the atrial systole, the left atrium contracts as well so the blood flows into the left ventricle. Simultaneously to the right ventricle, the left ventricle contracts during systole, pumping the blood into the aorta. The aorta transports the blood to all parts of the body after which the blood returns to the superior and inferior vena cava. An ECG shows the electrical activity during systole and diastole. This will be explained in the Section 2.3.

Figure 2.1: Blood flow in the heart [13]

Four valves separate the ventricles from the atria and the arteries. The main purpose of the valves is to prevent a back return of the blood. The tricuspid valve is located between the right atrium and ventricle, the mitral valve between the left atrium and ventricle, the pulmonary valve between the right ventricle and the pulmonary artery and the aortic valve between the left ventricle and the aorta.

In this project the focus is on the aortic valve. There exist different diseases of the aortic valve, bicuspid valve, aortic insufficiency, AoS or a combination of these diseases. In this project, the focus

is on AoS, which is a narrowing of the aortic valve. [14] This causes a higher pressure in the left ventricle which eventually can lead to heart failure. AoS can be caused by degenerative calcification, rheumatic heart disease or the cause can be congenital. [14] AoS is one of the most common valvular heart disease. [14, 15] In at least 2% of persons older than 70 years AoS is present. [16] AoS knows different stages, risk of AoS (mild), progressive hemodynamic obstruction (moderate) and severe AoS (both asymptomatic and symptomatic). Once symptoms of AoS occur, the mortality rate increases significantly, with about 25% per year. [14] Symptoms of AoS are dyspnea, angina pectoris, fatigue, syncope and heart failure.

Symptomatic AoS is diagnosed with a combination of the anamnesis, physical examination and an echocardiographic examination. During this examination different measurements are assessed: the thickness of the aortic valve, the area of the aortic valve that is opening during systole (AVA), the left ventricle function (LVF) with the ejection fraction (EF) and the mean and maximal transaortic pressure gradient. With Doppler the flow through the aortic valve and the ante grade flow across the aortic valve is examined. This can be used to calculate the maximum velocity over the aortic valve. These parameters are used to classify with stage of AoS is present. An overview is shown in Table 2.1.

Table 2.1: Stages of severe AoS with echocardiographic outcomes [17]

Class	**Description**	**Echocardiographic outcomes**
C1	Asymptomatic severe AoS with normal LVF	$V_{max} \geq 4$ m/sec or $\Delta P_{mean} \geq 40$ mmHg EF $\geq 50\%$
C2	Asymptomatic severe AoS with poor LVF	$V_{max} \geq 4$ m/sec or $\Delta P_{mean} \geq 40$ mmHg EF $< 50\%$
D1	Symptomatic severe high-gradient AoS	$V_{max} \geq 4$ m/sec or $\Delta P_{mean} \geq 40$ mmHg
D2	Symptomatic severe low-gradient AoS with poor LVF	AVA $\leq 1\ cm^2$ $V_{max} \geq 4$ m/sec or $\Delta P_{mean} \geq 40$ mmHg EF $< 50\%$
D3	Symptomatic severe low-flow, low-gradient AoS	Indexed $\text{AVA}_{\text{indexed}} \leq 0.6 cm^2$ SV index $< 35ml$ $\Delta P_{mean} \geq 40$ mmHg

V_{max}: maximal velocity through aortic valve, ΔP_{mean}: mean transaortic pressure gradient, EF: ejection fraction, AVA: aortic valve area, SV index: systolic volume index in the left ventricle.

2.2 Transcatheter aortic valve implantation

Patients with severe AoS can be treated by a TAVI. This is a minimally invasive procedure to replace the stenotic aortic valve with a biological valve. This procedure is in favor if the patient is old (> 75 yr.), has comorbidities, has had heart surgery before and/or has heart failure. [18, 19]

Since the first TAVI procedure in 2002, a lot of research has been performed to compare the outcomes of TAVI with the SAVR. Initially, the procedure was performed on high-risk patients only. In the past years also intermediate-risk and low-risk patients are included in the TAVI procedure. Different studies have shown that outcomes of the TAVI procedure are the same or even better than the outcomes of the SAVR. [1, 2, 3, 4, 5]

In Figure 2.2, the patient journey before and after a TAVI is shown. Before the TAVI, a period of screening exists. The patients are evaluated if they are fit for a TAVI procedure. If the patient has a murmur or gets symptoms, they are redirected to a general cardiologist by their general practitioner. This cardiologist diagnoses the patient with AoS and in the case of severe AoS, the patient can be redirected to a heart center. During the consult with the general cardiologist, the first ECG is made. This hospital is often not the Amsterdam UMC, but a referring center. These ECGs were not available for this study. During the first heart team, which is a multidisciplinary consultation, the decision is

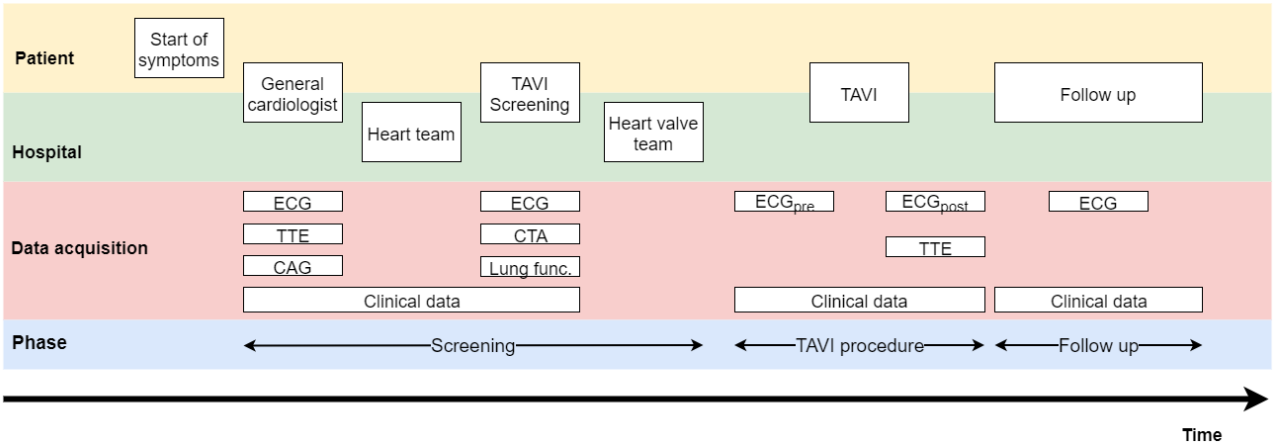


Figure 2.2: Patient journey before and after TAVI. ECG: electrocardiography, TTE: transthoracic echocardiography, CAG: coronary angiogram, CTA: computed tomography angiography, lung func.: lung function. The clinical data consists of the anamnesis, symptoms, comorbidities and laboratory results

made to perform a SAVR or a TAVI. This depends mostly on the age of the patient and the comorbidities. If a patient is selected for TAVI, first a specific TAVI screening is performed. This includes among other things a computed tomography angiography (CTA). This scan is used to evaluate the femoral arteries, the coronaries and the aortic valve. During the heart valve team, another multidisciplinary consultation, the decision for a TAVI is made and the approach of the TAVI is chosen. There are three approaches for a TAVI: transfemoral, transapical and transaortic. The preferred approach is the transfemoral (TF) approach. The other approaches are used when the femoral arteries are not suited for a catheter. An important criterion is the predicted survival time of more than one year. [20]

The TF approach is performed under local anesthesia. This procedure takes place in a cardiac catheterization laboratory (cath lab). The new valve is placed in the old aortic valve using X-ray. With rapid pacing, the flow through the aortic valve drops and the new valve is expanded. The old valve stays in place and the calcification is used to keep the new valve in place. Different complications can occur, like hemorrhage or rhythm disorders, like AF. It is founded that both baseline ECG before TAVI and the ECG after TAVI have a predictive value for long-lasting rhythm disorders. [8]

2.3 Electrocardiography

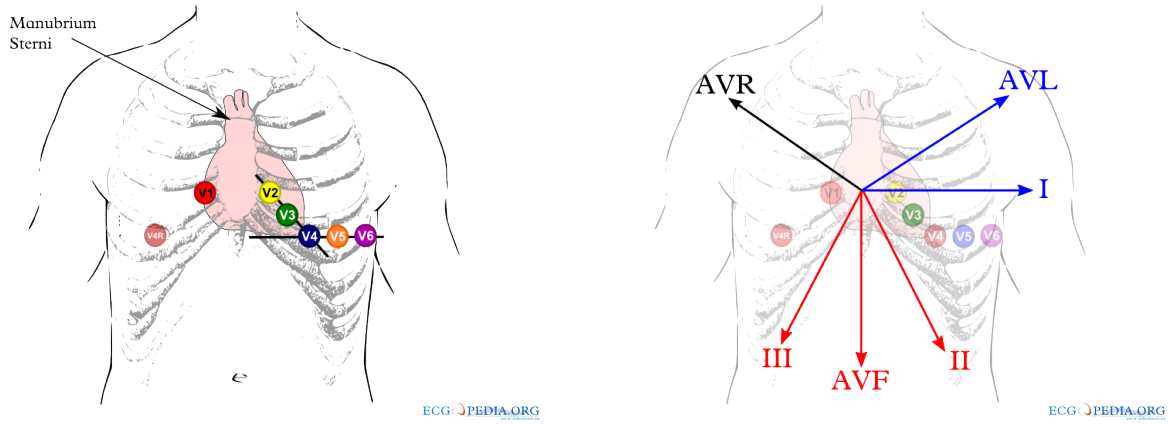
An ECG shows the electrical activity of the heart. This electrical activity is caused by activation of the cardiac muscles. An ECG is recorded with 12 leads: I, II, III, AVF, AVL and AVR are the extremity leads and V1 – V6 are the precordial leads. An example of a 12-leads ECG is shown in Figure 2.4. The direction of the extremity leads, and the places of the chest electrodes are shown in Figure 2.3. The extremity leads are registered by four electrodes: one on the left arm, one on the right arm, one on the foot (usually on the left leg) and one neutral one (usually the right leg). The ECG is saved as an 8 leads ECG. The 8 channels are I, II, V1, V2, V3, V4, V5 and V6. The channels III, AVF, AVL and AVR can be calculated with the following equations:

$$III = II - I$$

$$AVR = -\frac{1}{2}(I + II)$$

$$AVL = \frac{1}{2}(I - III)$$

$$AVF = \frac{1}{2}(II + III)$$



(a) Precordial leads (b) Extremity leads

Figure 2.3: The 12-lead ECG recordings [21]

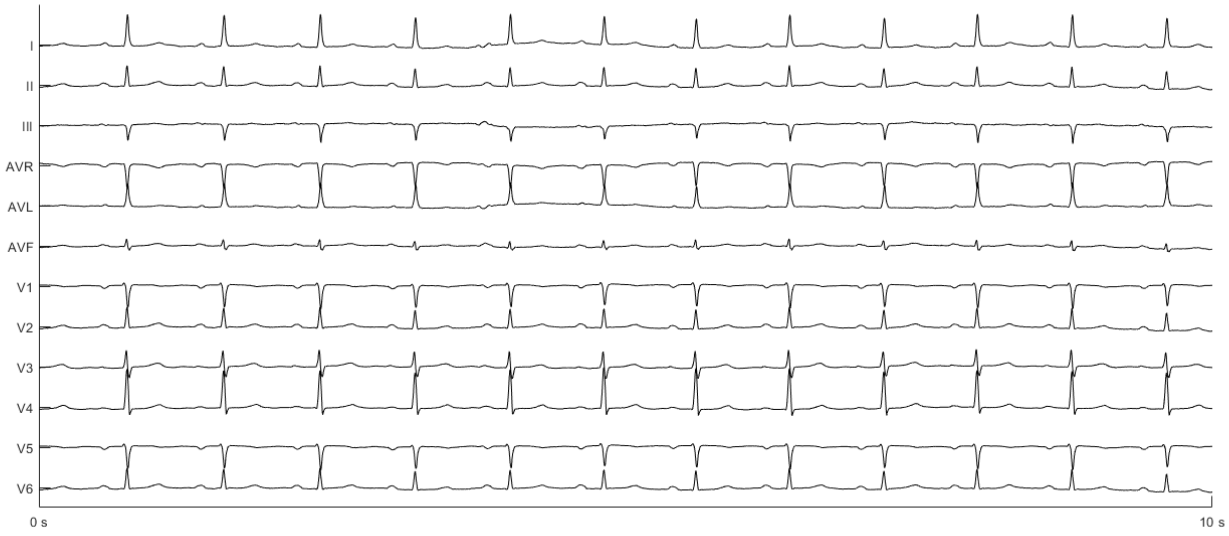


Figure 2.4: Example of an ECG. The different channels are shown over time (10 seconds). For every heartbeat, a P wave, a QRS complex and a T wave can be seen.

2.3.1 Clinical features in ECG

The ECG contains a lot of information about the electrical activity of the heart. It is a repetition of PQRST complexes (one complex for every heartbeat). A normal complex is shown in Figure 2.5. The depolarization of the atria is represented by the P wave, the QRS complex represents the depolarization of both the ventricles and the T wave represents the repolarization of the ventricles. The systole of a heartbeat is represented in the QRS complex. After the T wave, the diastole starts.

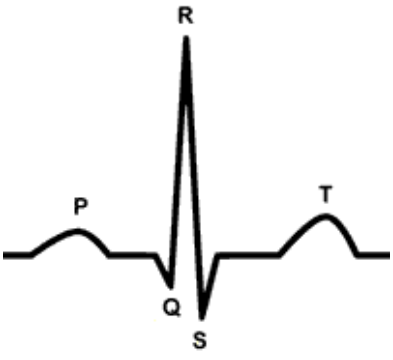


Figure 2.5: A normal PQRST complex [22]

A feature is a measurable property. In the ECG, different clinical features can be observed. The PR interval is the time between the

beginning of the P wave and the beginning of the QRS complex. The ST segment is the interval between the end of the QRS complex and the beginning of the T wave. The J-point is defined as the point where the QRS complex finishes and the ST segment begins. The QT time is defined as the time between the beginning of the QRS complex to the end of the T wave. This time may vary with the heart rate (HR), so the corrected QT interval (QTc) is the QT time divided by the square root of the RR interval. The RR interval is the time between two successive R peaks. A lot of conduction disorders can be seen in an abnormal ECG, like arrhythmia, bundle branch block (wide QRS complex), AV block (increased PR time) and QT syndromes (abnormal QT time). [23] An overview of normal conduction duration and the meaning of abnormalities is shown in Table 2.2.

Table 2.2: : Overview of features in the ECG, their normal duration and the meaning of abnormalities. [23]

Feature	**Duration (ms)**	**Pathology**
P wave	110	Normal P waves are upright in all leads except AVR and corresponds with the start of the depolarization in the sinus node. An abnormal P wave can be inverted in other leads. This indicates an ectopic atrial pacemaker. A biphasic P waves indicates hypertrophy of one of the atria.
PR interval	120 – 200	A decreased PR interval suggests that the AV node is bypassed. This can occur in the Wolf-Parkinson-White syndrome. An increased PR interval indicates an AV block. This block can be first to third order, where the first order indicates a delay of the electrical conduction through the AV node and a third is complete blockage of the electrical signals between the atria and the ventricles.
QRS complex	80 – 100	A wide QRS complex indicates a left or right bundle branch block or a ventricular rhythm. A low-amplitude QRS complex indicates a pericardial effusion or an infiltrative myocardial disease. High amplitudes in the precordial leads indicates hypertrophy of the left ventricle. A fragmented QRS is a marker for myocardial damage.
J-point		The existence of a J wave at the J-point is a characterization of hypothermia or hypercalcemia.
ST segment		The ST segment can be depressed or elevated e.g. in the case of a myocardial infarction or ischemia.
T wave	160	Inverted T waves (with respect to the QRS complex) are associated with myocardial ischemia.
QTc interval	Male: < 450 Female: < 460	An increased QTc interval is a risk for ventricular tachyarrhythmias and sudden death. It can also be related to genetic syndromes.

2.4 Data

In the Section 2.2, the patient journey is described. During this journey a lot of data is collected. All patient who underwent a TAVI in the Amsterdam UMC, location AMC between 2008 and 2016 are included in the TAVI database. Here, the data collected during screening, procedure and follow-up is saved. An overview of the data is shown in Table 2.3.

Table 2.3: Overview of the collected data of TAVI patients

Screening	**Procedure**	**Follow up**
Clinical data - Existence of symptoms - Laboratory data - Medical history - Risk factors Examinations - CTA - TTE - CAG - ECG	- Access route - Operators - Type of valve - Size of valve - Complications	Clinical data - Existence of symptoms - Laboratory data - Medical history - Risk factors - Mortality Examinations - TTE - ECG

All this data can be used to predict the outcome of patients after TAVI. In Figure 2.6 the amount of data is shown over time. The more data, the better the prediction is. For the clinician, the earlier a prediction can be made, the better. There is a mismatch between those. In this project data of different points in time is used, to learn what the added value of the data is, with a focus on ECG. The New York Heart Association (NYHA) [24] classification during screening and follow-up is used to describe the outcome of the patient. NYHA is a measure for dyspnea and can be classified between I and IV, where I is indicating the best condition and IV the worst. Also, the survival days after TAVI is used as an outcome. Mortality data from the municipal basic administration is used to calculate the survival days.

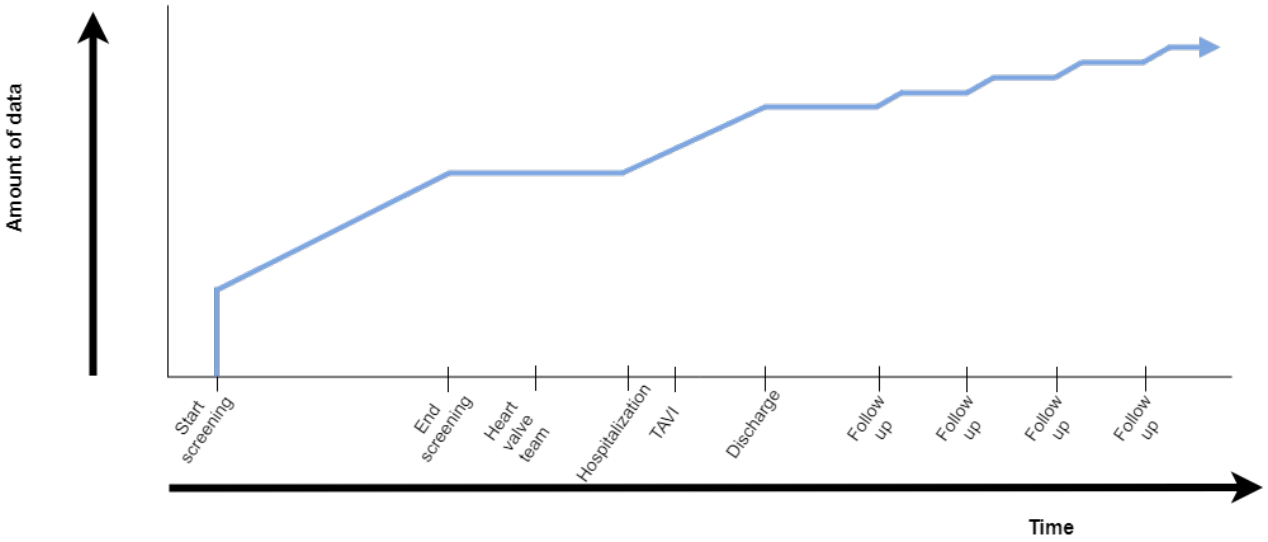


Figure 2.6: The growth of data over time per patient

A total of 5727 ECGs are available from 1003 patients. The ECGs have a sample frequency of either 250Hz or 500Hz. 8-leads ECGs are saved, which results in a matrix for every ECG with the size 2500x8 or 5000x8. For most of these ECGs, there are no labels (for example for atrium fibrillation) available. Also, NYHA scores or mortality data can be missing. An overview of the number of ECGs

per patient can be seen in Figure 2.7. If only one ECG is available, the patient either died before the TAVI or was denied for a TAVI. These patients are not included in this research.

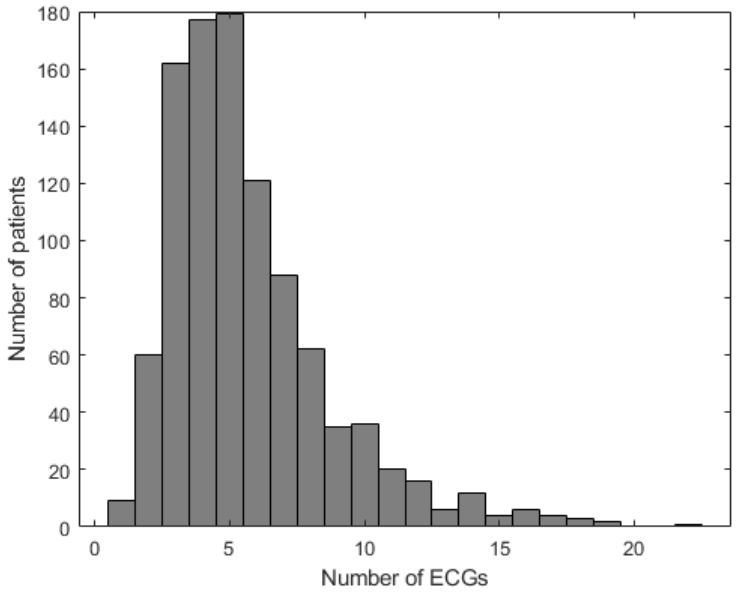


Figure 2.7: Number of patients combined with number of ECGs for each individual patient

Chapter 3

Machine learning

In this chapter, the focus is on ML. First, an introduction to ML is given. Then, the mechanism of a CNN is explained. Furthermore, different details for ML are described: training and testing, different loss functions, transfer learning and performance measures. After these sections, the mechanism of two visualization techniques are elaborated. Finally, the challenges in ML and learning from imbalanced data are discussed.

3.1 Introduction to machine learning

ML is a form of artificial intelligence. The goal of ML is to extract patterns out of data. [25] DL is a special form of ML. The last decades, ML and DL techniques are widely developed and applied, see Figure 3.1.

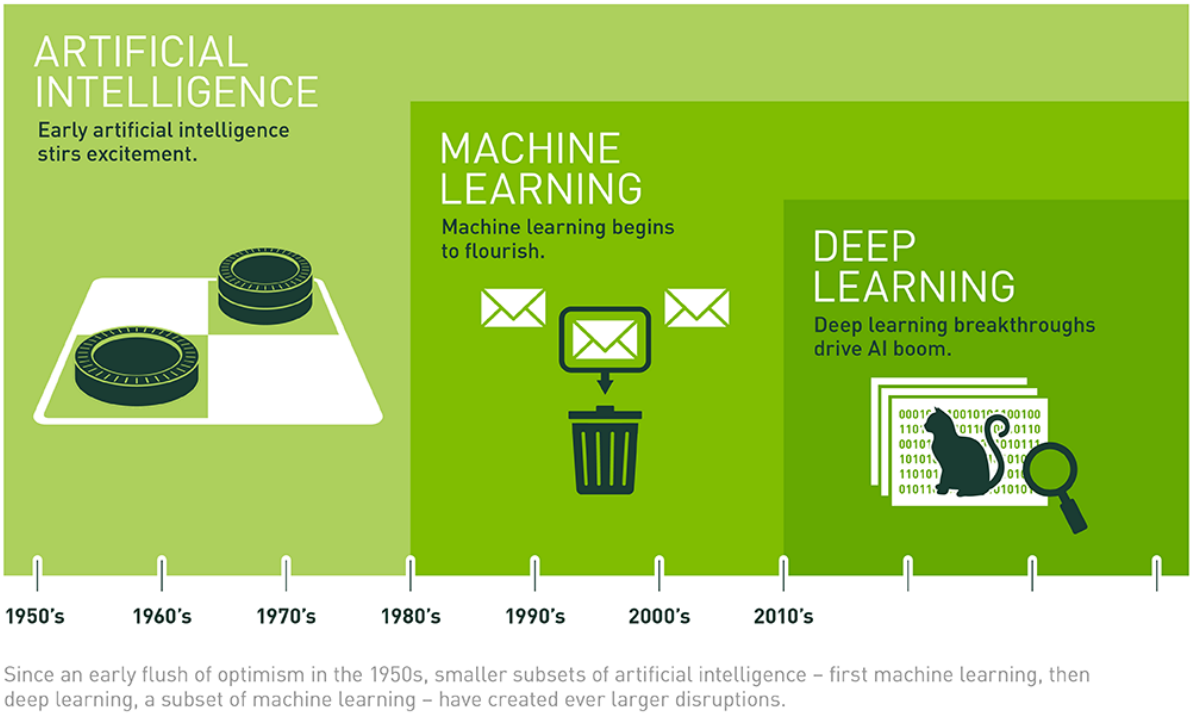


Figure 3.1: The development of artificial intelligence, ML and DL [26]

Three different problems can be tackled with ML techniques: [25]

1. Fitting data to a function or approximation: supervised learning
2. Exploring data and finding relations: unsupervised learning
3. Playing games with rewards and payoffs: reinforcement learning

An example to illustrate the first two different problems is the detection of spam email. Learning can be described as a process of 'using experience to gain expertise'. In the case of supervised learning, the experience contains significant information (in the form of labels). The goal is to learn from the training data, which contain labels, and predict these labels in the test data. In the example, the training data consist of emails which all have a label spam/not-spam. The test data consist of emails, for which the label is not known. In the case of unsupervised learning these labels are absent and there is no distinction between training and test data. The goal is to find patterns in the data itself. The patterns can be used to select the labels. In the example, the goal is to detect anomalies in the data. [25, 27]

Reinforcement learning is an intermediate learning setting. An illustrative example is a chess game. The training data contains all the positions during a game and the outcome: win or lose. The goal of such an algorithm is to learn which steps are necessary to eventually win the game. It is required to predict more than just the label. [27]

DL techniques are able to address more complex problems than ML techniques. They are better at processing natural data in their raw form. Another difference between ML and DL is the feature extraction. For ML techniques, the feature extraction needs to be performed before the application of ML techniques. DL techniques are able to extract features themselves. [25, 26, 28] The clinical features in the ECG are described in Section 2.3.1.

There exists a lot of different ML techniques. Examples of conventional techniques are logistic regression, support vector machine (SVM), random forest, hidden Markov models, clustering and filtering. Examples of DL techniques are neural network (NN), deep neural network, convolution neural network (CNN) and recurrent neural network. The models which are best suitable for sequential data, like ECG, are hidden Markov models, CNN and recurrent neural network. In this project a CNN is used in Chapter 4 and a NN and SVM are used in Chapter 5. The three pillars in DL are data, network architecture and network optimization. All three are evenly important and if one of them is corrupt, the problem will not be solved in a correct way.

3.2 Support vector machine

An SVM maps a vector of predictors into a higher dimensional plane. This mapping can be done by either linear or non-linear kernel functions. In the case of two classes, $\{+1\}$ and $\{-1\}$, the goal is to find a separating (linear) hyperplane, see Figure 3.2. This hyperplane is defined as

$$\mathbf{w}'\phi(\mathbf{x}) + b = 0$$

where $\mathbf{x}$ is the vector of predictors. This vector is mapped into a higher dimension feature space by a non-linear function ϕ, a vector $\mathbf{w}$ of weights and a bias b. The classification ($\{+1\}$ or $\{-1\}$) of an observation y_i is then based on the the classification function

$$f(y_i) = \text{sign}(\mathbf{w}'\phi(y_i) + b)$$

For the observation y_i the SVM gives a confidence value. This value lies between 0 and 1 and indicates how sure the SVM is about the observation being a fall event. For a confidence of 0, the SVM is sure it is not a fall action. For a confidence close to 1, the SVM is pretty sure the observation is a fall action.

In a binary classification problem, there exists infinite separation hyperplanes. The aim of training is to find the optimal hyperplane. This can be achieved by maximizing the distance of seperation between

the two planes $\mathbf{w}'\phi(\mathbf{x}) + b = -1$ and $\mathbf{w}'\phi(\mathbf{x}) + b = +1$, $r = \frac{2}{\|\mathbf{w}\|}$. This equivalent to minimizing the cost function

$$C(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2} + c\sum_{i=1}^{n}\xi_i = \frac{1}{2}\mathbf{w}'\mathbf{w} + c\sum_{i=1}^{n}\xi_i \tag{3.1}$$

with the constrains

$$y_i(\mathbf{w}'\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

where $c > 0$ and ξ_i are penalty parameters. c balances classification errors versus the complexity of the model. ξ_i is the slack-variable and controls how far a misclassified observation can lie on the wrong side. his is only the case when the training data cannot be completely separated by a linear hyperplane. [29]

The nonlinear mapping by the feature function ϕ is computed by kernels. These kernels are nonlinear and semi-positive. The cost function, Equation (3.1), is solved by

$$\min \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{n}\alpha_i$$

with the constraints

$$\textstyle\sum_{i=1}^{n} y_i\alpha_i = 0$$

$$0 \leq \alpha_i \leq C$$

where α_i are non-negative Lagrange multipliers and $K(.)$ is a kernel function. There are lots of different Kernel functions. In this study the kernel function are the linear,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i\mathbf{x}_j + \text{constant}$$

and the radial basis function

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{2\sigma^2}\right)$$

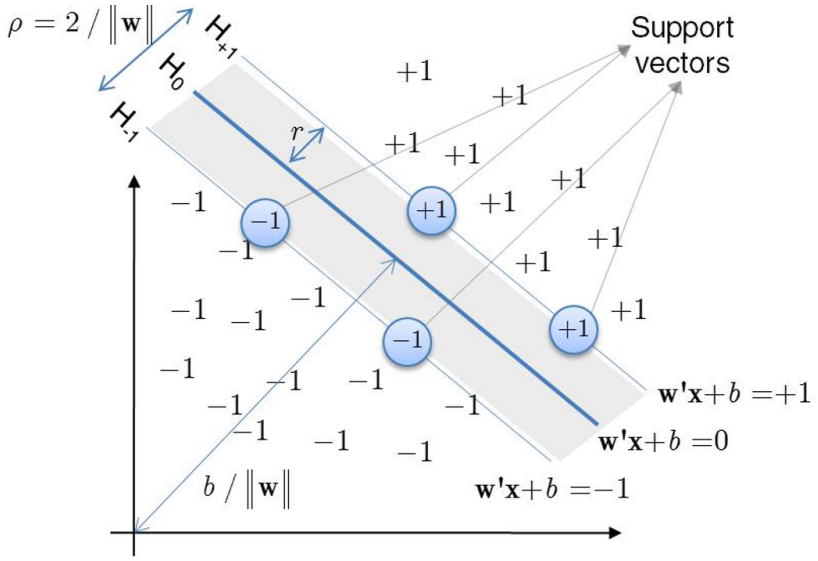


Figure 3.2: Schematic representation of an SVM in a two-dimensional space [29]

3.3 Convolutional neural network

A CNN is a type of DL method. The structure of a CNN is shown in Figure 3.3. The three main parts are the input, the feature learning and the classification. The input is an image or a signal. This input is a matrix filled with numbers. In the case of an image, these numbers represent the value of a pixel. In the case of a signal, these numbers represent the amplitude of the signal at each time point.

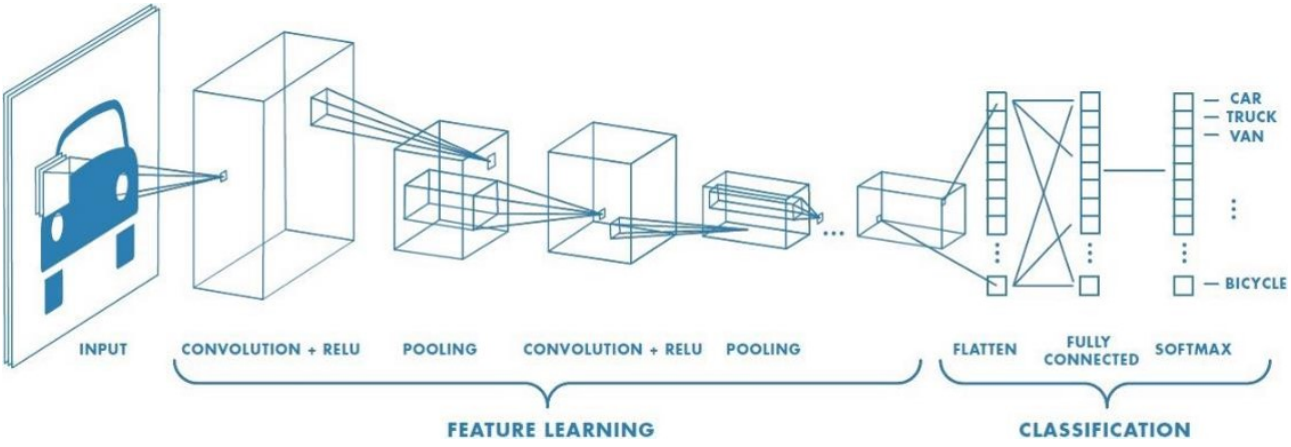


Figure 3.3: The structure of a CNN [30]

The feature learning, also called feature extraction, is done by a combination of convolutional layers and pooling layers. The convolutional layers consist of a number of filters. For each filter, the convolution between the input image and the kernel is calculated. This is shown in Figure 3.4. In this example, the kernel is a cross. For every 3x3 square, the convolution with the same kernel is calculated. This is done pixel wise, so in the example the equation of the convolution is (from left to right, top to bottom):

$$1 \cdot 1 + 0 \cdot 0 + 0 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 0 \cdot 0 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 = 4$$

If this $3 \cdot 3$ square looks like a cross, the result of the convolution is high (towards nine, because the size of the kernel is $3 \cdot 3 = 9$). Vice versa, if the square doesn't look like a cross, the result is low (towards zero). All these outcomes can be combined into a new matrix, which is the output for that filter.

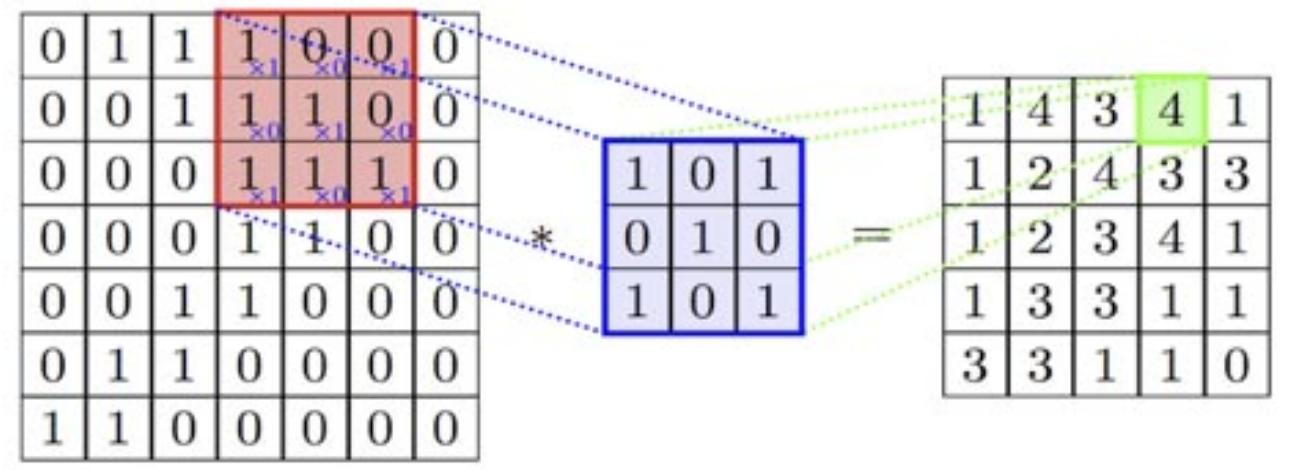
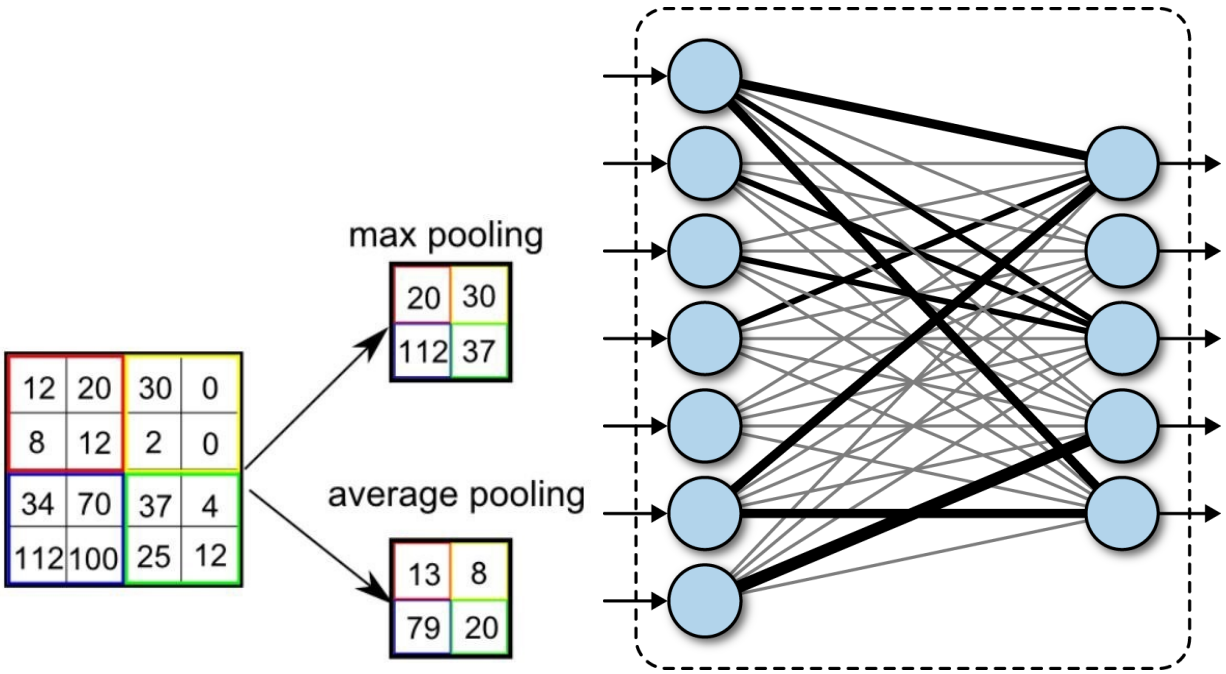


Figure 3.4: The working of a convolution in a convolutional layer. On the left, the input is shown. In the middle, the kernel (blue) is shown and on the right, the result is shown. In this example, the kernel represent a cross, because the ones are in all the corners and in the middle and the spaces in between are filled with zeros. [31]

The output of the whole convolutional layer is a 3D matrix where the number of filters equals the depth of the output and the height and the width equals the height and the width of one convolution outcome. This 3D matrix is used as input for the pooling layer. The pooling layer decreases the size of the matrix. This can be done by different pooling techniques: maximum pooling, average pooling or L2-norm pooling. The mechanism of pooling layers is shown in Figure 3.5a. In this work, maximum pooling is used. The maximum values of each 2x2 square are saved and the combination of all these outcomes is again the input for the next layer. The combination of convolutional layers and a pooling layer can be seen as one block, which is repeated throughout the network.



(a) The mechanism of pooling layers, maximum pooling and average pooling [32]

(b) A fully connected layer. The thicker lines represent the higher weights [33]

Figure 3.5: Explanation of two layers in a CNN.

With the convolutional layers and the pooling layers, the features are extracted. After these feature extraction part, the classification has to be done. The first step is to flatten the features into a vector, the feature vector. This vector can be used as input for one or more fully connected layers. As the name states, the neurons in the fully connected layer are all connected to the neurons in the flatten layer, see Figure 3.5b. During training, the weights of all these connections are calculated. The higher the weight, the more important the feature is for the classification. After the fully connected layer, a softmax layer finishes the CNN. This layer translates the output of the fully connected layer to the classification of the outcomes.

3.3.1 Training and testing

Above, the mechanism of a CNN is explained. To teach the network, it needs to see data. In this work, supervised learning is used. The total dataset is split in training data and test data. The network uses the training data to adjust the weights and to learn, the test data is used to measure the performance of the network.

To train, the network knows the label for each data point. So, each data point can be given to the network as input and the predictive outcome is calculated by the network, $\hat{y}$. The actual label, y, is known. With a loss function the loss over all these predicted outcomes and actual label can be calculated. The different loss functions are given in the next section. For the network to learn, this loss must be minimized. This is done by the stochastic gradient descent (SGD) method. The equation

of SGD is

$$\theta \longleftarrow \theta - \eta \nabla \mathcal{L}(\theta, x_i, y_i)$$

where θ are the weights, η is the learning rate and $\mathcal{L}(\theta, x_i, y_i)$ is the loss function with respect to θ from training data $\{x_i, y_i\}$. In every iteration the weights are adjusted. If the network sees the training data multiple times, the weights can be adjusted to minimize the loss.

After training, the test data is seen by the network. In this case, the network only sees the inputs and calculates the predicted outcome. A confusion matrix can be made and the performance of the network can be calculated. This is explained in Section 3.3.4.

3.3.2 Loss functions

As explained above, the loss function is used during training to measure the inconsistency between predicted outcome $\hat{y}$ and actual label y. In Table 3.1 different loss functions are stated. A different loss function means a different model, because the new weights depend on the loss.

Table 3.1: Overview of different loss functions. θ is the set of weights, $\hat{y}$ is the predicted outcome and y is the actual label.

Name	**Equation**
Mean squared error	$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$
Mean squared logarithmic error	$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\log\left(y^{(i)}+1\right) - \log\left(\hat{y}^{(i)}+1\right)\right)^2$
Mean absolute error	$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left\lvert y^{(i)} - \hat{y}^{(i)}\right\rvert^2$
Mean absolute percentage error	$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y^{(i)} - \hat{y}^{(i)}}{y^{(i)}}\right\rvert^2$
Kullback-Leibler (KL) divergence (also known as relative entropy)	$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} \cdot \log\left(y^{(i)}\right)\right) - \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} \cdot \log\left(\hat{y}^{(i)}\right)\right)$
Cross-entropy	$\mathcal{L}(\theta) = y^{(i)} \cdot \log\left(\hat{y}^{(i)}\right) - \left(1 - y^{(i)}\right)\log\left(1 - \hat{y}^{(i)}\right)$
Weighted cross-entropy (WCE)	$\mathcal{L}(\theta) = w_0 y^{(i)} \cdot \log\left(\hat{y}^{(i)}\right) - w_1\left(1 - y^{(i)}\right)\log\left(1 - \hat{y}^{(i)}\right)$
Negative log-likelihood (NLL)	$\mathcal{L}(\theta) = -\sum_{i=0}^{\lvert\theta\rvert}\log P\left(Y = y^{(i)} \left\lvert x^{(i)}, \theta\right.\right)$
Poisson	$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}^{(i)} - y^{(i)} \cdot \log\left(\hat{y}^{(i)}\right)\right)$
Cosine proximity	$\mathcal{L}(\theta) = \frac{\sum_{i=1}^{n} y^{(i)} \cdot \hat{y}^{(i)}}{\sqrt{\sum_{i=1}^{n}\left(y^{(i)}\right)^2} \cdot \sqrt{\sum_{i=1}^{n}\left(\hat{y}^{(i)}\right)^2}}$
Hinge	$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y^{(i)} \cdot \hat{y}^{(i)}\right)$
Squared hinge	$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\max\left(0, 1 - y^{(i)} \cdot \hat{y}^{(i)}\right)\right)^2$

3.3.3 Transfer learning

Transfer learning is a technique where a pre-trained model is used to predict another outcome. The theory is that the feature extraction is the same, but the classification part is different. So, the convolutional layers and pooling layers are used from the pre-trained model and new classification layers are added to perform new classification. One of the advantages is a lower computational time, because only the weights of the classification layers must be adjusted. Another advantage is that the network is trained on a lot of data and, thus, the feature extraction is good.

3.3.4 Perfomance measures

Performance measures are used to validate the model. A confusion matrix can be made with the outcome of the model from the test data. Such a confusion matrix is shown in Figure 3.6. In this work, there is a lot of imbalanced data. Imbalanced data means that the amount of data sample in each class is not in balance. This will be further explained in Section 3.5. Accuracy is not a good performance measure in to compare the performance of different models that case, because if the model classifies everything as the most common label, then the accuracy of the model is pretty good. Instead of accuracy, the F1-score is used as the performance measure, together with specificity and sensitivity. The meaning of F1-score can be shown by the example of a coin toss. Say a coin is tossed 100 times. The expected true positive (TP), false positive (FP), false negative (FN) and true negative (TN) are all 25. In that case the sensitivity is 0.5, the specificity is 0.5 and also the precision is 0.5. Then the F1-score is 0.5 as well. So, if the F1-score is higher than 0.5, the model is better than a coin toss. Since the F1-score calculates how good a TP can be predicted, the score can be low while the performance is good. This is the case with imbalanced data where the negative labels are more frequently present than the positive labels. Because this is not always the case in this study, the F1 score is used to compare different models and the accuracy is calculated on the final models. The equation for these performance measures are given below in Table 3.2.

Table 3.2: Performance measures and their meaning

Performance measure	**Equation**	**Meaning**
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	How well can the network predict the truth (0 – 100 %)
Sensitivity (TPR)	$\frac{TP}{TP+FN}$	How well does the network predict the TP (0 – 1)
Specificity (TNR)	$\frac{TN}{FP+TN}$	How well does the network predict the TN (0 – 1)
Precision (PPV)	$\frac{TP}{TP+FP}$	How much positive predictions are relevant
F1-score	$\frac{2 \cdot PPV \cdot TPR}{PPV+TPR}$	Combination of precision and sensitivity

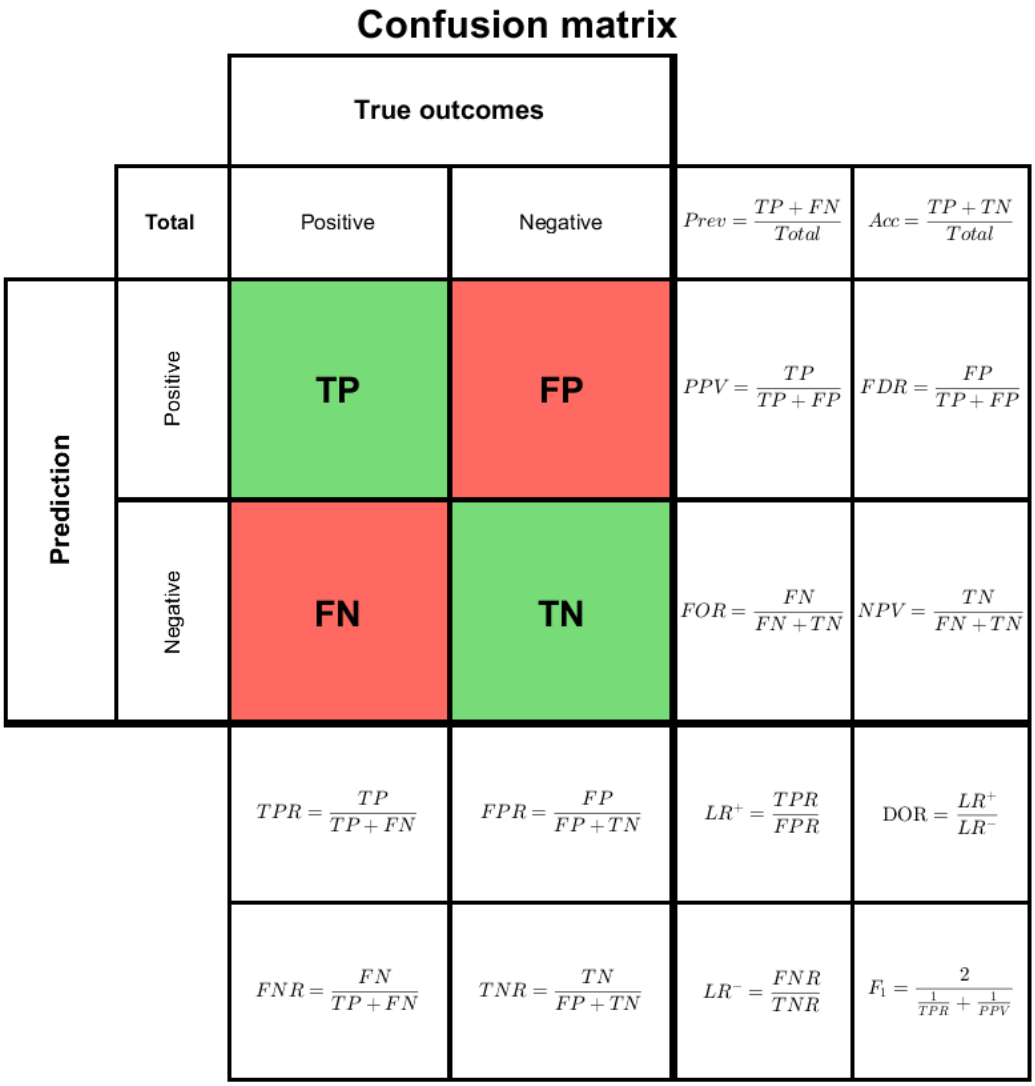


Figure 3.6: Confusion matrix. TP = true positive, FP = false positive, FN = false negative, TN = true negative. Prev = prevalence, Acc = accuracy, PPV = positive predictive value, FDR = false discovery rate, FOR = false omission rate, NPV = negative predictive value, TPR = true positive rate, FPR = false positive rate, FNR = false negative rate, TNR = true negative rate, LR+ = positive likelihood, LR- = negative likelihood, DOR = diagnostic odds ratio, F1 = F1-score

3.4 Visualization

A CNN is a black box, which means that it is not known what happens inside. To break open this black box, visualization techniques can be used. The most well-known example is the wolf vs husky problem. Researchers at the University of Washington tried to build a classifier to detect whether there is a wolf in a picture or a husky. They achieved great results: 90% accuracy. When they went deeper in the network, it turned out that the network was using snow in a picture to classify instead of the animal. They build a snow detector.[34] This is an example which shows why it is important to understand what a network uses for its classification. Another reason is the clinical practice. If it can be explained to a doctor why a network predicts an outcome, it is easier for a doctor to use that results.
There exist different techniques to make visualizations of the network's behavior. They are mostly developed for images, instead of signals. This must be remembered when interpreting the visualizations. In this work, the focus is on two techniques: filter visualization by maximum activation and occlusion maps. These will be explained in the next section.

3.4.1 Filter visualization

The aim of filter visualization is to visualize what maximizes a certain filter in a certain layer, the feature. As shown in Figure 3.7, the hypothesis is that the deeper the filter lays in the network, the more complex the feature is. This visualization can be done by a technique called activation maximization. It is an iterative process, where a noisy input is being adjusted until the output of the specific filter is maximized. The input can be adjusted with a gradient for each iteration. This technique is mostly applied with images as input, as shown in Figure 3.7. For a signal as input, the interpretation can be less obvious and has to be thought trough well.

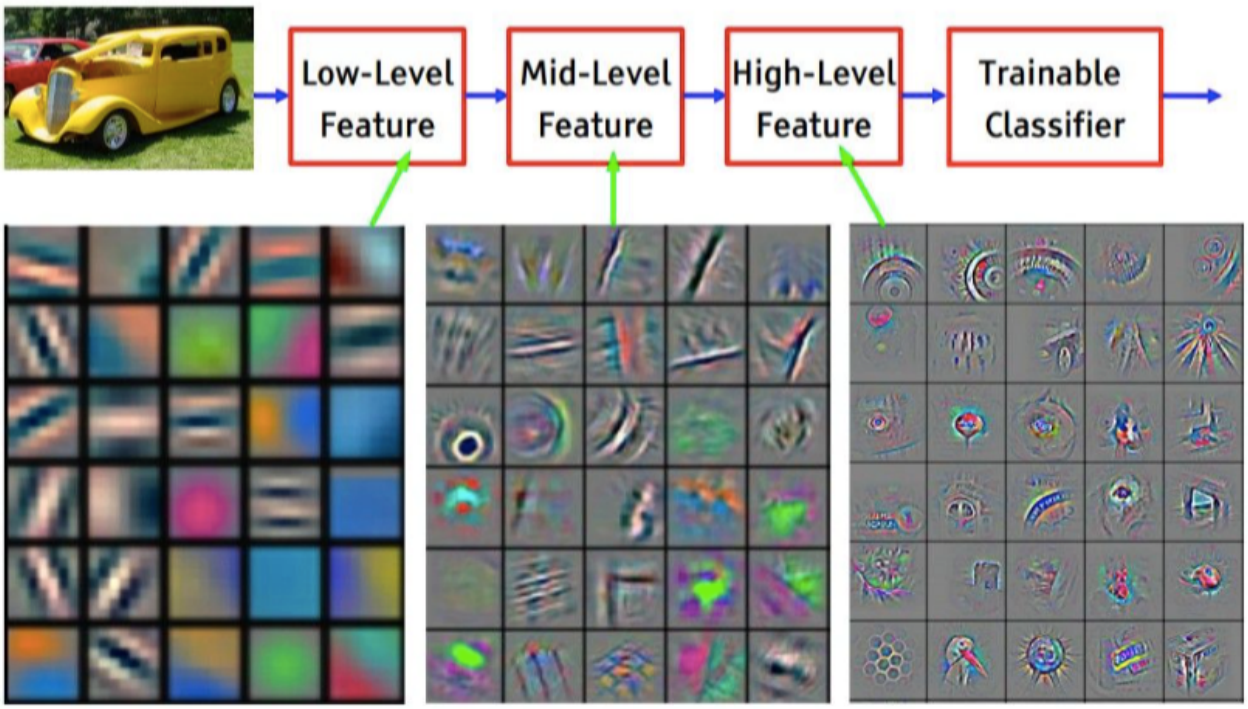


Figure 3.7: Features in different layers in a CNN [35]

3.4.2 Occlusion maps

The aim of occlusion maps is to visualize which part of the input is important for the network to calculate the outcome. This can be done by deleting a part of the input and analyze what the network predicts. If the certainty of the prediction is different from the prediction with the whole input, the

deleted part is important for the network. To make the occlusion maps, different ways of deleting a part of the input can be chosen. The easiest way is to set a part of the signal to zero. Another approach can be to delete certain frequencies from the signal. In this report, the choice is made to set a part of the signal to zero, because it can be more easily visualized and explained to the cardiologist.

3.5 Learning from imbalanced data

The ideal data set exists of the same number of data points for each label. In real problems, this is almost never the case. Especially in health care, the number of data points for each label will be very imbalanced. For example, 1% of the labels is positive and the rest of the labels is negative. In this case the model can easily learn to classify everything as negative and reach an accuracy of 99%. This is not the desired result, because the 1% is the interesting part and the network is made to help classify this 1%. This is for example the case in the prediction of mortality in a certain population. If it known that 1% of this population will die within 10 years, the doctor is interested in the population with the highest changes to be in this 1% so the treatment can be adjusted, or the frequency of visits can be higher. The goal is to create a model which can classify correctly. From this example two things can be learned: accuracy is not a good performance measure for imbalanced data and some techniques are needed to handle imbalanced data sets. Two techniques are explained in the next sections, the performance measures are described in Section 3.3.4.

3.5.1 Resampling techniques

The goal of resampling is that there are equal number of data points for each label. This can be done by two methods: down sampling and up sampling. With down sampling the number of data points in the bigger group are reduced to the number of data samples in the smaller group. With this method, data is removed and not used for training. Up sampling means that with data augmentation the number of data samples from the minor group are increased. Different techniques for data augmentation exist, but these are not used in this work.

3.5.2 Class weighted learning

Another technique to handle an imbalanced data set is class weighted learning, also called cost sensitive learning. This technique influences the loss function instead of resampling the data. By giving a higher weight to the minority class and a lower weight to the majority class, the classifier is made aware of the imbalance. This can be easily illustrated by the cross-entropy loss function. The equation for this loss function is

$$\mathcal{L}(\theta) = y^{(i)} \cdot \log\left(\hat{y}^{(i)}\right) - \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right)$$

To make this function the WCE loss function, weights w_0 and w_1 are added. Then, the equation is

$$\mathcal{L}(\theta) = w_0 y^{(i)} \cdot \log\left(\hat{y}^{(i)}\right) - w_1 \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right)$$

This weighted learning can also be applied to other loss functions, for example the KL loss function.

3.6 Machine learning challenges

There are a lot of challenges when applying ML to a real-life problem. Some of the challenges are stated here and tricks are given to prevent models from related problems. The three pillars to create a ML model are data, architecture and optimization. These three pillars can be handled individually, but they all need to be correct to tackle the problem. The challenges in the data pillar are dealing with unstable data and the labels. Unstable data can be a database with a lot of missing values. The data points with missing values can either be deleted from the database or techniques like substitution or imputation can be applied. [36] In the architecture pillar the challenge is to choose the right ML technique and in case of DL, the right network architecture. After this choice is made, the challenge

is to find the right depth of the network. If the network is too deep, the information in the features will be lost. If the network is too shallow, the network cannot get all the information out of the input. The depth of the network correlates with the dimensionality of the input data. The final pillar, optimization, has mainly the challenge to prevent the network from under- or overfitting. If the network is under fitting, the network has not learned everything it could have learned. In the case of overfitting, the network learns too much on the training set and the performance of the network is not good on the test set. In other words, the network doesn't generalize. Another challenge is an unpredictable future. It may be the case that the used data doesn't contain the information needed for the prediction.

Chapter 4

The use of an electrocardiogram for prediction

In this chapter, the use of a CNN is investigated for prediction outcomes of TAVI patients. First, the data used in this chapter is elucidated. Then the method to find an optimal network for predicting outcomes is described. The results show how the different settings of the network influence its performance. This chapter ends with a discussion, where the results are interpreted, the limitations are elaborated and the connection with the current literature is shown.

4.1 Introduction and motivation

AoS is one of the most common heart diseases. [15] As described earlier, TAVI is a relative new treatment for patient with AoS and high surgical risk. [18, 19] Even though strict patient selection criteria and planning tools are developed [37, 38, 39], there are patients with limited benefit of a TAVI. [40, 41] Current prediction models, to choose between SAVR and TAVI, mainly use clinical data and traditional linear regression models. Instead of using clinical data, the ECG signals might have predictive value for TAVI outcomes. This hypothesis is also stated in the Introduction of this work. A previous study used ECG to predict rhythm disorder post-procedural. [8] Alternatively for these models, other ML approaches can be used. These methods have shown superior predictive value in various clinical areas. [42]

In this chapter, the post-procedural ECG and CNNs are used to predict the outcome of TAVI patients. The post-procedural ECG is used here, because it is closer to the future which is predicted. The set-up is shown in Figure 4.2. The outcomes are either mortality or improvement of dyspnea. In Chapter 3, it is explained that there are a lot of parameters which can be changed to increase the performance of the network. Here, a trial and error method to find these parameters is proposed and executed. A CNN is trained to detect AF, set-up shown in Figure 4.1, and transfer learning is applied to learn the features which are important to predict the outcome of TAVI. The features will be illustrated using feature visualization techniques.

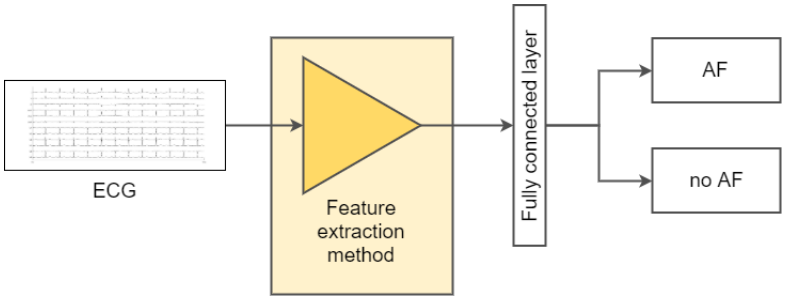


Figure 4.1: Set-up of the method used in experiment 1 of this chapter. An 8-leads ECG is used to detect AF or no AF. Here, the feature extraction method which is used is a CNN.

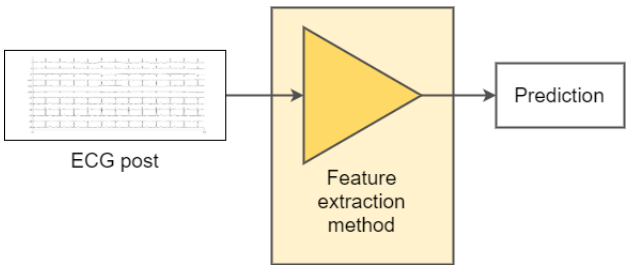


Figure 4.2: Set-up of the method used in experiment 3 and 4 of this chapter. An 8-leads ECG is used to predict the outcome. Here, the feature extraction method which is used is CNN, either trained in experiment 1 or trained from scratch.

4.2 Method

4.2.1 Study population

The outcomes used in this study are mortality and improvement of dyspnea. Dyspnea is measured using the NYHA functional score [24] (classified between I-IV, where I is indicating the best condition and IV the worst). Since patients with NYHA class III and IV are hindered in daily life, these two classes are split out in Table 4.1. Mortality is defined as patients who died within one year after the procedure. Patients with missing data are excluded. Missing data can either be NYHA scores at the baseline or at 60 days follow-up or mortality data. The presence of permanent AF is known in 947 ECGs (220 AF, 727 no AF) from 419 patients. There can be multiple ECGs for individual patients in the used dataset. The labels AF or no AF are given by an electrophysiologist. The NYHA score, at baseline and 60-days follow up, is known for 700 patients (547 improved, 153 did not improve). Mortality, within one year after the procedure, is known for 932 patients (785 survivors, 147 non-survivors). The flowchart in Figure 4.3 gives the overview of used and missing data for each experiment.

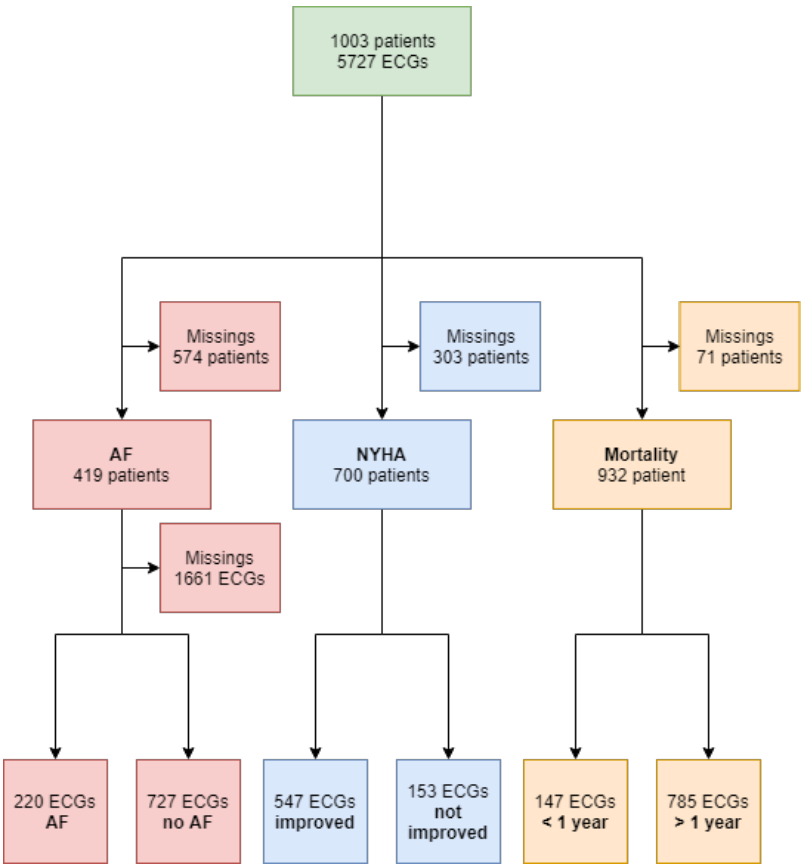


Figure 4.3: Flowchart of the data. The AF data is used for the first experiment. In the database, 419 patients have one or more ECGs with an AF label. Of these patients (2588 ECGs) have 927 ECGs a label. For the NYHA data, the NYHA score before and after TAVI is known for 700 patients. The mortality is known for 932 patients. The NYHA data and mortality data is used in experiment 3 and 4.

Table 4.1: Demographic and clinical characteristics per data group

Parameters *Patients*	AF data ($n = 419$)	dyspnea data ($n = 700$)	Mortality data ($n = 932$)
Age (years)	82 (77 – 85)	81.5 (77 – 85)	82 (77 – 85)
Female gender	252 (60.1 %)	400 (57.1 %)	532 (57.1 %)
BMI (kg/m^2)	27.1 (24.4–30.7)	27.1 (24.4–30.7)	26.9 (24.5–30.7)
NYHA class III or IV	291 (69.4 %)	468 (66.9 %)	644 (69.1 %)
Cardiovascular medical history			
Hypertension	331 (79.0 %)	575 (82.1 %)	770 (82.6 %)
AF	150 (35.8 %)	275 (39.3 %)	372 (39.9 %)
Valve surgery	21 (5.0 %)	27 (3.9 %)	41 (4.4 %)
CABG	57 (13.6%)	92 (13.1 %)	126 (13.5 %)
PCI	134 (32.0 %)	192 (27.4 %)	255 (27.4 %)
Pacemaker	49 (11.7 %)	84 (12.0 %)	113 (12.1 %)
Stroke	36 (8.6 %)	66 (9.4 %)	98 (10.5 %)
PAD	110 (26.3 %)	190 (27.1 %)	264 (28.3 %)
Other medical history			
COPD	150 (35.8 %)	216 (30.9 %)	304 (32.6 %)
Diabetes mellitus	129 (30.8 %)	215 (30.7 %)	292 (31.3 %)
eGFR < 60 ml/min/1,73 m^2	203 (48.4 %)	338 (48.3 %)	461 (49.5 %)
Riskscores			
STS	4.5 (3.2 – 7.1)	4.4 (3.2 – 7.1)	4.6 (3.2 – 7.1)
EuroSCORE I	14.6 (10.0 – 22.2)	14.3 (10.0 – 22.2)	14.4 (10.0 – 22.2)
EuroSCORE II	4.2 (2.5 – 7.5)	4.0 (2.5 – 7.5)	4.2 (2.5 – 7.5)
TTE			
Poor LVF	34 (8.1 %)	53 (7.6 %)	74 (7.9 %)
Poor RVF	3 (0.7 %)	8 (1.1 %)	12 (1.3 %)
SPAP > 55 mmHg	155 (37.0 %)	258 (36.9 %)	341 (36.6 %)

BMI: body mass index, AF: atrial fibrillation, CABG: coronary artery bypass grafting, PCI: percutaneous coronary intervention, PAD: peripheral artery disease, COPD: chronic obstructive pulmonary disease, eGFR: estimated glomerular filtration rate, STS: society of thoracic surgeons, EuroSCORE: European System for Cardiac Operative Risk Evaluation, TTE: transthoracic echocardiography, LVF: left ventricle function, RVF: right ventricle function, SPAP: systolic pulmonary artery pressure. The definitions can be found in Appendix A. In case of categorical data, the count and the percentage of the total populations are given. In case of numerical data, the median is given with the 25% and 75% percentiles.

4.2.2 Data

The ECGs are 8-leads (I, II, V1-V6) and recorded for 10 seconds using a sample frequency of 500 Hz, resulting in 8x5000 samples. To reduce computational costs, the signal is resampled to 8x625 samples. In Figure 4.4 can be seen that the same frequencies are present in the original signals as in the resampled signals. There are no filters applied, the resampled signal is used as an input for the CNN.

Before training the model, the data is split into a train set and a test set, respectively 80% and 20%. In case of the AF data, there are patients with multiple ECGs. This has been taken into account in splitting of the data. All the ECGs of a patient can be either in the train set or the test set, but not in both. The train set is used to train the network. The test set is used to analyze the performance of the network.

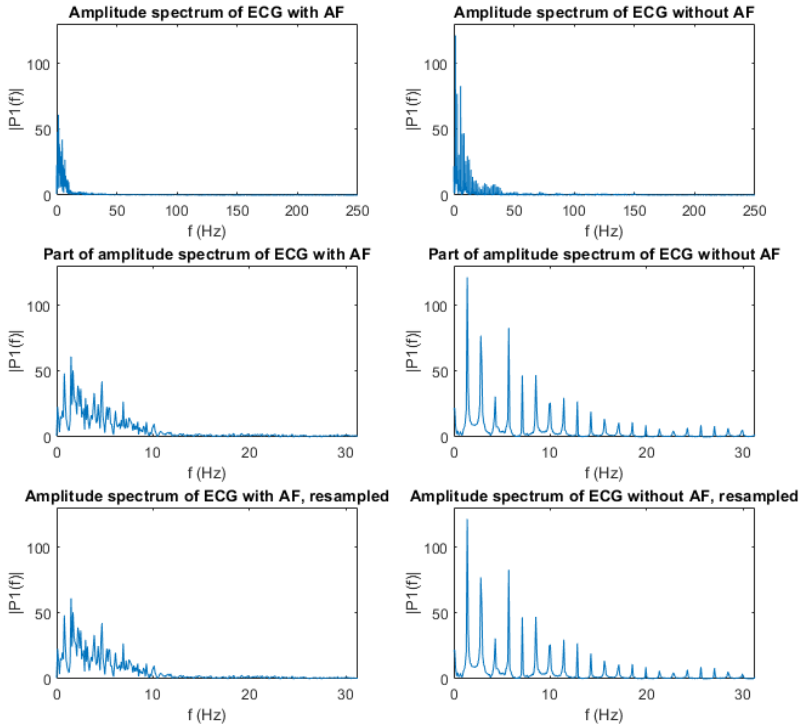


Figure 4.4: Amplitude spectrum of two original signals with the amplitude spectrum of their resampled signals as a comparison. It can be seen that the frequencies which are present in the original signals are also present in the resampled signals.

The measures used are explained in the Section 3.3.4. The meaning of the performance measures in this chapter is explained in Table 4.2. The measures used are calculated on the test set. The accuracy is only calculated on the final models, because it is not a good measure to compare two different models.

Table 4.2: Equations and meaning of performance measures used in this chapter.

Performance measure	**Equation**	**Meaning**
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	How well can the network predict the truth (0 – 100 %)
Sensitivity	$\frac{TP}{TP+FN}$	How well does the network predict AF (0 – 1)
Specificity	$\frac{TN}{FP+TN}$	How well does the network predict no AF (0 – 1)
Precision (PPV)	$\frac{TP}{TP+FP}$	How much positive predictions are relevant
F1-score	$\frac{2 \cdot PPV \cdot TPR}{PPV+TPR}$	Combination of precision and sensitivity

4.2.3 Experiments

Experiment 1: CNN architecture for AF prediction

Experiment 1 consists of experiment 1a and 1b, which both contain three trials. In this experiment the architecture of the CNN is found (experiment 1a) and the architecture is optimized (experiment 1b). The architecture of the CNN consists of an input layer, followed by n clusters of 2 convolutional layers with each a rectified linear unit and these layers are followed by a maximal pooling layer. The value of n is determined in the first trial of experiment 1a. The architecture ends with a fully connected layer with 20% dropout and a Softmax classification layer. The global architecture is shown in Figure 4.5. The initial loss function is a WCE. The initial chosen parameters of the CNN are shown in Table 4.3. The most important parameters are the learning rate and the momentum. The values of these will be tested in the second and third trial of experiment 1b. The learning rate is the rate which the optimizer uses for the backpropagation. The momentum is a tool which can be added to the SGD method. It can help to find the global minimum instead of a local minimum. The number of epochs is initially chosen at 150.

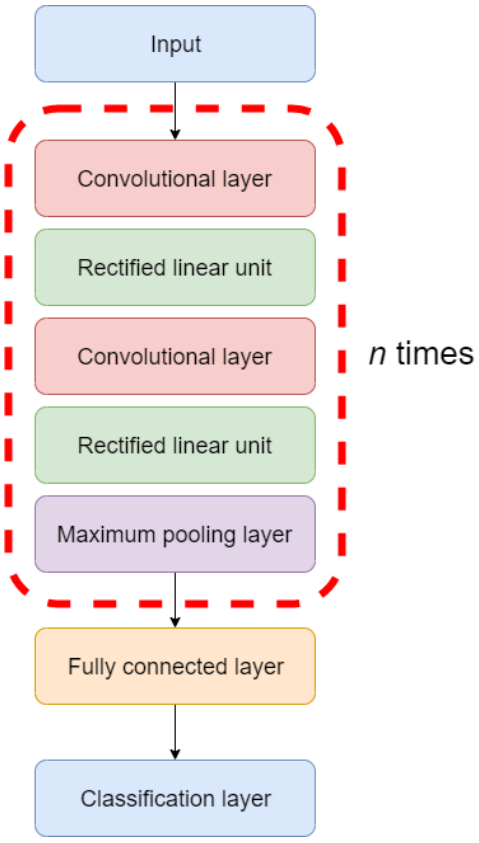


Figure 4.5: Architecture of CNN. In experiment 1a the value of n is found.

Table 4.3: Initial values of the parameters of the CNN.

Parameter	**Initial value**
Kernel initializer, convolutional layer	Gaussian distribution with $\mu = 0$, $\sigma = 0.01$
Kernel size maximum pooling layer	(2,1)
Number of epochs	150
Batch size	32
Optimizer	SGD
Learning rate	0.0001
Learning rate decay	0.000001 (1e-6)
L2-regularization	0
Dropout	20%
Loss function	WCE

During experiment 1a and 1b, the architecture and the parameters of the CNN are found by a trial and error method. For both experiments different trials are done with ten repetitions and for each repetition the F1-score is determined after 150 epochs. The mean and the standard deviation of the ten repetitions are calculated and used to choose the parameter for every next trial.

Experiment 1a During the first experiment the global architecture of the CNN is determined. The depth of the network, the number of fully connected layers and the kernel parameters are found in three different trials. During the first trial, different depths are tried (4 till 22 convolutional layers). The second trial is focused on the neurons and the number of fully connected layers. Different number of neurons and one or two layers are tried. In the final trial of experiment 1a different kernel sizes and different number of filters are tried. The set of parameters which are tried during this experiment can be found in Table 4.4 and Table 4.5.

Table 4.4: Parameters of the second trial of experiment 1a

Exp.	**FC 1**	**FC 1**
1	128	None
2	256	None
3	128	64
4	128	32
5	256	128
6	256	64

Table 4.5: Parameters of the third trial of experiment 1a

Exp.	**Kernel size**	**Filters layer 1,2**	**Filters layer 3 − 4**	**Filters layer 5 − 12**
1	(5,1)	128	64	32
2	(5,1)	128	128	64
3	(5,1)	128	128	32
4	(5,1)	128	32	32
5	(10,1)	128	64	32
6	(10,1)	128	128	64
7	(15,1)	128	64	32
8	(15,1)	128	128	64

Experiment 1b When the architecture of the network is found, the optimization of the network is analyzed. In the first trial, different loss functions are tried. The equation of the different loss functions can be found in Section 3.3.2. Also, the addition and absence of class weights in the cross-entropy loss function is tested. All loss functions are shown in Table 4.6.

Table 4.6: Overview of the loss functions which are tested in the first trial of experiment 1b

Exp.	Loss function
1	Mean squared error
2	Mean squared logarithmic error
3	Mean absolute error
4	Mean absolute percentage error
5	KL divergence
6	Cross entropy
7	WCE
8	NLL
9	Poisson
10	Cosine proximity
11	Hinge
12	Squared hinge

During the second trial, the value of the learning rate is tested. Different values are tested, and again the mean and the standard deviation of the F1-score of the ten repetitions is used to select the learning rate. The values of the learning rates that are tried are in the range between $1e^6$ and 0.05. In the third trial, the parameters of the optimizer are further analyzed. In the SGD optimizer, a (Nesterov) momentum can be added to accelerate the optimizing process. Different values for both normal momentums and Nesterov momentums are tried. For the final model, the model with the highest F1-score is selected.

Experiment 2: feature visualization

In experiment 1, the parameters for the CNN are found. Now, it is interesting to visualize what the CNN has learned, because in that way the working of the CNN can be explained. To do so, different feature visualization techniques are used. First, the filters are visualized. The difference between filters in the lower layers and the filters in the higher layers are shown. Filters can be visualized by the activation maximization method. Another technique to visualize what the network learns, is the use of occlusion maps. This technique shows which input is most important according to the network. The techniques are explained in Section 3.4. These maps are shown for four examples: a true positive, a false positive, a true negative and a false negative.

Experiment 3 and 4: predict mortality and improvement of dyspnea

The goal of this project is to predict outcomes of TAVI patients. In experiments 3 and 4, this is done. The network of experiment 1 is trained to predict both outcomes using transfer learning. To compare the results of transfer learning, the network is also trained from scratch to predict the outcomes.

The calculations were done on a 64-bit Windows desktop in Python with Keras (Tensorflow backend). The GPU, Quadro P6000, used for this research was donated by the NVIDIA Corporation.

4.3 Results

4.3.1 Experiment 1

For each trial in this experiment the mean and the standard deviation of the F1-scores of the test set are shown in Figures below. Based on these results, the choice is made which parameters are used for the next experiment.

Experiment 1a

In experiment 1a, the architecture of the network is determined. The first trial is to define the number of layers in the network. In Figure 4.6, the results of this trial are shown. An optimum can be seen with ten convolutional layers. Lower number of layers show a decrease of the mean of the F1-scores, but an increase of the standard deviation. The same effect can be seen if the number of layers is greater than ten. In the next trials, the CNN consists of ten convolutional layers.

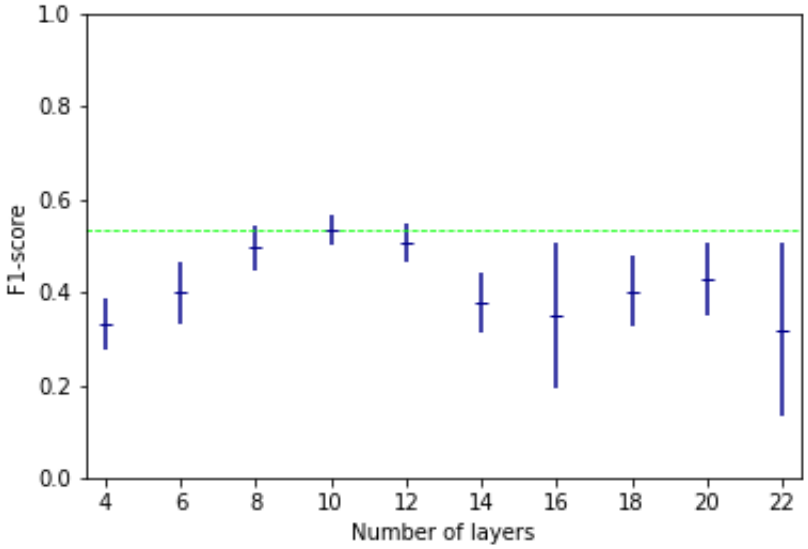


Figure 4.6: Results of the first trial of experiment 1a. The red dotted line indicated the highest mean F1-score. In this experiment, the highest F1-score corresponds to a CNN with 10 layers.

In the second trial of experiment 1a, the number of fully connected layers are tested and the number of neurons in these layers. The parameters which are tested are shown in Table 4.4. The results are shown in Figure 4.7. The highest mean is of the second set of parameters: one fully connected layer with 256 neurons. If the number of neurons is lower, the standard deviation increases. Also, the addition of an extra fully connected layer does not contribute to a higher F1-score. One fully connected layer with 256 neurons is used in the next trials.
In the final trial of experiment 1, different trials are done to test the kernel size and the number of filters in each layer. To reduce the number of trials, the filters in layer 1 and 2, 3 and 4 and 5 till 12 are chosen the same. The combinations of parameters are shown in Table 4.5. The results are shown

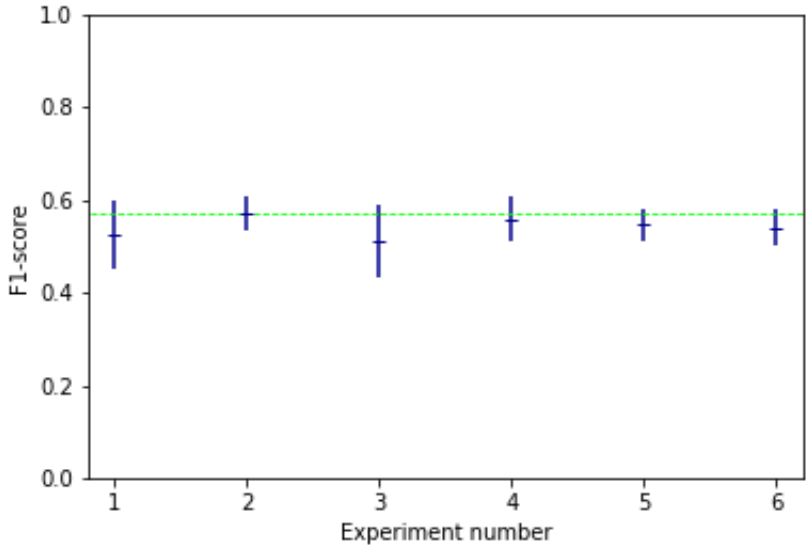


Figure 4.7: Results of the second trial of experiment 1a. The red dotted line indicated the highest mean F1-score. In this experiment, the highest F1-score corresponds to a CNN with one fully connected layer with 256 neurons.

in Figure 4.8. The differences are small, but the highest mean is reached with a kernel size of (5,1), 128 neurons in layers 1-4 and 64 neurons in layers 5-12. An increase of the kernel size leads to an increase of the standard deviation.

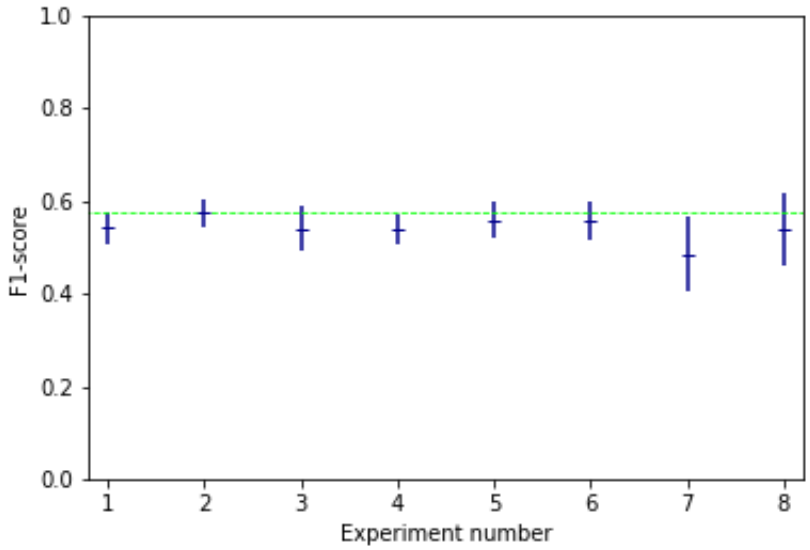


Figure 4.8: Results of third trial of experiment 1a. The red dotted line indicated the highest mean F1-score. In this experiment, the highest F1-score corresponds to a CNN with a kernel size of (5,1), 128 neurons in layers 1-4 and 64 neurons in layers 5-12.

Experiment 1b

In experiment 1a, the architecture of the CNN is chosen. The next step is to examine the best optimization parameters. This includes the loss function, the learning rate and the addition of a momentum. This is done in four trials. In the first trial 12 different loss functions are tested for their F1-score. The different loss functions are shown in Table 4.6. The results are shown in Figure 4.9.

The highest mean of the F1-score is found in the WCE loss function. This loss function is chosen for the rest of the trials.

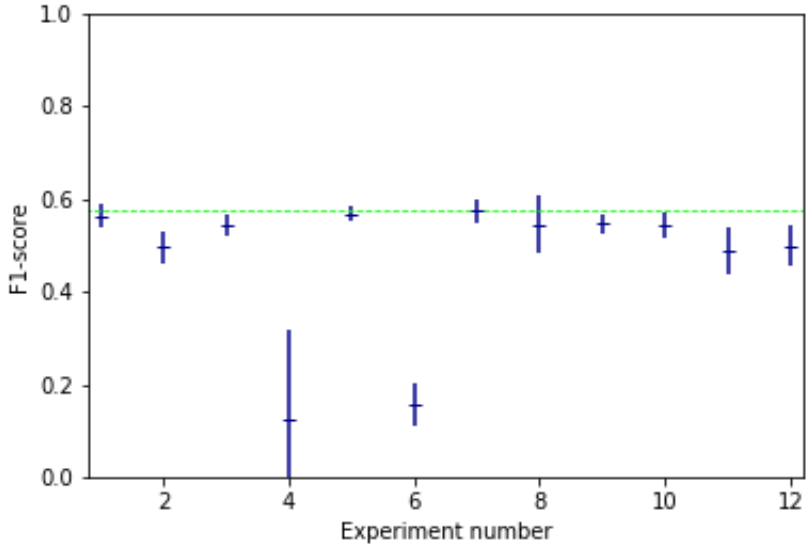


Figure 4.9: Results of the first trial of experiment 1b. The red dotted line indicated the highest mean F1-score. In this experiment, the highest F1-score corresponds to the WCE loss function.

In Figure 4.10 the results of the second trial are shown. There is an optimum at 1e-4. A high learning rate or a very low learning rate leads to an increase of the standard deviation.

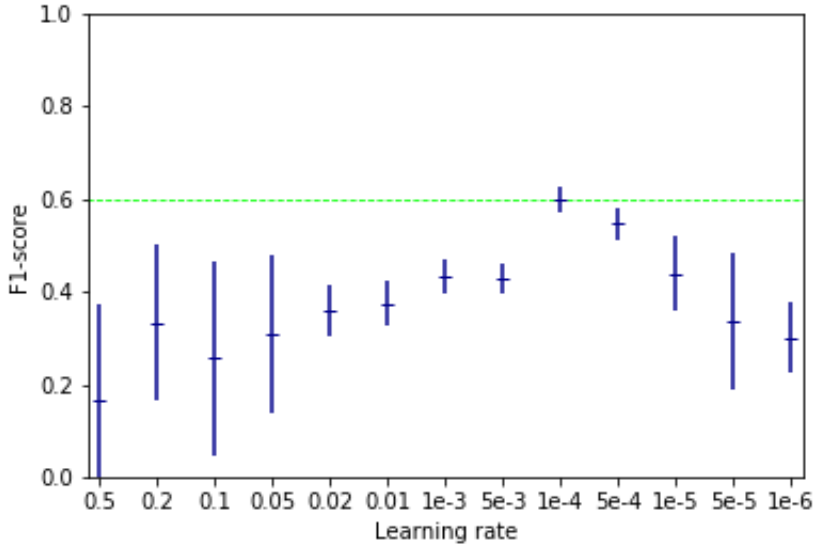


Figure 4.10: Results of the second trial of experiment 1b. The red dotted line indicated the highest mean F1-score. In this experiment, the highest F1-score corresponds to a learning rate of 1e-4.

In the final trial different momentums are tried. The results are shown in Figure 4.11. A large momentum is associated with a lower mean and a higher standard deviation of the F1-score. The smaller momentums are very close. The Nesterov momentum with a value of 0.01 has the highest mean and the lowest standard deviation and is therefore chosen as the last parameter.

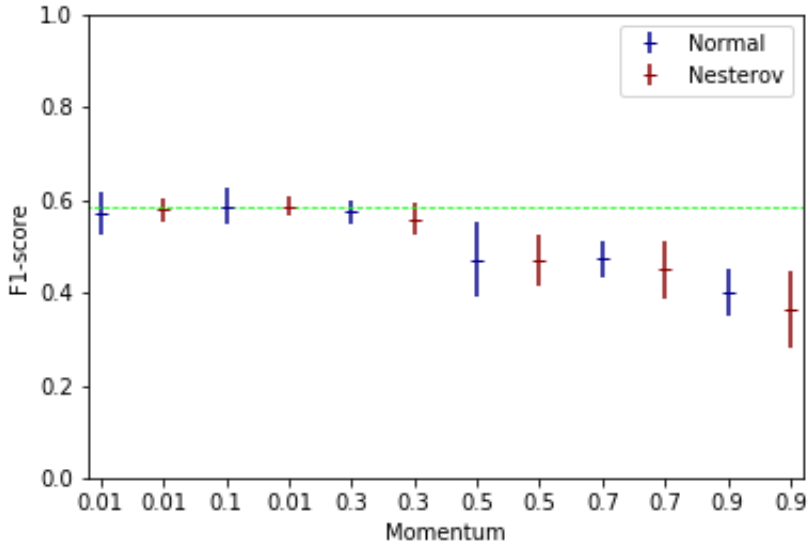


Figure 4.11: Results of the third trial of experiment 1b. The red dotted line indicated the highest mean F1-score. In this experiment, the highest F1-score corresponds to a Nesterov momentum of 0.01

The final network has a F1-score of 0.652 and an accuracy of 82.6%. The confusion matrix is shown in Figure 4.12.

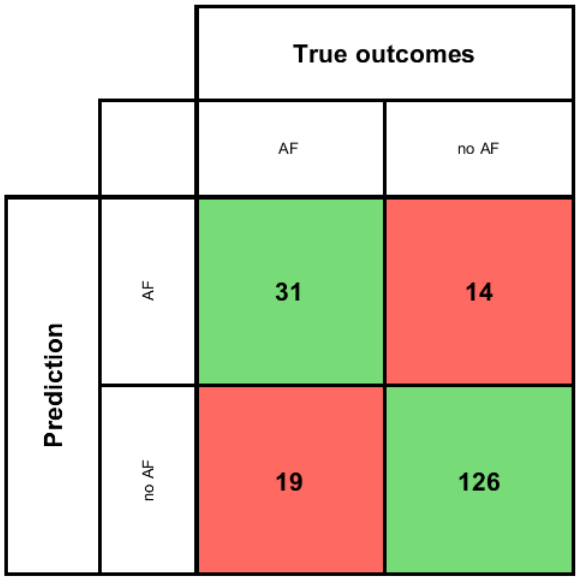


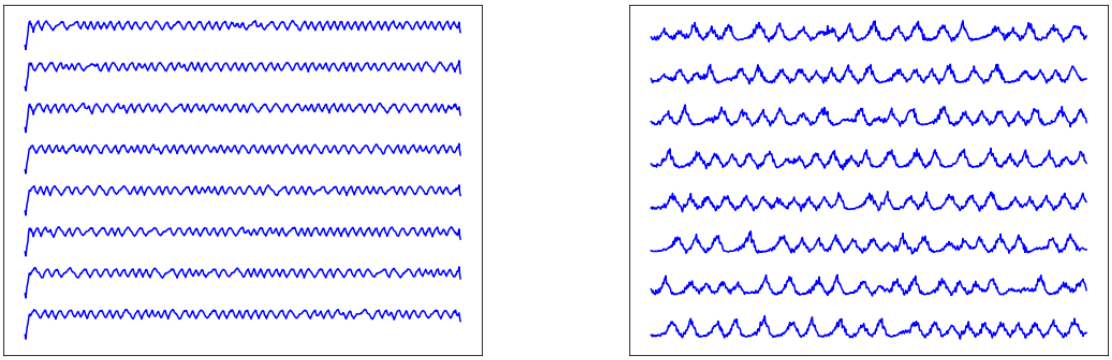
Figure 4.12: Confusion matrix of the final model for AF detection. A true positive is an ECG with AF which is classified by the model as AF, a false positive is an ECG without AF which is classified by the model as an ECG with AF. A false negative is the opposite: an ECG with AF which is classified as an ECG without AF and a true negative is an ECG without AF which is classified as an ECG without AF.

4.3.2 Experiment 2

Filter visualization

In Figure 4.13 a filter of layer 2 (on the left) and a filter of layer 4 are shown. It can be seen that the signals in layer 4 are more complex than the signals in layer 2. This can be seen in three things: the amplitude of the signal, the irregularity and the peaks. The amplitude in layer 4 is higher than

the amplitude in layer 2. The filter in layer 2 is very regular, the waves look a lot like each other and there is a regular repetition of the waves. The irregularity is seen in layer 4. There is more difference in the different waves and also the rhythm is less regular than in layer 2. The final difference is the presence of peaks in layer 4 where they are absent in layer 2.



(a) Layer 2 (b) Layer 4

Figure 4.13: Results of filter visualization

Occlusion maps

The occlusion maps are made for a TP, TN, FP and FN. Figure 4.14 shows the occlusion map of a TP. It can be seen that there are two QRS complexes that are less interesting than the rest of the signal (blue is a lower value than yellow).

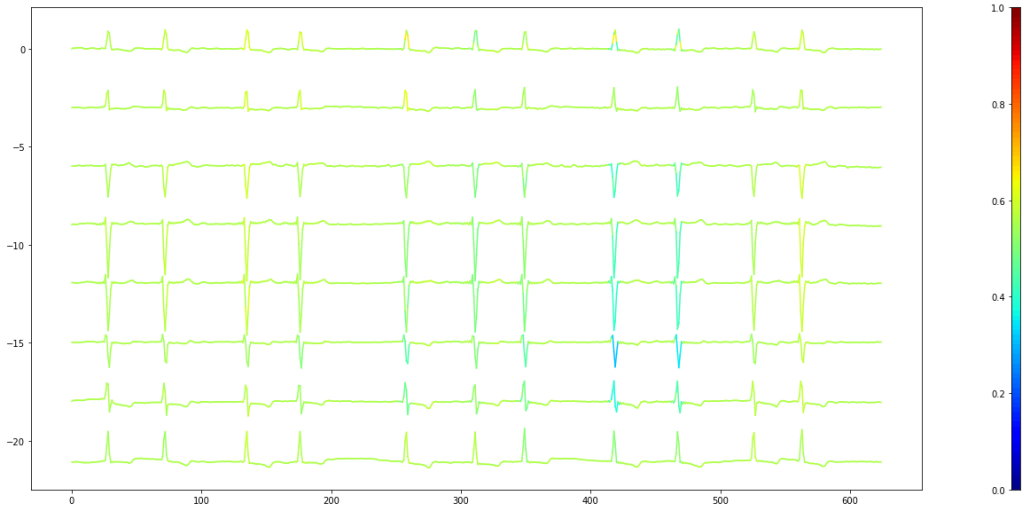


Figure 4.14: An occlusion map of a true positive

Figure 4.15 shows the occlusion map of a TN. Here, some QRS complexes are more interesting than the rest of the signal. This is in contrary with the occlusion map of a TP. Another thing that can be noticed is that the V1-V4 are more colored than I, II, V5 and V6. This is less the case in the TP, where only II and V6 are less colored. Figure 4.16 shows the occlusion map of a FP. Here, an extra ventricular heartbeat can be seen, and this is also picked as interesting by the network. The rest of the signal is not colored. In case of a false negative, shown in Figure 4.17, the network has no parts in the signal which it finds more interesting than other parts.

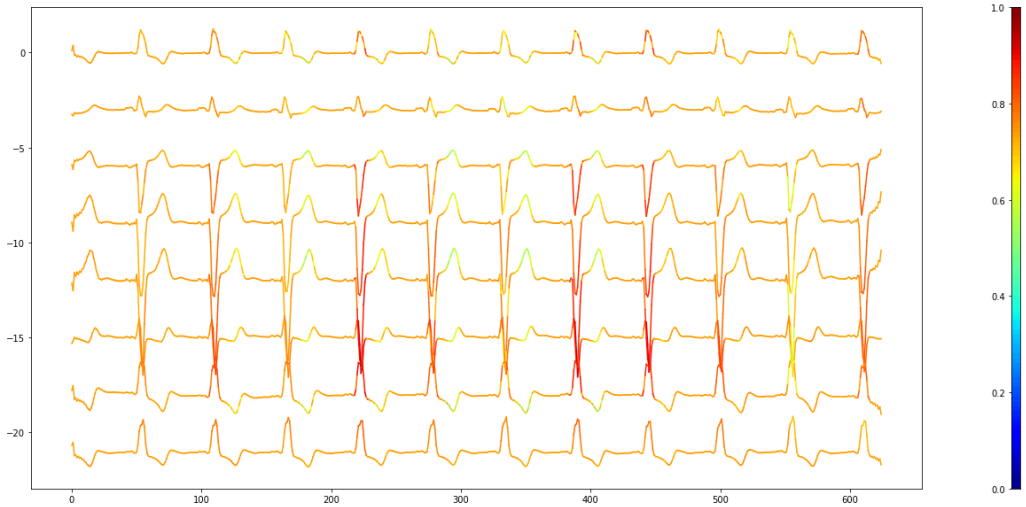


Figure 4.15: An occlusion map of a true negative

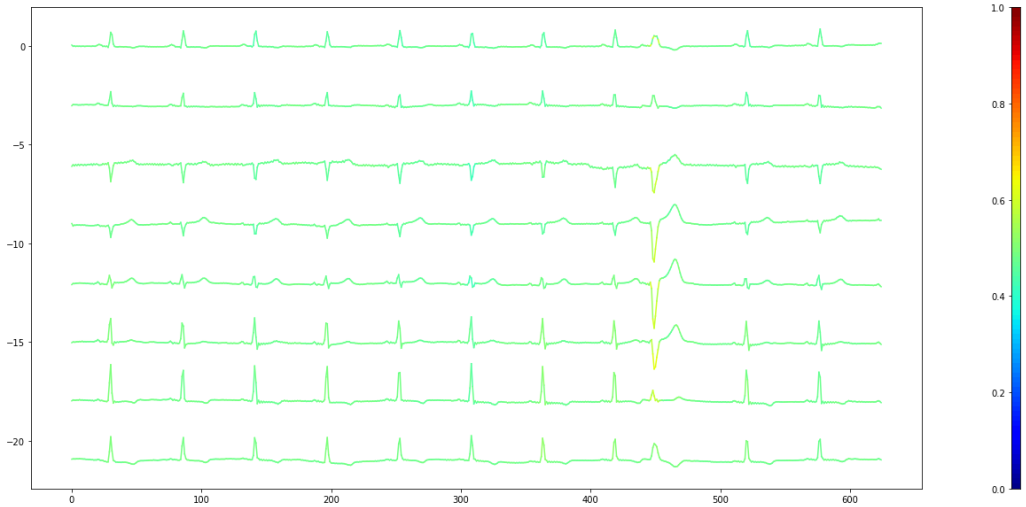


Figure 4.16: An occlusion map of a false positive

Figure 4.17: An occlusion map of a false negative

4.3.3 Experiment 3

In Figure 4.18 the results of experiment 3, prediction of mortality, are shown. The means of the F1-scores are very low. The mean of the F1-scores with transfer learning (0.286) is a fraction higher than the mean without transfer learning (0.23). The standard deviation is lower in the case of transfer learning. The confusion matrices are shown in Figure 4.19 of the models with the highest F1-score. The model with transfer learning has a F1-score of 0.293 and an accuracy of 72.3% and the model without transfer learning has a F1-score of 0.242 and an accuracy of 60.7%.

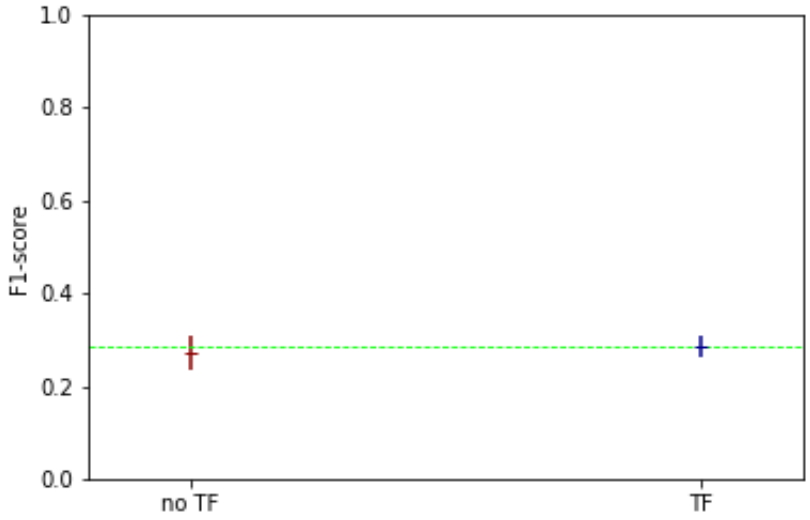
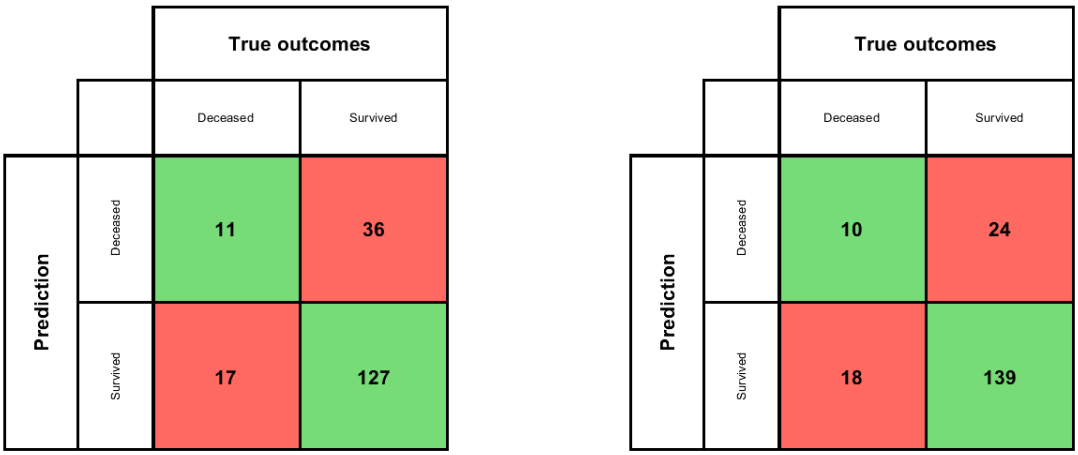


Figure 4.18: Results of experiment 3. The mean without transfer learning is 0.23 and with transfer learning is 0.286.



(a) Confusion matrix of the model trained with transfer learning

(b) Confusion matrix of the model trained without transfer learning

Figure 4.19: Confusion matrices of the final model for mortality prediction. A true positive is an patient who died within one year and is classified by the model as a patient who died within one year, a false positive is a patient who didn't die within one year and is classified as a patient who died within one year. A false negative is the opposite: a patient who died within one year and is classified as a patient who survived one year. A true negative is a patient who survived one year and is classified as such.

4.3.4 Experiment 4

In Figure 4.20 the results of the prediction of the improvement of dyspnea are shown. The result of the case with transfer learning is better (0.649) than in the case of training the network from scratch, without transfer learning (0.148). The mean of the F1-scores for the prediction are even higher than the AF detection.

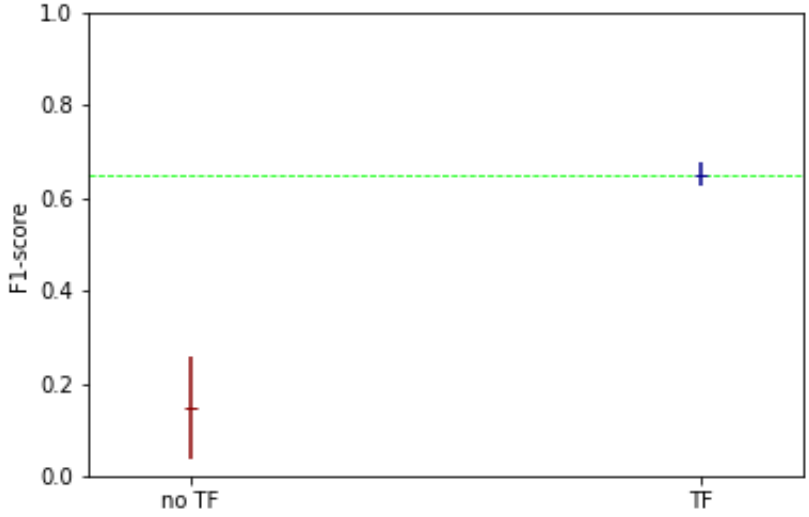
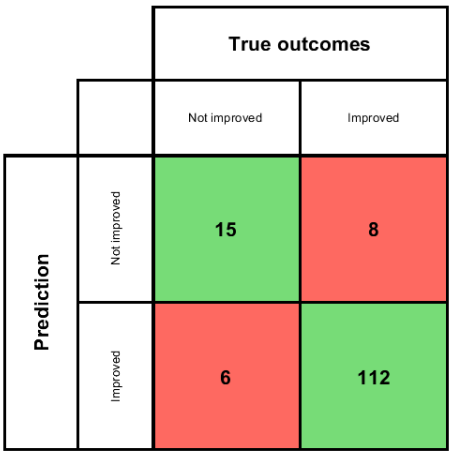
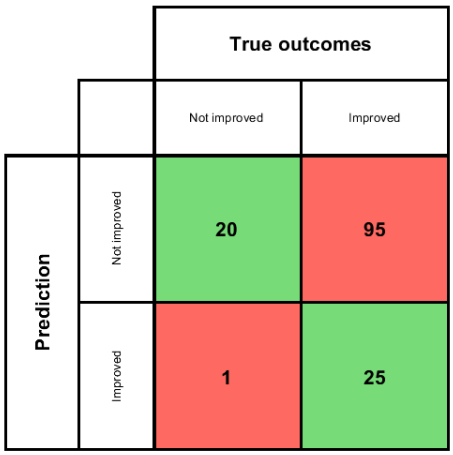


Figure 4.20: Results of experiment 4. The mean without transfer learning is 0.148 and with transfer learning is 0.649.

The confusion matrices are shown below of the models with the highest F1-score. The model with transfer learning has a F1-score of 0.682 and an accuracy of 90.1% and the model without transfer learning has a F1-score of 0.294 and an accuracy of 31.9%.



(a) Confusion matrix of the model trained with transfer learning



(b) Confusion matrix of the model trained without transfer learning

Figure 4.21: Confusion matrix of the final model for dyspnea prediction. A true positive is an patient who didn't improve after TAVI and is classified by the model as a patient who didn't improve, a false positive is a patient who improved and is classified as a patient who didn't improve. A false negative is the opposite: a patient who didn't improve and is classified as a patient who would have improved. A true negative is a patient who improved and is classified as such.

4.4 Discussion

4.4.1 Interpretation of results

Experiment 1

Experiment 1a is focused on finding the architecture of the CNN. The first trial indicates that ten convolutional layers are a good choice. The increase of the standard deviation for a high number of layers can be explained by the fact that the network has learned enough with ten layers and the convolutions do not make sense after the tenth layer. Besides that, with ten layers, the standard deviation is quite small, which means that the choice of ten layers is a steady situation.

In the second trial of experiment 1a, different numbers and sizes of the fully connected layer are analyzed. The results are very close, but the first and the third experiment have quite a high standard deviation. Both have 128 neurons in the first (and in case of the first experiment, the only) fully connected layer. The standard deviation is smaller with 256 neurons in the first fully connected layer. The network creates, thus, a lot of features, which are needed for the classification. The choice for one fully connected layer is also a good choice to prevent the network from overfitting.

In the final experiment of experiment 1a, different parameters are tested. The first thing that can be noticed is, that the standard deviation is high with a kernel size of 15. A lot of information is then missed because of the large convolutions. The other results (experiment 1-6) give all the same result more or less. This can be explained by the fact that the differences in the parameter choice are small.

In the first trial, different loss functions are tested. There is a lot of difference in the results. Three loss functions give similar results: the KL divergence, the WCE and the NLL. The KL divergence is also known as relative entropy. NLL and WCE are two different interpretations of the same formula, so that explains the close results. The results show that the NLL is less stable than the WCE. A cause of this could be, that the NLL is programmed manually, where the WCE was a built-in function in Keras. The WCE is chosen over the KL divergence, because of the computational time. On the amount of data used in this research, the difference is minimal, but with larger amounts of data, this could be relevant.

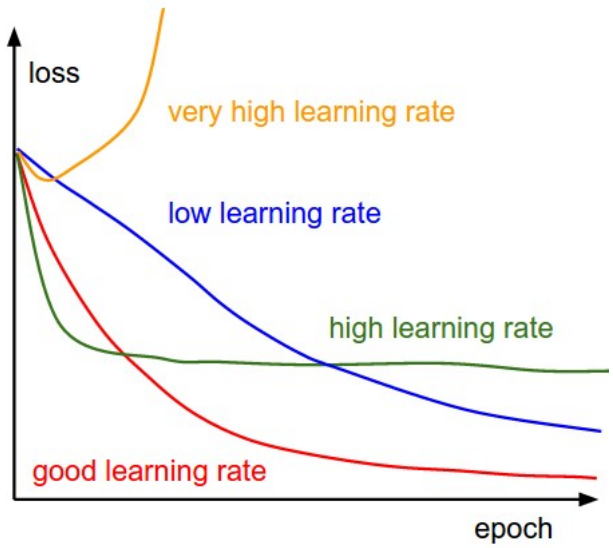


Figure 4.22: Different choices of the learning rate and their effect on the loss [43]

In the second trial of experiment 1b, the learning rate is tested. The step size is chosen very small so the results can be easily compared. However, a smaller step size could have given a slightly better result with another value in between 1e-4 and 5e-4. This can be tested in another experiment. A well-known figure for the learning rate is shown in Figure 4.22. The results of this work can be related

to this figure. A very high learning rate (yellow line in the Figure below and values of > 0.05 in this work) have a low mean and a high standard deviation of the F1-score. If the learning rate decrease a bit (green line in the Figure below and values between 0.02 and 0.005 in this work) the mean is still low, but the result is more stable. This can be seen by the lower standard deviation. Then there exists an optimum (red line in Figure 4.22 and a value of 0.0001 in this work) where the mean is maximal and the standard deviation is small. If the learning rate is further decreased (blue line in Figure 4.22 and values of < 0.0005 in this work), the mean decreases again and the standard deviation increases.

In the final trial of experiment 1b, a momentum is added to the SGD. A little momentum of 0.1 is optimal. If the momentum is increased the mean decreases and the standard deviation increases, so it becomes less stable.

Experiment 2

With the maximum activation technique, a lot of images are created (for each filter). It is hard to interpret these images, because they are signals instead of pictures. In Figure 3.7 these features are shown for a CNN, which processes images. If we compare this to the result in Figure 4.13, the signals in layer 4 can be described as more complex, because of the bigger amplitude, the more irregularity and the peaks.

For the occlusion maps, it can be seen that in both the TP and the TN the QRS complex is important for classification. For the interpretation it is important that if a part of the signal is deleted and the probability of the classification changes, the signal is either shown as more important or less important. In case of the TP, if a QRS complex is deleted, the probability increases, because it is shown as less important. If a QRS complex is deleted in an irregular heartbeat, the heartbeat becomes even more irregular and the network can classify it even better as AF. In case of a TN, it is the other way around. The deletion of a QRS complex results in an irregular heartbeat, which is associated with AF. Then, the network has a harder task to classify it as no AF. From this it can be seen that the network learns to identify an irregular heartbeat and associates this with AF. Another thing that may be noticed is that the different colors are also associated with a higher amplitude in the signal. This is not entirely true, because there are other examples in which the amplitude doesn't play a role, see Appendix B. This can also be seen in the TP, Figure 4.14, where V5 is colored, but the amplitude is low.

For the FP and the FN, two things can be noticed. For the FP, in the example shown there is another irregularity in the ECG; a extra ventricular heartbeat. Other examples are bundle track blockage or just an irregular heartbeat. In case of the FN, there is nothing classified as important, so the network doesn't know what to use for the decision.

Experiment 3 and 4

For the prediction of mortality after TAVI, the F1-scores of the model in this report are not good. To train the model with transfer learning is a little bit better than without transfer learning (0.286 vs 0.275).The F1-scores are not looking good, but if the accuracy is calculated, the model is performing quite well. An accuracy of 72.3 % is reached for the prediction of mortality based on one ECG. This shows that F1-scores are very useful to compare two different networks, but the accuracy tells more about the actual performance of a network.

For the prediction of improvement of dyspnea, the F1-scores for transfer learning are better and the accuracy is even better. Without transfer learning, the results are much worse. To see what the difference is in what the network has learned, again occlusion maps are made. In Figure 4.23 and Figure 4.24 the occlusion maps are shown of a true positive from the model with transfer learning and without transfer learning, respectively. As can be seen, the model with transfer learning is learning features from the ECG, but the model without transfer learning, the model is learning nothing from the ECG.

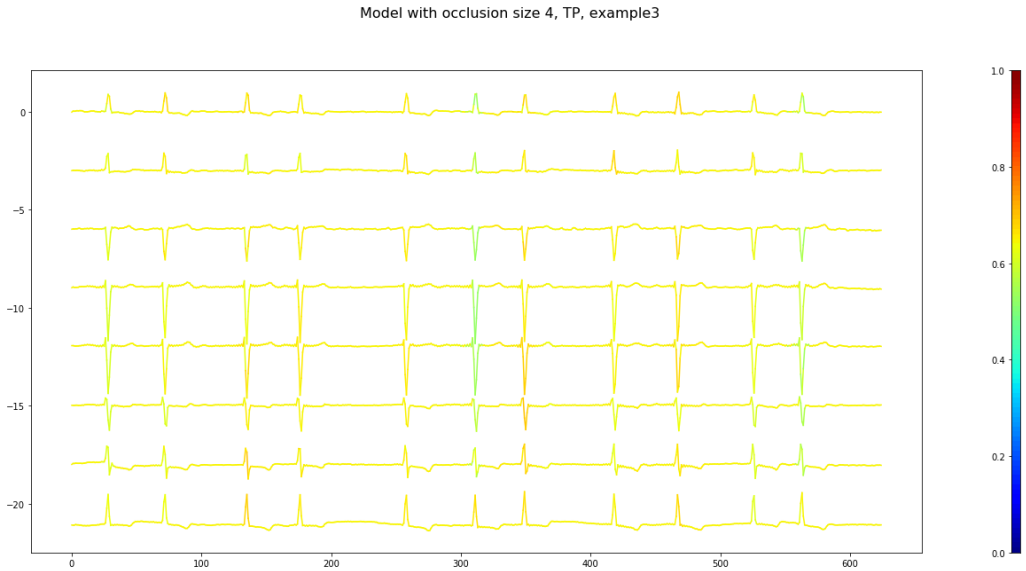


Figure 4.23: Occlusion map of a true positive from the model with transfer learning.

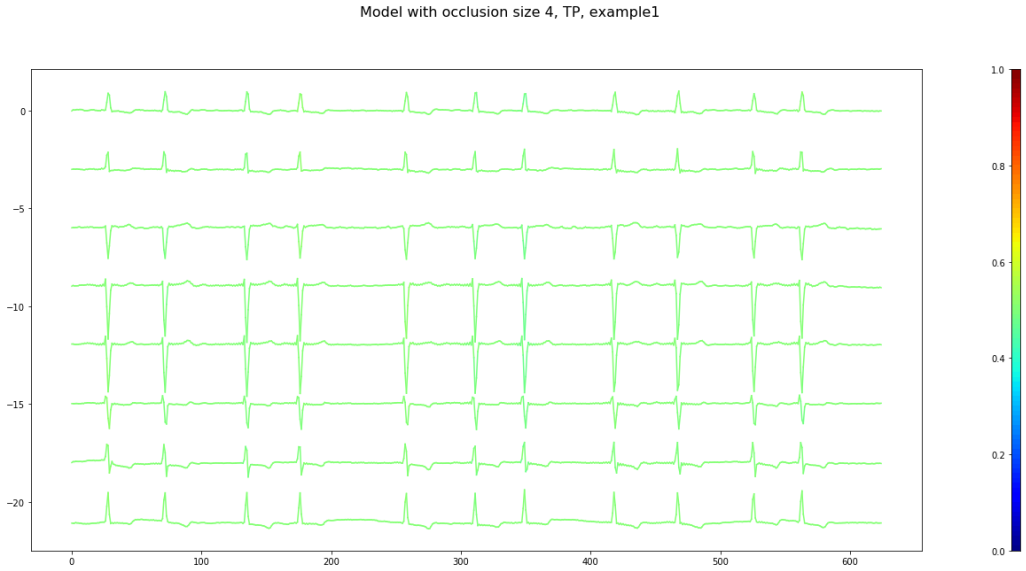


Figure 4.24: Occlusion map of a true positive from the model without transfer learning.

If this results are compared with the results by Lopes et al.[7], some interesting difference can be seen. The models by Lopes are made with data from the same patient population used in this research, but the clinical and laboratory values are used instead of the ECG. Where in this work the prediction of improvement of dyspnea is much better, in that work, the prediction of mortality is better. A reason for this could be that the ECG contains more information to predict improvement in dyspnea and the clinical and laboratory values contain more information to predict mortality. This is interesting, because it is known that dyspnea can be present as a symptom of rhythm disorders. This might be an underlying reason for the good results of the dyspnea prediction in the network that is trained for AF detection.

4.4.2 Limitations

In this work a CNN is trained to classify AF. In this dataset, the presence of AF is labeled. In the ECGs without AF, other arrhythmias could be presented. In the dataset, variations in abnormalities is present. There are ECGs with extra ventricular beats, artifacts and pacemaker induced heart beats.

For a CNN to reach a higher performance, the variation could be reduced or the number of data points must be increased.

The performance measure used to compare the results between different models is F1-score. This measure is chosen because of the imbalanced dataset. The interpretation of this score is harder than the interpretation of, for example, accuracy. Accuracy is a linear measure, where the F1-score is non-linear. Where a network with an accuracy of 0.5 is equal to a network with a F1-score of 0.5, it is very different if both scores are 0.6. Then, the network with an accuracy of 0.6 is not that good, where a network with a F1-score of 0.6 is already quite well and also a model with a low F1-score can perform well. This can be seen in the results of experiment 3. To better show the performance of the final networks (for AF classification, mortality prediction and prediction of improvement of dyspnea) the accuracy of these models is calculated.

In the first experiment, the choice is made to do ten repetitions for each experiment in each trial. With ten repetitions, you can say something about the mean and the standard deviation, but with more repetitions these values become more valuable. Due to computational time, the choice is made to do ten repetitions. The different outcomes for the repetitions show there is some randomness within the results. This randomness is partly caused by the choice for the optimization technique: SGD. It is also caused by the unreliability of some of the parameters. Therefore, the standard deviation is chosen as the marker for repeatability. If the standard deviation is low, the change for a same outcome the next time is high. Another limitation in the first experiment is the step by step optimization of both the architecture and the optimization. The problem is a convex problem. If the order of trials was chosen differently, the outcome could have been different. The choices for these orders are made by the detail of the parameter. The more detailed the parameter is, the later it will be tested.

In the second experiment, two visualization techniques are applied. There exists a lot of other visualization techniques, which are not applied. They might give new insights. A limitation of the maximization technique is the noisy input which can affect the outcome of this technique. A limitation of the occlusion maps is that it doesn't show connections which are used by the CNN, only little areas in the ECG. In experiment 3 and 4, the network is used which is optimized for AF detection. It is not optimized for prediction purposes. However, the network is optimized on the same kind of data, an 8 leads ECG.

4.4.3 Comparison with literature

Quite a lot is published for AF detection using CNNs. Most papers describe methods for one- or two-lead ECG. Most of them use longer recordings (Holter recordings). [44] If 12-leads ECGs are used, it is most of the time for biometric purposes. [45] Different approaches are used for AF detection: the whole one-lead signal or a beat-to-beat algorithm is used. The results of both are quite good, but these are not applicable in this situation. The only available data are the 12-leads ECG. There is not much published about the results of 12-leads ECG with CNN for prediction outcomes. Prediction outcomes are mainly done with clinical data, genomics or imaging data. [46] Clinical data is used in prediction of heart diseases and diabetes. [47, 48] Genomics and imaging data is used in cancer predictions. [49]

Chapter 5

Prediction based on multiple data points

In this chapter, the predictive value of ECG is further investigated. The last chapter used one ECG to predict the outcome of TAVI patients. In this chapter the course of the ECG is used. First, the choices in the feature extraction method are exemplified. Then a method for combining multiple ECGs is proposed. This is followed by de method for prediction. The results show the additional value for using more than one ECG. The chapter finishes with a discussion, where the results are interpreted, the limitations are elaborated and the connection with the current literature is shown.

5.1 Introduction and motivation

To further investigate the predictive value of ECG on the outcome of TAVI patients, the possibilities of using multiple data points is investigated. As a start for multiple data points, two ECGs are used. To describe the course of these two ECGs, the set-up shown in Figure 5.1, is used. The course of the ECG can also be described as mapping the features from one point in time to another point in time. The features of two ECGs are extracted and then combined into one fully connected layer, which is used for prediction. The features calculated by the CNN could be used. Other features could be more related to the clinical features, as described in the Section 2.3.1. This method is chosen in this part, because of the easy visualization of these features and the low dimensionality of the data. If a method is found where these features can be mapped, the higher dimensional features, calculated by the CNN, could be used. This step is not done in this study.

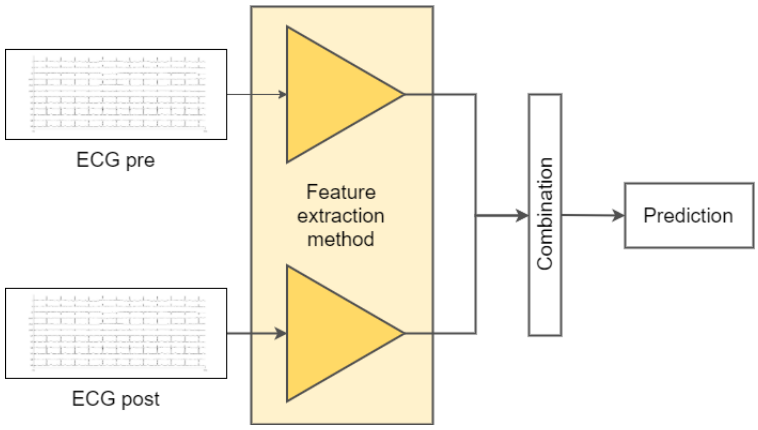


Figure 5.1: Set-up of the method of using two ECGs for prediction of outcome

5.2 Method

5.2.1 Feature extraction

The feature extraction method used in this chapter is signal analysis. A method for R peak detection and P and T wave detection are described. The method for R peak detection is developed by Pan and Tompkins. [50] The method for P and T wave detection is developed by Elgendi et al. [51] The next sections describe how the features are extracted with these methods and what the correlation between those features is.

R peak detection

This R peak detection is developed by Pan and Tompkins [50] and is also known as the Pan-Tompkins algorithm. It consists of four steps. As an input, the ECG is used. The first step is filtering, the second step is a differentiation, the third step is squaring and the final step is moving window integration. The steps are explained below.

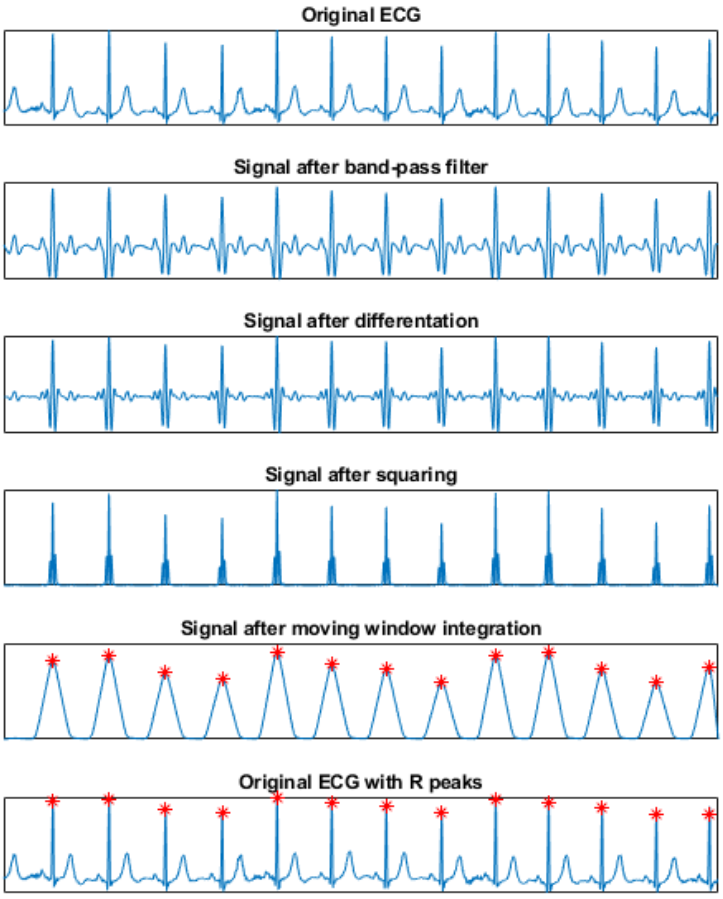


Figure 5.2: All steps of the Pan-Tompkins algorithm

The first step is a band-pass filter. This filter reduces the influence of noise, baseline drift and interference of the P- and T waves. The bandwidth of the filter is 5-15 Hz to maximize the QRS power, see Figure 5.3. After this filter, the signal is differentiated. With this technique, the slope of the QRS complex is provided. This can be done by the transfer function

$$H(z) = \frac{1}{8}T\left(-z^{-2} - 2z^{-1} = 2z^{1} + z^{2}\right).$$

And the difference equation is

$$y(nT) = \frac{1}{8}T\left(-x(nT-2T) - 2x(nT-T) + 2x(nT+T) + x(nT+2T)\right)$$

The third step is to square the signal elementwise. Now all data points are positive. To obtain waveform feature information a moving window integration is applied. The difference equation of this calculation is

$$y(nT) = \frac{1}{N}\left(x(nT-(N-1)T) + x(nT-(N-2)T) + \cdots + x(nT)\right).$$

Here, N is the width of the integration window. In Figure 5.2 the signal is shown after every step. To find the R peaks, a function can be applied to the signal after the integration step. This function calculates the first derivative and analyzes where this derivative cross the x-axis (equal to zero).

P and T wave detection

The detection method developed by Elgendi et al. [51] claims to be effective in arrhythmic ECG signals. It needs prior information about the location of the R peaks. For this, the Pan-Tompkins algorithm can be used. The algorithm consists of four steps. The ECG is used as input and the first step is the application of a band-pass filter. After this filter, the QRS complexes are removed. The third step is to generate blocks of interest and the final step uses thresholds to detect the P and T wave. The signal after each step is shown in Figure 5.4.

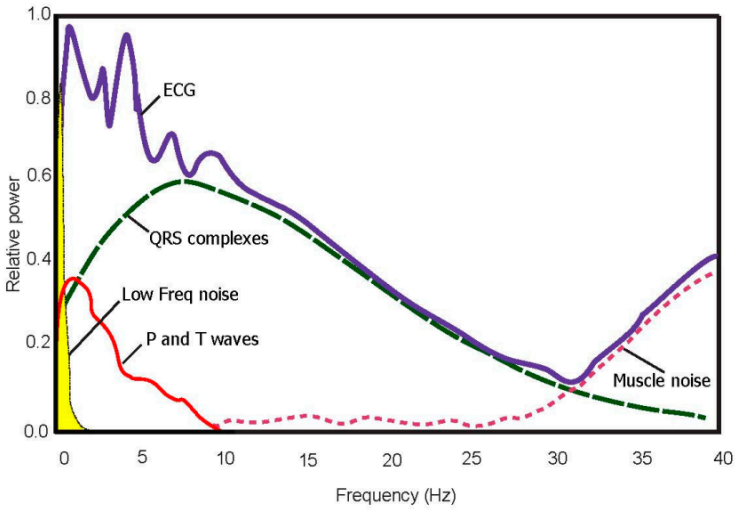


Figure 5.3: Frequencies which are present in the ECG signal [51]

The first step is to apply a band-pass filter to reduce the influence of noise like above. The frequency band is different in this method, because the focus is on P and T waves. The bandwidth is 0.5 – 10 Hz. It can be seen in Figure 5.3 that the P and T waves are present in this bandwidth. The next step is the removal of the QRS complexes. The Pan-Tompkins finds the R peaks. To make sure the QRS complex is completely removed, 83 ms before the R peak and 166 ms after the R peak are set to zero.

During the third steps, blocks of interest are generated. The P and T wave are in such a block, but so might other parts of the signal. So first, the blocks are generated and then thresholds are set to find the P and T wave in these blocks. To generate the blocks, two moving averages filters are applied on the signal after QRS removal. The aim of the first moving average filter (MA_{peak}) is to differentiate the P and T waves. The second moving average filter ($MA_{P_{wave}}$) is used as a threshold

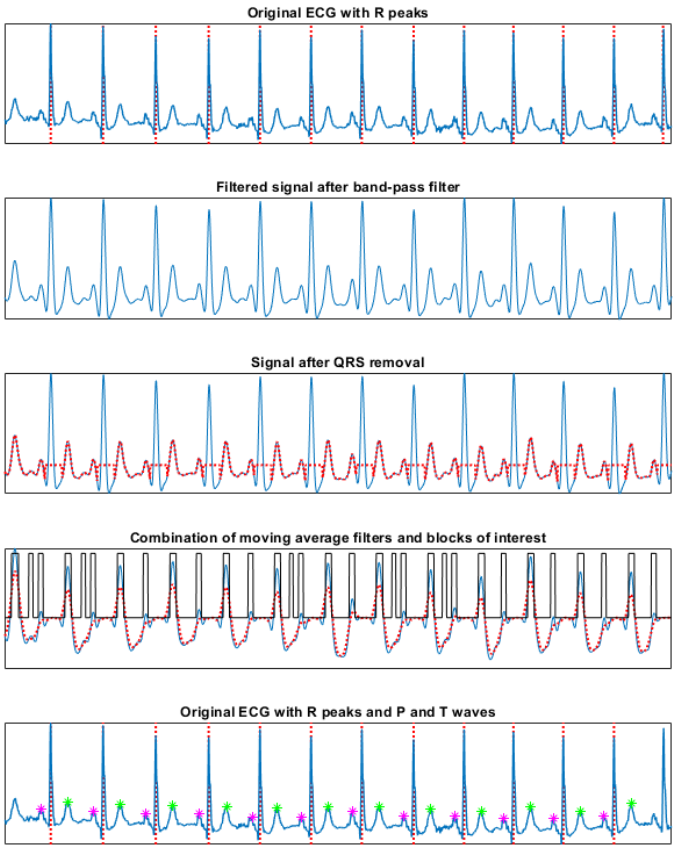


Figure 5.4: Different steps of the P and T wave detection algorithm. The last R peak is ignored, because it is too close to the border.

for the first moving average filter. The equations for both filters are stated below.

$$\mathrm{MA}_{\mathrm{peak}}[n] = \frac{1}{W_1}\left(x\left[n-\frac{W_1-1}{2}\right]+\cdots+x[n]+\cdots+x\left[n+\frac{W_1-1}{2}\right]\right)$$

$$\mathrm{MA}_{\mathrm{P_{wave}}}[n] = \frac{1}{W_2}\left(x\left[n-\frac{W_2-1}{2}\right]+\cdots+x[n]+\cdots+x\left[n+\frac{W_2-1}{2}\right]\right)$$

where $W_1 = 55$ ms, which is half the size of the P wave and $W_2 = 110$ ms, which is the size of the P wave. For the first filter, halve the size of the P wave is chosen, because the P wave can have different forms and sizes in arrhythmias. Now, the blocks are generated on the time steps where $\mathrm{MA}_{\mathrm{peak}} > \mathrm{MA}_{\mathrm{P_{wave}}}$. There are three possibilities after this step:

1. No blocks of interest
   No detection of P or T wave in the RR interval.
2. One block of interest between two R peaks
   The P and T wave are most likely merged within one block.
3. More than one block detected between two R peaks
   The blocks with P and T wave are in the intervals given in Table 5.1. If multiple blocks are within this interval, the block with the highest amplitude is selected.

Table 5.1: Prior knowledge about the features in the ECG that is used in the P and T wave detection algorithm

Feature	**Normal interval**	**Samples** ($F_s = 250Hz$)	**Samples** ($F_s = 500Hz$)
P wave	90 – 130 ms	23 — 33	45 — 65
PR interval	120 — 200 ms	30 — 50	60 — 100
QRS complex	80 — 120 ms	20 — 30	40 — 60
QTc interval	360 — 440 ms	90 — 110	180 – 220

Feature extraction

In the previous sections, the method to extract the R peaks and the P and R peaks are described. Once these peaks are known, the HR, the PR time and the RT time can be calculated. The meaning of these clinical features is described in Section 2.3.1 and can be seen in Figure 5.5. To compare the features in the ECG before TAVI and the ECG after TAVI, the mean is taken from the PR time and the RT time. The correlations are investigated and described in the next section

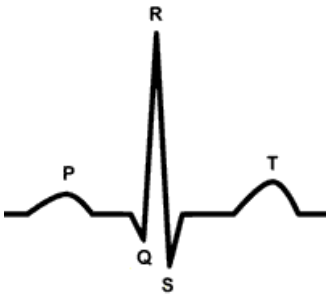


Figure 5.5: ECG Complex. PR: time between the P peak and the R peak, RT: time between R peak and T peak. [22]

Feature correlation

To check whether the feature extraction method is valid, the correlation between all the features is calculated. It is expected that there exists a correlation between the equivalent features before and after the TAVI, but there is no correlation expected between the different features. The calculated features are shown in Figure 5.6. The expected is seen.

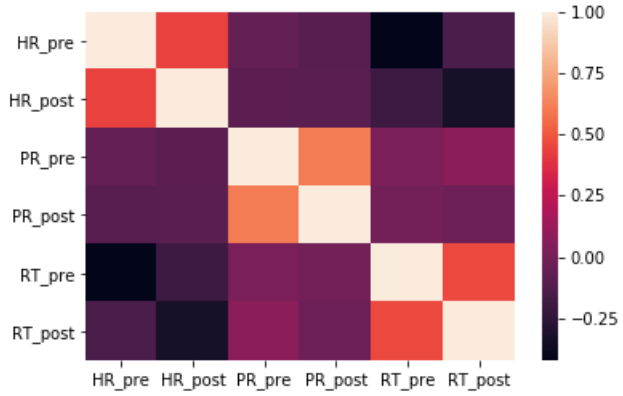


Figure 5.6: Correlation between the extracted features with signal analysis

5.2.2 Model

Mathematical problem description

Now that the features are extracted, the problem can be described mathematically. There exist two inputs u_0, $u_1 \in \mathbb{R}^3$, ECG pre and ECG post. The size of u_i is either 2500x8 or 5000x8, based on the sample frequency ($F_s = 250$ Hz or $F_s = 500$ Hz). There exists a feature extraction function f (described in the sections above), such that

$$x = f(u) = \begin{bmatrix} HR \\ HR_{IR} \\ PR \\ RT \end{bmatrix}$$

where x_i is a feature vector, $x_i \in \mathbb{R}^4$, $HR \in \mathbb{R}$, the HR, $HR_{IR} \in \{0, 1\}$, irregular HR, $PR \in \mathbb{R}$, PR interval, and $RT \in \mathbb{R}$, RT interval. The function f can also be another feature extraction method, like a CNN. Then,

$$\begin{aligned} x_0 &= f(u_0) \\ x_1 &= f(u_1). \end{aligned}$$

For each couple $X = \{x_0, x_1\}$ there exist an outcome label y That results in a dataset:

$$\{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$$

X can be defined as the difference between x_0 and x_1 or x_0 and x_1 can be stacked to one vector. The two different methods are described in the following subsections. For the outcome, the same outcomes as in the previous chapter are used: mortality and improvement of dyspnea. Here, mortality is defined as dead within one year after the TAVI procedure. Improvement of dyspnea is defined as an improved NYHA score after the TAVI in comparison with the NYHA score before TAVI.

Distance measure

A distance measure must be defined for the distance between x_0 and x_1. For this distance measure the following assumptions are made:

- $d(x_0, x_1) = 0$ if and only if $x_0 = x_1$
- $d(x_0, x_1) \neq d(x_1, x_0)$

Table 5.2: Overview of non-symmetric distance measures

Name	Equation
Relative J-divergence [53]	$J_R(x_0, x_1) = \sum_{i=1}^{n} (x_{0_i} - x_{1_i}) \log \frac{x_{0_i} + x_{1_i}}{2x_{1_i}}$
KL [54]	$KL(x_0, x_1) = \sum_{i=1}^{n} x_{0_i} \log \frac{x_{0_i}}{x_{1_i}}$
Relative AG divergence [55]	$G(x_0, x_1) = \sum_{i=1}^{n} \left(\frac{x_{0_i} + x_{1_i}}{2} \right) \log \frac{x_{0_i} + x_{1_i}}{2x_{01_i}}$

The last assumption makes it a non-symmetric distance measure. [52] There exist a lot of distance measures, three non-symmetric measures are given in Table 5.2.
The relative J-divergence is not applicable, because $J_R(x_0, x_1) \geq 0$. The choice is made to continue with the KL measure. Then the equations for the continuous features are

$$KL_{HR}(HR_{\text{pre}}, HR_{\text{post}}) = HR_{\text{pre}} \log \frac{HR_{\text{pre}}}{HR_{\text{post}}}$$

$$KL_{PR}(PR_{\text{pre}}, PR_{\text{post}}) = PR_{\text{pre}} \log \frac{PR_{\text{pre}}}{PR_{\text{post}}}$$

$$KL_{RT}(RT_{\text{pre}}, RT_{\text{post}}) = RT_{\text{pre}} \log \frac{RT_{\text{pre}}}{RT_{\text{post}}}$$

For the binary feature, HR_{IR} another measure must be defined.

$HR_{\text{IR}_{\text{pre}}}$	$HR_{\text{IR}_{\text{post}}}$	$d_{\text{IR}} = d\left(HR_{\text{IR}_{\text{pre}}}, HR_{\text{IR}_{\text{post}}}\right)$
0	0	0
1	1	1
0	1	2
1	0	3

That results in the following datasets:

$$\{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$$

where X_i is defined as

Dyspnea prediction	**Mortality prediction**
$X = \begin{matrix} KL_{HR} \\ d_{IR} \\ KL_{PR} \\ KL_{RT} \end{matrix}$	$X = \begin{matrix} KL_{HR} \\ d_{IR} \\ KL_{PR} \\ KL_{RT} \end{matrix}$
$y = NYHA_{improved}$	$y = mortality_{1year}$

Stack the features

Another method to combine the two feature vectors is stacking them into one vector and make a prediction based on this new feature vector. Then the operation done is

$$x_0, x_1 \mapsto x_n$$

where

$$x_0 = \begin{bmatrix} HR_{\text{pre}} \\ HR_{\text{IR}_{\text{pre}}} \\ PR_{\text{pre}} \\ RT_{\text{pre}} \end{bmatrix}, x_1 = \begin{bmatrix} HR_{\text{post}} \\ HR_{\text{IR}_{\text{post}}} \\ PR_{\text{post}} \\ RT_{\text{post}} \end{bmatrix}$$

and X_S is defined as

$$X_S = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} HR_{\text{pre}} \\ HR_{\text{post}} \\ HR_{\text{IR}_{\text{pre}}} \\ HR_{\text{IR}_{\text{post}}} \\ PR_{\text{pre}} \\ PR_{\text{post}} \\ RT_{\text{pre}} \\ RT_{\text{post}} \end{bmatrix}.$$

Now, the dataset is

$$\{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$$

where $X_i = X_{S_i}$ and $y_i = NYHA_{improved}$, in case of NYHA prediction, or $y_i = mortality_{1year}$, in case of mortality prediction.

The feature extraction part is done in MATLAB. The further analysis and experiments are done in Jupyter from Anaconda.

5.2.3 Experiments

To check whether multiple data points have additional value for prediction, two ECGs are used for prediction. Three experiments are conducted: two with stacked features and one with the features where a distance is measured. All these experiments are conducted for the prediction of improvement of dyspnea and the mortality prediction.

The first experiment is focused to investigate the use of a simple NN with the stacked features. This is like the method described in Figure 5.1, the NN consists of one fully connected layer and one classification layer. A simple parameter search is done to find the network with the highest accuracy. The number of neurons in the hidden layer can be 32 or 64. There are only eight features to train on, so the hidden layer doesn't need to be big. For the optimization, the loss function is the WCE. The optimal learning rate is found by experimenting between 1e-6 and 0.05. The best result of each experiment is shown in the Section 5.3.1. To compare the outcome of the NN, the stacked features are used as input for a SVM to predict the outcome during the second experiment. Different parameters can be chosen. Here, the results with different kernels and the parameter C are compared. The accuracy and the confusion matrix of the best result for both outcomes are shown in Section 5.3.2. The final experiment investigates how well the distance measure can be used to predict the outcome. For this, again an SVM is used and the same parameters are tested on their outcome, the kernel and parameter C. The hypothesis is that the results of the stacked features and the NN, experiment 1, are the best, since a NN can find more patterns than an SVM. Another hypothesis is that the SVM with stacked features is outperforming the SVM with distance features, since the stacked features contain more information than the distance features.The results of the final experiment are shown in Section 5.3.3.

5.3 Results

5.3.1 Experiment 1

In this experiment, a simple NN with one fully connected layer is trained to predict both mortality after one year and the improvement of dyspnea after a TAVI procedure.

Mortality prediction

The confusion matrix for the prediction of mortality by a NN is given in the confusion matrix below. The model has an accuracy of 81.1 % with 32 neurons in the hidden layer. A model with 64 neurons in the hidden layer is also performing well: an accuracy of 79.4 %. The confusion matrix of this model is not shown here.

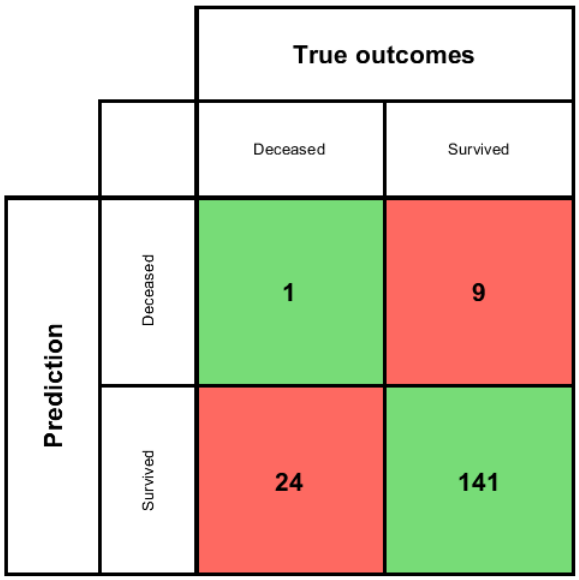


Figure 5.7: Confusion matrix of the stacked features and NN for mortality prediction. A true positive is an patient who died within one year and is classified by the model as a patient who died within one year, a false positive is a patient who didn't die within one year and is classified as a patient who died within one year. A false negative is the opposite: a patient who died within one year and is classified as a patient who survived one year. A true negative is a patient who survived one year and is classified as such.

Prediction of improvement of dyspnea

The confusion matrix for the prediction of improvement of dyspnea by a NN is shown in Figure 5.8. The model has an accuracy of 75.8 % with 64 neurons in the hidden layer. A model with 32 neurons in the hidden layer is not performing well.

5.3.2 Experiment 2

In this experiment the stacked features are used with an SVM to predict both mortality and improvement of dyspnea.

Mortality prediction

The confusion matrix for the prediction of mortality by an SVM s shown in Figure 5.9. The model has an accuracy of 78.9 %. The kernel used is the radial basis function kernel and the value of parameter C is 100.

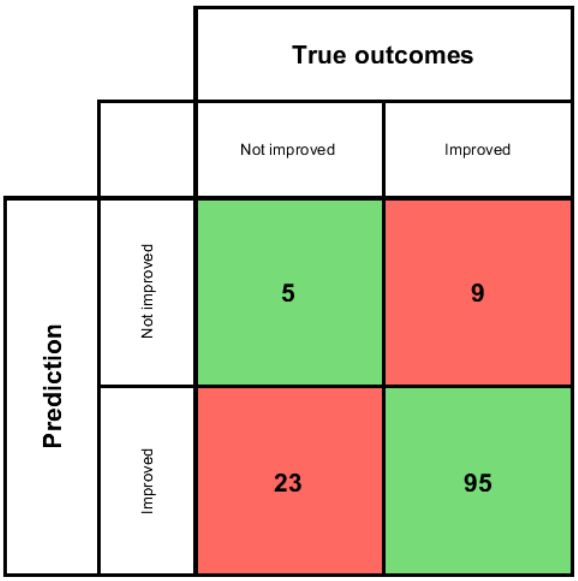


Figure 5.8: Confusion matrix of the stacked features and NN for dyspnea prediction. A true positive is an patient who didn't improve after TAVI and is classified by the model as a patient who didn't improve, a false positive is a patient who improved and is classified as a patient who didn't improve. A false negative is the opposite: a patient who didn't improve and is classified as a patient who would have improved. A true negative is a patient who improved and is classified as such.

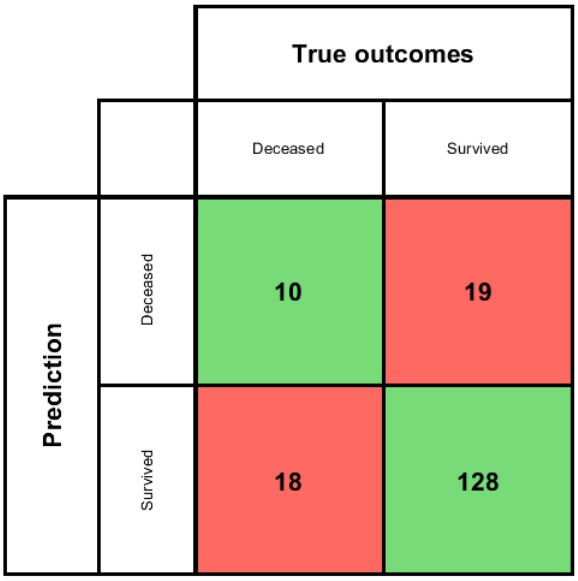


Figure 5.9: Confusion matrix of the stacked features and SVM for mortality prediction. A true positive is an patient who died within one year and is classified by the model as a patient who died within one year, a false positive is a patient who didn't die within one year and is classified as a patient who died within one year. A false negative is the opposite: a patient who died within one year and is classified as a patient who survived one year. A true negative is a patient who survived one year and is classified as such.

Prediction of improvement of dyspnea

In Figure 5.10, the confusion matrix of the prediction of dyspnea is shown. The model is an SVM with a radial basis function kernel and the value of parameter C is 10. The SVM reaches an accuracy of 68.2 %.

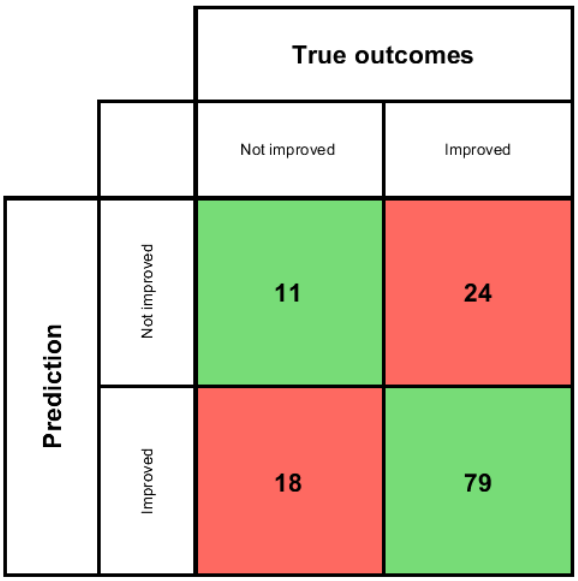


Figure 5.10: Confusion matrix of the stacked features and SVM for dyspnea prediction. A true positive is an patient who didn't improve after TAVI and is classified by the model as a patient who didn't improve, a false positive is a patient who improved and is classified as a patient who didn't improve. A false negative is the opposite: a patient who didn't improve and is classified as a patient who would have improved. A true negative is a patient who improved and is classified as such.

5.3.3 Experiment 3

In the final experiment the distance features are used with an SVM to predict both mortality and improvement of dyspnea.

Mortality prediction

The confusion matrix is shown in Figure 5.11 for the prediction of mortality. The SVM is based on a linear kernel and the value for parameter C is 1. This SVM reaches an accuracy of 71.4 %.

Prediction of improvement of dyspnea

The SVM with the distance features reaches an accuracy of 69.7 % for the prediction of improvement of dyspnea. The confusion matrix is shown in Figure 5.12. The SVM has a linear kernel and the value of parameter C is 1.

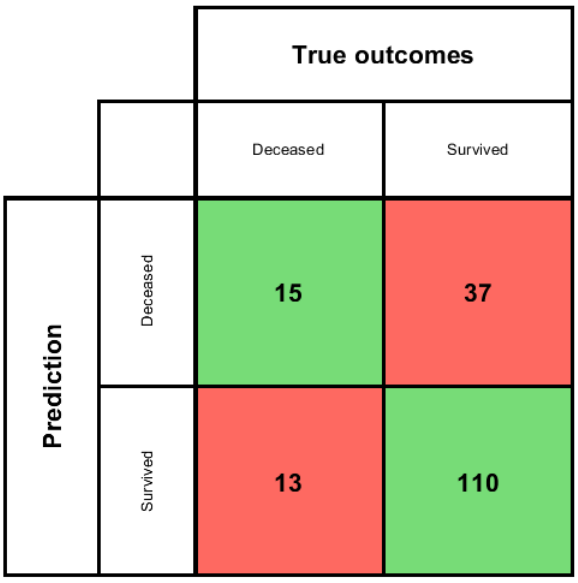


Figure 5.11: Confusion matrix of the stacked features and SVM for mortality prediction. A true positive is an patient who died within one year and is classified by the model as a patient who died within one year, a false positive is a patient who didn't die within one year and is classified as a patient who died within one year. A false negative is the opposite: a patient who died within one year and is classified as a patient who survived one year. A true negative is a patient who survived one year and is classified as such.

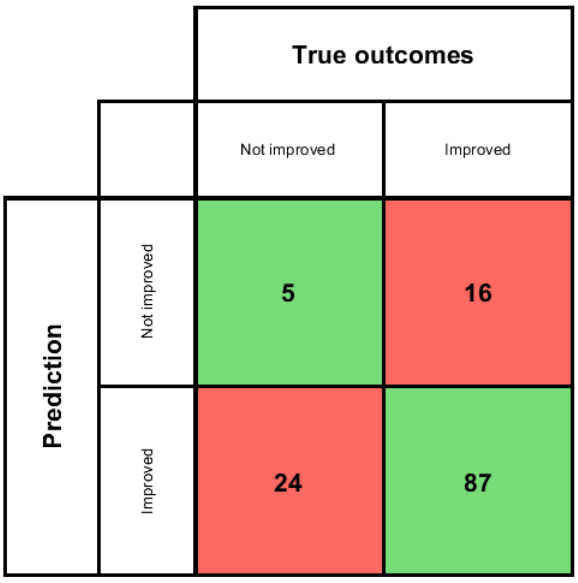


Figure 5.12: Confusion matrix of the stacked features and SVM for dyspnea prediction. A true positive is an patient who didn't improve after TAVI and is classified by the model as a patient who didn't improve, a false positive is a patient who improved and is classified as a patient who didn't improve. A false negative is the opposite: a patient who didn't improve and is classified as a patient who would have improved. A true negative is a patient who improved and is classified as such.

5.4 Discussion

5.4.1 Interpretation of results

For the best interpretation, the results of all prediction models for each outcome are summarized in Table 5.3. For mortality the highest accuracy is reached with the stacked features from two ECGs. This result matches the hypothesis that the course in the ECG has an additional value for prediction. This additional value is reached with a much simpler model, with the stacked features versus the complete CNN for one ECG. The prediction with multiple ECG with a more complete feature extraction method, such as a CNN, will probably have even more additional value.

Table 5.3: Summary of the results of all prediction models throughout this work

	1 ECG	2 ECGs		
		Stacked features		Distance features
	CNN	NN	SVM	SVM
Mortality	72.3 %	81.1 %	78.9 %	71.4 %
Improvement of dyspnea	90.1 %	75.8 %	68.2 %	69.7 %

For the prediction of improvement of dyspnea, the best result is reached with the complete feature extraction method with one ECG. The results with two ECGs are also promising. In both outcomes it can be seen that the NN outperforms the SVM with stacked features and the SVM with the distance features. This is as expected. An explanation for the better result with one ECG in predicting improvement of dyspnea is the correlation between rhythm disorders and dyspnea. The high accuracy of 90.1 % was reached with the CNN which was trained to classify AF in an ECG. The features, extracted with signal analysis, have more value for the mortality prediction than for the prediction of improvement of dyspnea.

5.4.2 Limitations

In this work different models are trained to predict mortality and improvement of dyspnea after TAVI. In Chapter 4, a CNN was used with the raw ECG signal. Here, a feature extraction algorithm is applied before training the model. This lead to easier trainable models, but it misses the enhanced feature extraction method a CNN can apply. Even though less complicated features are used, the results are impressive.

Another limitation is the imbalance of the data. In this chapter the accuracy is used as the performance measure instead of the F1-score, used in chapter 4. This performance measure is easy to interpret, but it is less fit for imbalanced data.

5.4.3 Comparison with heart rate variability

The combination of data from multiple time points for outcome prediction is not applied much. An application which could be compared with this type of prediction is the use of heart rate variability (HRV). Here, the variation in HR is taken into account. An example where they used HRV for prediction is an article of Semenova et al. [56] where they use HRV data and blood pressure to predict the short-term health outcome in preterm infants. They reached an area under the curve of 0.97 with the use of multiple heart rate variability features. The use of HRV has become more popular as a predictive measure. [57, 58]

Chapter 6

Clinical implementation of a data-driven prediction model

In this chapter, the possibilities and difficulties of the implementation of a data-driven prediction model are elaborated. First, the current workflow is described. Here, the focus is on the decision-making moments and the decision parameters within the current guidelines. Also, the possibilities for the use of a data-driven model in the current workflow are accessed. With all this knowledge a future workflow will be proposed. With this future workflow, the challenges are mainly technical. These challenges are pointed out and possible solutions are given. This chapter is written according to the workflow in the Amsterdam UMC, location AMC. The protocols which are used are based on European guidelines. The goal is to give an overview which can be used in other hospitals as well, but the details must then be changed to the workflow in that hospital.

6.1 Current workflow

The current workflow is based on the Guidelines for the management of valvular heart disease of the European Society of Cardiology.[19] An overview of this workflow is shown in Figure 6.1. The workflow starts with the screening phase, followed by the TAVI procedure, completed with the follow-up. During the screening phase are two important decision-making moments: the discussion with the Heart team and the meeting with the Heart valve team.

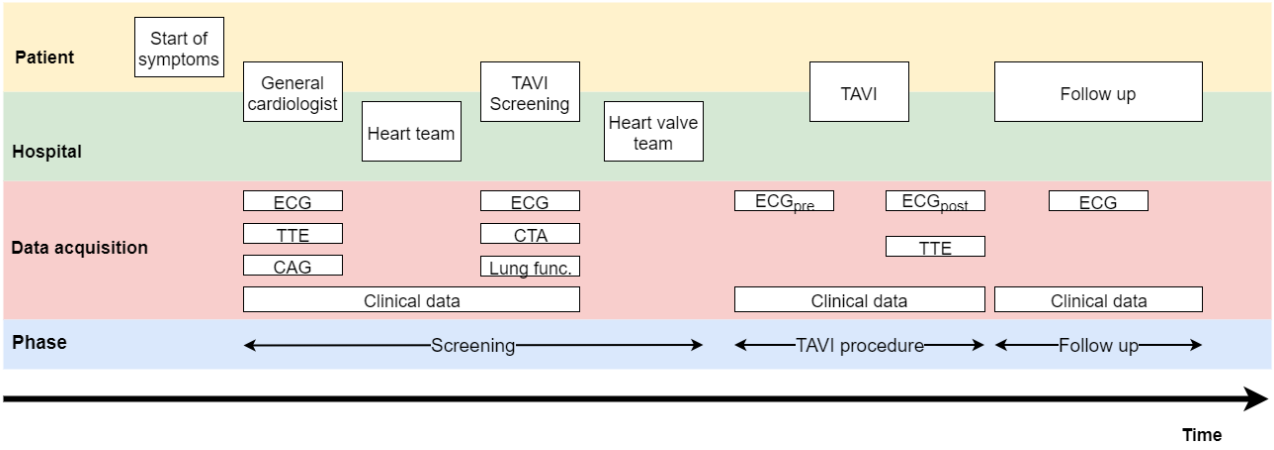


Figure 6.1: Current workflow of the TAVI procedure. A patient is evaluated to proceed in the trajectory during the Heart team and the Heart valve team discussions. ECG: electrocardiogram, TTE: transthoracic echocardiography, CAG: coronary angiography, CTA: computed tomography angiography.

6.1.1 Clinical decision making

During the Heart team meeting a decision is made, if the patient will continue for the screening of SAVR or TAVI. The necessary examinations for this meeting are a transthoracic echocardiography (TTE) and the knowledge whether the patient is symptomatic. With this information the decision is made. In case there is doubt about the severity of the AoS, the CTA is used. An overview of the decision path is given in Figure 6.2. An overview of which parameters are favorable for TAVI or SAVR is given in Table 6.1.

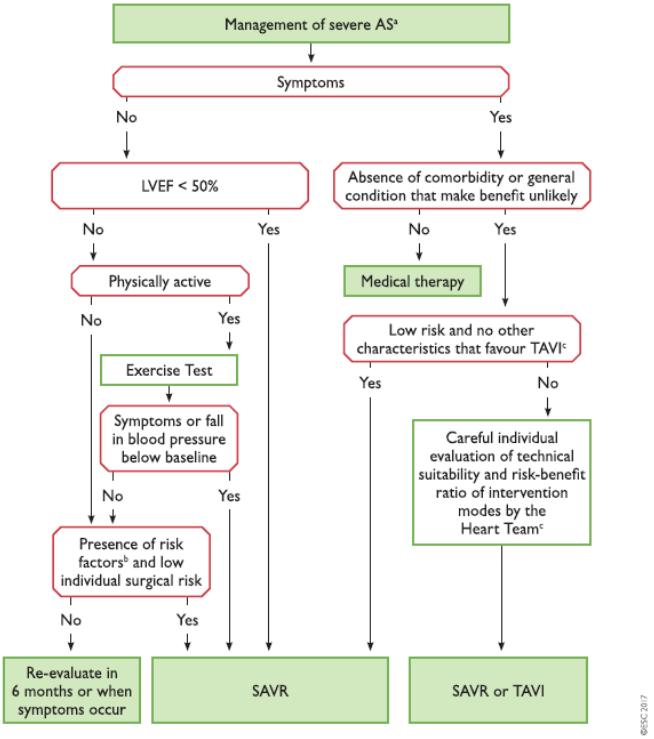


Figure 6.2: Management of AoS. [19] AS: Aortic valve stenosis, LVEF: Left ventricular ejection fraction, SAVR: Surgical aortic valve replacement, TAVI: Transcatheter aortic valve implantation.

The second decision making moment is during the Heart valve team meeting. The necessary examinations for this meeting are a CTA, laboratory examination and an ECG. These three examinations will be done in the Amsterdam UMC, location AMC. The TTE used in the Heart team meeting can be made in any hospital. In most cases, the TTE is not performed in the same way, so the parameters and the quality of the images can differ. The important parameters of the CTA are anatomy and dimensions of aortic root, size and shape of aortic valve annulus, distance between aortic valve and the coronary ostia, distribution of calcification, the number of aortic valve cusps and the access route for the catheter: femoral arteries, aorta or the apex of the heart.

In Table 6.1 the STS and EuroSCORE II are mentioned. These are risk scores to access the risk of mortality. The STS is a risk model, with the complete name: Society of Thoracic Surgeons Adults Cardiac Surgery Risk Models. It is based on the adult cardiac surgery database. The risk models are focused on the surgeries CABG, valve surgery and a combination of these. The valve replacement can either be an aortic valve replacement, a mitral valve replacement or a mitral valve repair. The EuroSCORE II is an improvement of the additive and logistic EuroSCORE I. [59, 60] EuroSCORE is an abbreviation for European System for Cardiac Operative Risk Evaluation. This risk score is also based on a database. It has more procedures included than the STS risk score, but, like the STS risk score, all the procedures are open heart surgeries. [61] These risk scores are focused on open heart surgery and not on transcatheter interventions. Therefore, different TAVI clinical prediction models are developed: UK TAVI registry, American College of Cardiology (ACC) TAVI prediction model and The French Aortic National CoreValve and Edwards (FRANCE 2) risk score. [62, 63] These risk

Table 6.1: Guidelines for the choice between TAVI and SAVR [19]

	Favor TAVI	**Favor SAVR**
Clinical		
STS/ EuroSCORE II < 4 % (logistic EuroSCORE I < 10 %)		+
STS/ EuroSCORE II > 4 % (logistic EuroSCORE I > 10 %)	+	
Presence of severe comorbidity	+	
Age <75 years		+
Age ≥ 75 years	+	
Previous cardiac surgery	+	
Frailty	+	
Restricted mobility and conditions that may affect the rehabilitation process after the procedure	+	
Suspicion of endocarditis		+
Anatomical and technical aspects		
Favorable access for TF-TAVI	+	
Unfavorable access (any) for TAVI		+
Sequelae of chest radiation	+	
Porcelain aorta	+	
Presence of intact coronary bypass grafts at risk when sternotomy is performed	+	
Expected patient-prosthesis mismatch	+	
Severe chest deformation or scoliosis	+	
Distance between coronary ostia and aortic valve annulus < 10 mm *		+
Size of aortic valve annulus out of range for TAVI		+
Aortic root morphology unfavorable for TAVI		+
Valve morphology unfavorable for TAVI		+
Presence of thrombi in aorta of LV		+
Cardiac conditions in addition to aortic stenosis that require consideration for concomitant intervention		
Severe CAD requiring revascularization by CABG**		+
Severe primary mitral valve disease, which could be treated surgically		+
Severe tricuspid valve disease ***		+
Aneurysm of the ascending aorta		+
Septal hypertrophy requiring myectomy		+

* In case of a distance < 10 mm and the choice is made for TAVI, a coronarography is made before implantation of the valve during a balloon dilatation.
** In the future a hybrid procedure will be performed in patients with severe CAD.
*** The expectation is that a severe tricuspid valve disease can be treated percutaneously in the near future.

scores are based on specific TAVI registries from those countries. The FRANCE 2 score is based on the following factors: age $\geq$ 90 years, BMI $\leq$ 30, NYHA class IV, pulmonary hypertension, critical hemodynamic state, two or more pulmonary edemas in the last year, respiratory insufficiency, dialysis and the TAVI approach. The outcome is a risk score for 30-day mortality. A recent study showed modest performance of the risk scores STS, EuroSCORE II and FRANCE 2. Also, the FRANCE showed no better performance than the surgical risk scores. [64] Either, the UK and ACC risk calculators are no better than the STS and EuroSCORE II. [41] An advantage of the UK risk score is the inclusion of frailty in the prediction.

6.1.2 Possibilities in current workflow

In the current risk score models, the clinical parameters are the most important parameters. Since, there is a lot of other data collected in the current workflow, this could be incorporated in a new risk model. In this work, the focus is on information extraction of ECG, but this could also be done of the TTE and the CTA.

6.2 New workflow

The workflow described here is not much divergent from the current workflow. This is done, because the decision to perform a TAVI or not must always be done by a cardiologist. This new workflow is focused on incorporating the data that is collected in the decision-making process.

6.2.1 Risk score

To provide the clinician with a simple measure to incorporate the data in the decision, a new risk score based on data can be developed: the Amsterdam Data-Driven Model for Prediction (ADDMoP) risk score. This risk score develops over time, depending on the amount of data available, see Figure 6.3. The outcome of the risk score can either be clinical improvement, for example the NYHA score, or the risk of mortality.

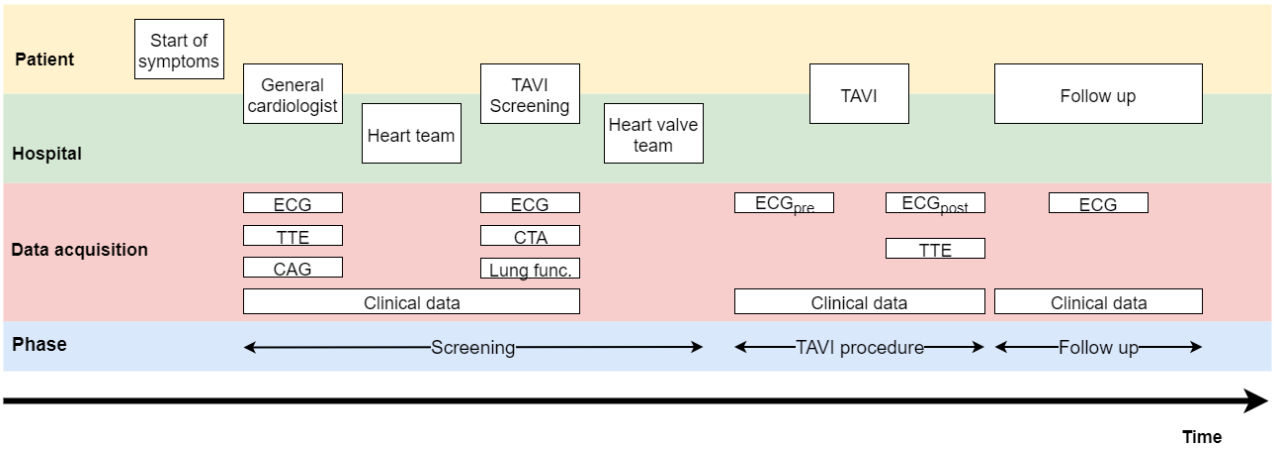


Figure 6.3: The amount of data which can be used in the ADDMoP. The more data available, the better the value of the ADDMoP.

The clinical decision-making will be intact, but the addition of the ADDMoP risk score should give the cardiologist and the surgeon more support during the Heart Team meeting and the Heart Valve Team meeting.

The risks score can also be used to determine the follow-up of a patient. If the risk of complications, e.g. pacemaker implantation, after TAVI is low, the frequency of appointments with a cardiologist can be lower than if the risk of complications is high. Based on this risk score, the choice between follow-up in an academic hospital versus a peripheral hospital can also be made. Another choice can be the necessity of TTEs and cardiac rehabilitation.

6.3 Technical challenges for implementation

To provide the cardiologist with the most up-to-date risk score, a few technical challenges are present. The goal is that the cardiologist can access the most up-to-date risk score with almost no effort. The least clicks he or she must do, the better and the sooner it will be used in the clinic.

The technical workflow is shown in Figure 6.4. The clinician selects a patient. All the data available for this patient must be collected and send to the model. The outcome of the model must be communicated back to the clinician. All the communication must be safe.

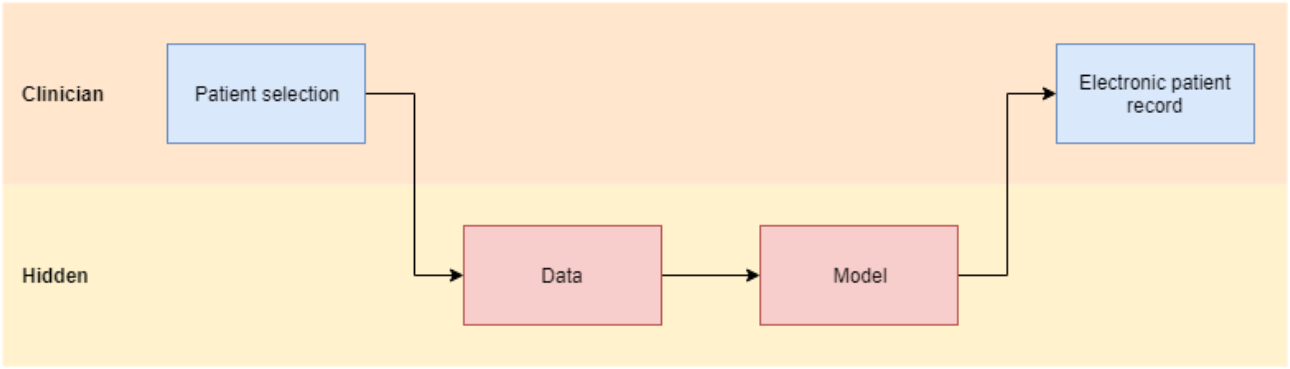


Figure 6.4: Overview of the new workflow. The clinician selects a patient, then the data must be available and must be communicated to the model, which runs on a server. The outcome of the model must be communicated back to the clinician and the electronic patient record.

6.3.1 Data

To run the model to predict the outcome, the raw ECG signal is needed. In the Amsterdam UMC, ECGs are stored in the data warehouse MUSE. Individual ECGs can be extracted from this warehouse, but it is very time consuming. An automatic extraction from MUSE would be ideal. With this extraction, the anonymity of the patient must be taken in account. If the ECG is available, the data is ready for the model. This is explained in the next section.

6.3.2 Server to run model

The two requirements for the server are the safety and the computational power. The safety is mainly the anonymity of the data. So, if the model needs an ECG and only the ECG is sent without any further patient details then it is safe. The computational power is needed to run the model quickly.

An example for such a server can be made by using a Representation State Transfer Application Programming Interface (REST API) service. A server is created on a computer with enough computational power. A request is made to this server to calculate the outcome with a certain input. An overview of this service is given in Figure 6.5. The cardiologist selects a patient and the data is sent to the web using a JavaScript Object Notation (JSON) request to the server. This server calculates the outcome of the model. This outcome is communicated back to the cardiologist by the web and a JSON response. JSON is a file format, which uses human-readable text.[65]

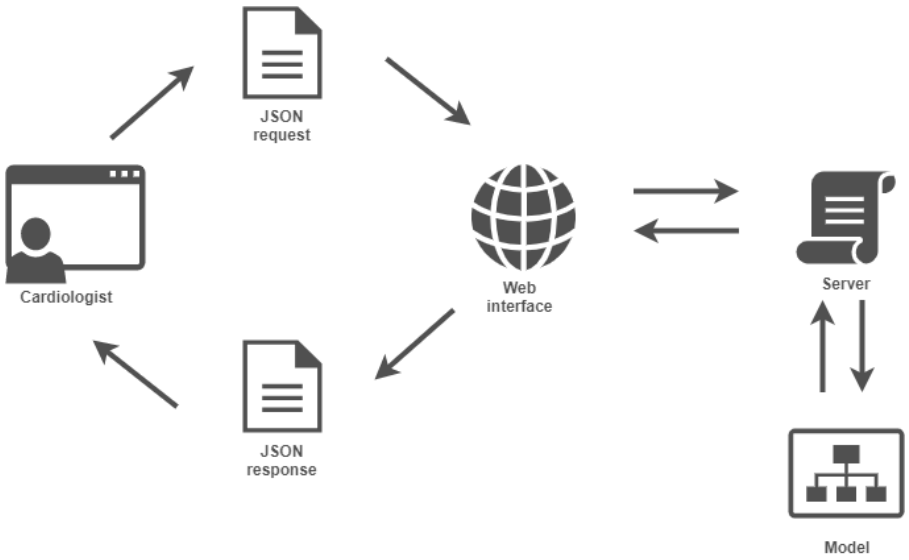


Figure 6.5: The technical details of the REST API service. The JSON request contains the ECG and the JSON response contains the prediction score.

To make it easy usable for the cardiologist, a web application can be made where the data points or raw data can be added and the ADDMoP risk score is given back in this web application. In the future, it would be ideal to have a link between the electronic patient record and such a server. Then the cardiologist would just push a button in the electronic patient record and the risk score will appear in the same screen.

Bibliography

[1] H. Thyregod, D. Steinbrüchel, N. Ihlemann, H. Nissen, B. Kjeldsen, P. Petursson, Y. Chang, O. Franzen, T. Engstrøm, and P. Clemmensen, "Transcatheter versus surgical aortic valve replacement in patients with severe aortic valve stenosis: 1-year results from the all-comers NOTION randomized clinical trial," *Journal of the American College of Cardiology*, vol. 65, no. 20, pp. 2184–2194, 2015.

[2] G. Siontis, F. Praz, T. Pilgrim, D. Mavridis, S. Verma, G. Salanti, L. Søndergaard, P. Jüni, and S. Windecker, "Transcatheter aortic valve implantation vs. surgical aortic valve replacement for treatment of severe aortic stenosis: a meta-analysis of randomized trials," *European heart journal*, vol. 37, no. 47, pp. 3503–3512, 2016.

[3] S. Kapadia, M. Leon, R. Makkar, E. Tuzcu, L. Svensson, S. Kodali, J. Webb, M. Mack, P. Douglas, and V. Thourani, "5-year outcomes of transcatheter aortic valve replacement compared with standard treatment for patients with inoperable aortic stenosis (PARTNER 1): a randomised controlled trial," *The Lancet*, vol. 385, no. 9986, pp. 2485–2491, 2015.

[4] M. Mack, M. Leon, V. Thourani, R. Makkar, S. Kodali, M. Russo, S. Kapadia, S. Malaisrie, D. Cohen, P. Pibarot, J. Leipsic, R. Hahn, P. Blanke, M. Williams, J. McCabe, D. Brown, V. Babaliaros, S. Goldman, W. Szeto, P. Genereux, A. Pershad, S. Pocock, M. Alu, J. Webb, and C. Smith, "Transcatheter Aortic-Valve Replacement with a Balloon-Expandable Valve in Low-Risk Patients," *New England Journal of Medicine*, vol. 380, pp. 1695–1705, mar 2019.

[5] J. Popma, G. Deeb, S. Yakubov, M. Mumtaz, H. Gada, D. O'Hair, T. Bajwa, J. Heiser, W. Merhi, N. Kleiman, J. Askew, P. Sorajja, J. Rovin, S. Chetcuti, D. Adams, P. Teirstein, G. Zorn, J. Forrest, D. Tchétché, J. Resar, A. Walton, N. Piazza, B. Ramlawi, N. Robinson, G. Petrossian, T. Gleason, J. Oh, M. Boulware, H. Qiao, A. Mugglin, and M. Reardon, "Transcatheter Aortic-Valve Replacement with a Self-Expanding Valve in Low-Risk Patients," *New England Journal of Medicine*, vol. 380, pp. 1706–1715, mar 2019.

[6] F. van Kesteren, M. van Mourik, E. Wiegerinck, J. Vendrik, J. Piek, J. Tijssen, K. Koch, J. Henriques, J. Wykrzykowska, and R. de Winter, "Trends in patient characteristics and clinical outcome over 8 years of transcatheter aortic valve implantation," *Netherlands Heart Journal*, vol. 26, no. 9, pp. 445–453, 2018.

[7] R. Lopes, M. van Mourik, E. Schaft, L. Ramos, J. Baan Jr, J. Vendrik, B. de Mol, M. Vis, and H. A. Marquering, "Value of machine learning in predicting TAVI outcomes," *Netherlands Heart Journal*, pp. 1–6, 2019.

[8] S. Toggweiler, S. Stortecky, E. Holy, K. Zuk, F. Cuculi, F. Nietlispach, Z. Sabti, R. Suciu, W. Maier, and P. Jamshidi, "The electrocardiogram after transcatheter aortic valve replacement determines the risk for post-procedural high-degree AV block and the need for telemetry monitoring," *JACC: Cardiovascular Interventions*, vol. 9, no. 12, pp. 1269–1276, 2016.

[9] M. Bakator and D. Radosav, "Deep learning and medical diagnosis: A review of literature," *Multimodal Technologies and Interaction*, vol. 2, no. 3, p. 47, 2018.

[10] M. Monteiro, A. Fonseca, A. Freitas, T. Pinho e Melo, A. Francisco, J. Ferro, and A. Oliveira, "Using machine learning to improve the prediction of functional outcome in ischemic stroke patients," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 15, no. 6, pp. 1953–1959, 2018.

[11] S. Nemati, A. Holder, F. Razmi, M. Stanley, G. Clifford, and T. Buchman, "An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU.," *Critical care medicine*, vol. 46, no. 4, pp. 547–553, 2018.

[12] O. Faust, Y. Hagiwara, T. Hong, O. Lih, and U. Acharya, "Deep learning for healthcare applications based on physiological signals: A review," *Computer methods and programs in biomedicine*, vol. 161, pp. 1–13, 2018.

[13] Hartstichting, "Bouw en werking hart," accessed on August 6, 2019. `https://www.hartstichting.nl/hart-en-vaatziekten/bouw-en-werking-hart`.

[14] B. Carabello and W. Paulus, "Aortic stenosis," *The Lancet*, vol. 373, no. 9667, pp. 956–966, 2009.

[15] B. Iung, G. Baron, E. Butchart, F. Delahaye, C. Gohlke-Bärwolf, O. Levang, P. Tornos, J.-L. Vanoverschelde, F. Vermeer, E. Boersma, P. Ravaud, and A. Vahanian, "A prospective survey of patients with valvular heart disease in Europe: The Euro Heart Survey on Valvular Heart Disease," *European Heart Journal*, vol. 24, pp. 1231–1243, jul 2003.

[16] P. Supino, J. Borer, J. Preibisz, and A. Bornstein, "The epidemiology of valvular heart disease: a growing public health problem," *Heart failure clinics*, vol. 2, no. 4, pp. 379–393, 2006.

[17] C. Otto and B. Prendergast, "Aortic-valve stenosis—from patients at risk to severe valve obstruction," *New England Journal of Medicine*, vol. 371, no. 8, pp. 744–756, 2014.

[18] Nederlandse Vereniging voor Cardiologie and Nederlandse Vereniging voor Thoraxchirurgie, "Indicatiedocument Transcatheter Aortaklep Interventie," 2017.

[19] H. Baumgartner, V. Falk, J. Bax, M. De Bonis, C. Hamm, P. Holm, B. Iung, P. Lancellotti, E. Lansac, and D. Rodriguez Muñoz, "2017 ESC/EACTS Guidelines for the management of valvular heart disease," *European heart journal*, vol. 38, no. 36, pp. 2739–2791, 2017.

[20] C. Otto, D. Kumbhani, K. Alexander, J. Calhoon, M. Desai, S. Kaul, J. Lee, C. Ruiz, and C. Vassileva, "2017 acc expert consensus decision pathway for transcatheter aortic valve replacement in the management of adults with aortic stenosis: a report of the american college of cardiology task force on clinical expert consensus documents," *Journal of the American College of Cardiology*, vol. 69, no. 10, pp. 1313–1346, 2017.

[21] "ECGpedia," accessed June 25,2019. `https://en.ecgpedia.org`.

[22] S. Karimifard and A. Ahmadian, "A robust method for diagnosis of morphological arrhythmias based on hermitian model of higher-order statistics," *Biomedical engineering online*, vol. 10, no. 1, p. 22, 2011.

[23] D. Dubin, "Snelle interpretatie van ECG's," 2013.

[24] Criteria Committee of the New York Heart Association, *Nomenclature and criteria for diagnosis of diseases of the heart and great vessels*, vol. 253. Boston: Little, Brown & Co, 1994.

[25] M. Kirk, "Thoughtful Machine Learning," 2015.

[26] M. Copeland, "What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?," 2016, accessed on September 6, 2018. `https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/`.

[27] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[29] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. de Mendonça, "Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests," *BMC research notes*, vol. 4, no. 1, p. 299, 2011.

[30] S. Saha, "A comprehensive guide to convolutional neural networks," 2018, accessed on May 31, 2019. `https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53{\%}0A`.

[31] S. Chatterjee, "Different kinds of convolutional filters," 2017, accessed on August 1, 2019. `https://www.saama.com/blog/different-kinds-convolutional-filters/`.

[32] J. Ricco, "What is max pooling in convolutional neural networks?," 2017, accessed on August 1, 2019. `https://www.quora.com/What-is-max-pooling-in-convolutional-neural-networks`.

[33] R. Bosagh Zadeh and B. Ramsundar, "Fully Connected Deep Networks," accessed on August 1, 2019. `https://www.oreilly.com/library/view/tensorflow-for-deep/9781491980446/ch04.html`.

[34] M. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, ACM, 2016.

[35] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, Springer, 2014.

[36] H. Kang, "The prevention and handling of the missing data," *Korean journal of anesthesiology*, vol. 64, no. 5, p. 402, 2013.

[37] S. Grbic, T. Mansi, R. Ionasec, I. Voigt, H. Houle, M. John, M. Schoebinger, N. Navab, and D. Comaniciu, "Image-based computational models for TAVI planning: from CT images to implant deployment," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 395–402, Springer, 2013.

[38] S. Grbic, R. Ionasec, T. Mansi, B. Georgescu, F. Vega-Higuera, and D. Navab, N.and Comaniciu, "Advanced intervention planning for Transcatheter Aortic Valve Implantations (TAVI) from CT using volumetric models," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pp. 1424–1427, IEEE, 2013.

[39] M. Elattar, E. Wiegerinck, F. van Kesteren, L. Dubois, N. Planken, E. Vanbavel, J. Baan, and H. Marquering, "Automatic aortic root landmark detection in CTA images for preprocedural planning of transcatheter aortic valve implantation," *The international journal of cardiovascular imaging*, vol. 32, no. 3, pp. 501–511, 2016.

[40] R. Puri, B. Iung, D. Cohen, and J. Rodes-Cabau, "TAVI or no TAVI: identifying patients unlikely to benefit from transcatheter aortic valve implantation," *European heart journal*, vol. 37, no. 28, pp. 2217–2225, 2016.

[41] G. Martin, M. Sperrin, P. Ludman, M. de Belder, C. Gale, W. Toff, N. Moat, U. Trivedi, I. Buchan, and M. Mamas, "Inadequacy of existing clinical prediction models for predicting mortality after transcatheter aortic valve implantation," *American heart journal*, vol. 184, pp. 97–105, 2017.

[42] Z. Obermeyer and E. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," *The New England journal of medicine*, vol. 375, no. 13, p. 1216, 2016.

[43] "Convolutional Neural Networks for Visual Recognition," accessed on June 3, 2019. `http://cs231n.github.io/neural-networks-3/?source=post_page`.

[44] B. Pourbabaee, M. Roshtkhari, and K. Khorasani, "Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 12, pp. 2095–2104, 2017.

[45] R. Labati, E. Muñoz, V. Piuri, R. Sassi, and F. Scotti, "Deep-ECG: Convolutional neural networks for ECG biometric recognition," *Pattern Recognition Letters*, vol. 126, pp. 78–85, 2018.

[46] B. Nithya and V. Ilango, "Predictive analytics in health care using machine learning tools and techniques," in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 492–499, IEEE, 2017.

[47] S. Amin, K. Agarwal, and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," in *2013 IEEE Conference on Information & Communication Technologies*, pp. 1227–1231, IEEE, 2013.

[48] M. Gandhi and S. Singh, "Predictions in heart disease using techniques of data mining," in *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, pp. 520–525, IEEE, 2015.

[49] J. Das, K. Gayvert, and H. Yu, "Predicting cancer prognosis using functional genomics data sets," *Cancer informatics*, vol. 13, pp. CIN–S14064, 2014.

[50] J. Pan and W. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng*, vol. 32, no. 3, pp. 230–236, 1985.

[51] M. Elgendi, M. Meo, and D. Abbott, "A proof-of-concept study: Simple and effective detection of P and T waves in arrhythmic ECG signals," *Bioengineering*, vol. 3, no. 4, p. 26, 2016.

[52] K. Jain and P. Chhabra, "Bounds on Nonsymmetric Divergence Measure in terms of Other Symmetric and Nonsymmetric Divergence Measures," *International scholarly research notices*, vol. 2014, p. 820375, oct 2014.

[53] S. Dragomir, J. Kim, and Y. Cho, *Inequality Theory and Applications*. Nova Science Publishers Incorporated, 2012.

[54] S. Kullback and R. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[55] I. Taneja, "New developments in generalized information measures," in *Advances in Imaging and Electron Physics*, vol. 91, pp. 37–135, Elsevier, 1995.

[56] O. Semenova, G. Carra, G. Lightbody, G. Boylan, and A. Dempsey, E.and Temko, "Prediction of short-term health outcomes in preterm neonates from heart-rate variability and blood pressure using boosted decision trees," *Computer methods and programs in biomedicine*, vol. 180, p. 104996, 2019.

[57] F. Sessa, V. Anna, G. Messina, G. Cibelli, V. Monda, G. Marsala, M. Ruberto, A. Biondi, O. Cascio, G. Bertozzi, *et al.*, "Heart rate variability as predictive factor for sudden cardiac death," *Aging (Albany NY)*, vol. 10, no. 2, p. 166, 2018.

[58] S.-A. Cha, Y.-M. Park, J.-S. Yun, S.-H. Lee, Y.-B. Ahn, S.-R. Kim, and S.-H. Ko, "Time-and frequency-domain measures of heart rate variability predict cardiovascular outcome in patients with type 2 diabetes," *Diabetes research and clinical practice*, vol. 143, pp. 159–169, 2018.

[59] F. Roques, S. Nashef, P. Michel, E. Gauducheau, C. De Vincentiis, E. Baudet, J. Cortina, M. David, A. Faichney, and F. Gavrielle, "Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients," *European Journal of Cardio-thoracic Surgery*, vol. 15, no. 6, pp. 816–823, 1999.

[60] F. Roques, P. Michel, A. Goldstone, and S. Nashef, "The logistic EuroSCORE.," may 2003.

[61] S. Nashef, F. Roques, L. Sharples, J. Nilsson, C. Smith, A. Goldstone, and U. Lockowandt, "EuroSCORE II†," *European Journal of Cardio-Thoracic Surgery*, vol. 41, pp. 734–745, apr 2012.

[62] H. Ribeiro and J. Rodés-Cabau, "The multiparametric FRANCE-2 risk score: one step further in improving the clinical decision-making process in transcatheter aortic valve implantation," *Heart*, vol. 100, pp. 993–995, 2014.

[63] V. Auffret, T. Lefevre, H. Van Belle, E.and Eltchaninoff, B. Iung, R. Koning, P. Motreff, P. Leprince, J. Verhoye, and T. Manigold, "Temporal trends in transcatheter aortic valve replacement in France: FRANCE 2 to FRANCE TAVI," *Journal of the American College of Cardiology*, vol. 70, no. 1, pp. 42–55, 2017.

[64] J. Carmo, R. Teles, S. Madeira, A. Ferreira, J. Brito, T. Nolasco, P. de Araújo Gonçalves, H. Gabriel, L. Raposo, and N. Vale, "Comparison of multiparametric risk scores for predicting early mortality after transcatheter aortic valve implantation," *Revista Portuguesa de Cardiologia (English Edition)*, vol. 37, no. 7, pp. 585–590, 2018.

[65] "Introducting JSON," accessed on November 4, 2019. `www.json.org`.

[66] Nederlands Huisartsen Genootschap, "NHG Standaarden," accessed November 19, 2019. `https://www.nhg.org/nhg-standaarden`.

[67] European Society of Cardiology, "Clinical Practice Guidelines," accessed on November 19, 2019. `https://www.escardio.org/Guidelines/Clinical-Practice-Guidelines`.

Appendices

Appendix A

Clinical definitions

The clinical definitions are derived from the Dutch general practitioner society [66] and from the Clinical Practice Guidelines from the European Society of Cardiology. [67]

Diseases

Hypertension

Systolic blood pressure ≥ 140 mmHg.

Chronic obstructive pulmonary disease (COPD)

A condition characterized by a not fully reversible airway obstruction that is generally progressive. It is caused by an abnormal inflammatory response of the lungs to inhalation of harmful particles or gases. Exacerbations and comorbidity contribute to the severity of the condition in individual patients. The criteria can be found in Table A.1.

Table A.1: Criteria for distinguishing between mild (absence of all criteria) and moderate (presence of ≥ 1 criterion) disease burden

Parameter	**Cut-off point**
Complaints, hindrance or limitations	MRC* ≥ 3 or CCQ** ≥ 2
Exacerbations	≥ 2 exacerbations per year which are treated with oral corticosteroids or ≥ 1 hospital stay because of COPD
Lung function	FEV1 after bronchus dilatation < 50 % of the predicted value or $< 1{,}5$ liters absolute or progressive loss of lung function (e.g. $\downarrow FEV1 > 150$ ml/year) over three years or more (≥ 3 measurements)
Nutritional status	Unintentional weight loss > 5 % per month or > 10 %/ 6 months, or less nutritional status (BMI < 21) without other causes.

* MRC: Medical Research council dyspnea scale (range 1 – 5)
** CCQ: Clinical COPD Questionaire (range 0 – 6)

Diabetes mellitus

Diabetes mellitus (DM) can be diagnosed if two fasting plasma glucose values are found ≥ 7.0 mmol/L on two different days. Diabetes mellitus is also present with a fasting plasma glucose value ≥ 7.0 mmol / l or a random plasma glucose value ≥ 11.1 mmol/L in combination with complaints that are compatible with hyperglycemia, see Table A.2.

Table A.2: Reference values for diagnosing diabetes mellitus, impaired fasting glucose and impaired glucose tolerance can occur in combination.

		Venous plasma
Normal	Fasting glucose [mmol/l]	< 6.1
	Not fasting glucose [mmol/l]	< 7.8
Impaired fasting glucose	Fasting glucose [mmol/l]	≥ 6.1 and < 7.0 and
	Not fasting glucose [mmol/l]	< 7.8
Impaired glucose tolerance	Fasting glucose [mmol/l]	< 6.1 and
	Not fasting glucose [mmol/l]	≥ 7.8 and < 11.1
Diabetes mellitus	Fasting glucose [mmol/l]	≥ 7.9
	Not fasting glucose [mmol/l]	≥ 11.1

Atrial fibrillation

Atrial fibrillation is a heart rhythm disorder when electrical impulses from different places in the atria are firing. This leads to an irregular heartbeat, which is regularly increased. Symptoms of AF are palpitations, shortness of breath, fatigue and dizziness.

Peripheral artery disease

For peripheral arterial disease (PAD), a distinction is made between:
- acute ischemia of the (lower) leg with a threat to the viability of the leg within a few hours to days
- chronic obstructive arterial disease, subdivided into intermittent claudication and critical ischemia.

Stroke

Stroke is a collective term for sudden symptoms of focal failure in the due to ischemia or spontaneous intracerebral bleeding.

Treaments

Valve surgery Surgery in one or more valves. It can either be a replacement or a repair of the valve. The four heart valves are the aortic valve, the mitral valve, tricuspid valve and the pulmonary valve.

Coronary artery bypass grafting (CABG)

Open heart surgery where a bypass for one or more coronaries is made to improve vascularization of the heart.

Percutaneous coronary intervention (PCI)

Surgery where a stent is placed in a coronary artery with an occlusion to improve the blood flow in that coronary artery.

Clinical parameters

Estimated glomerular filtration rate (eGFR)

This is an estimation of the function of the kidneys. It is calculated based on the amount of creatinine in the blood. The formula is

$$eGFR = 186 \cdot \left(\frac{\text{creatinine}}{88.4}\right)^{-1.154} \cdot \text{age}^{-0.203} \cdot (0.742 \text{ if female}) \cdot (1.210 \text{ if black}).$$

Echocardiography

Ejection fraction (EF)

A measure of how well the heart is pumping. It is the percentage of the blood leaving the left ventricle during every contraction. It is normal if the percentage is > 55 %.

Left ventricle function (LVF)

A measure of how well the left ventricle is functioning. This measure is described from poor to normal.

Right ventricle function (RVF)

A measure of how well the right ventricle is functioning. This measure is described from poor to normal.

Systolic pulmonary artery pressure (SPAP)

The pressure in the pulmonary artery during systole. If this pressure is increased, the afterload of the right ventricle is increased and the preload of the left atria is increased. The pressure is elevated if it is > 25 mmHg. Then it is called pulmonary hypertension. This can cause heart failure.

Appendix B

Visualization

In this appendix some more occlusion maps are shown. These visualizations are made on the final model of experiment 1 in Chapter 4. This model is trained to classify AF in an ECG. A true positive is an ECG with AF which is classified by the model as AF, a false positive is an ECG without AF which is classified by the model as an ECG with AF. A false negative is the opposite: an ECG with AF which is classified as an ECG without AF and a true negative is an ECG without AF which is classified as an ECG without AF.

First, two occlusion maps of true positives are shown in the Figures below. In these maps the same can be seen as in the occlusion map shown in Section 4.3.2 If a part of the ECG is blue, it means that it is less important for classification. If such a QRS complex is deleted, the signal becomes more irregular and the network is more convinced on his choice for AF. The occlusion map in Figure B.1 shows that the QRS complexes in all channels are evenly important for the classification.

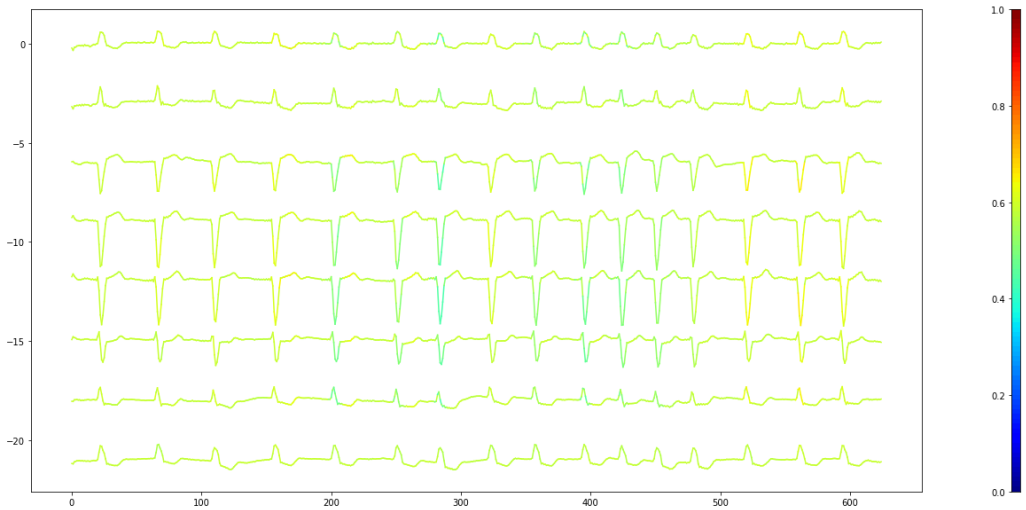


Figure B.1: Occlusion map of a true positive of the final network of experiment 1 in Chapter 4

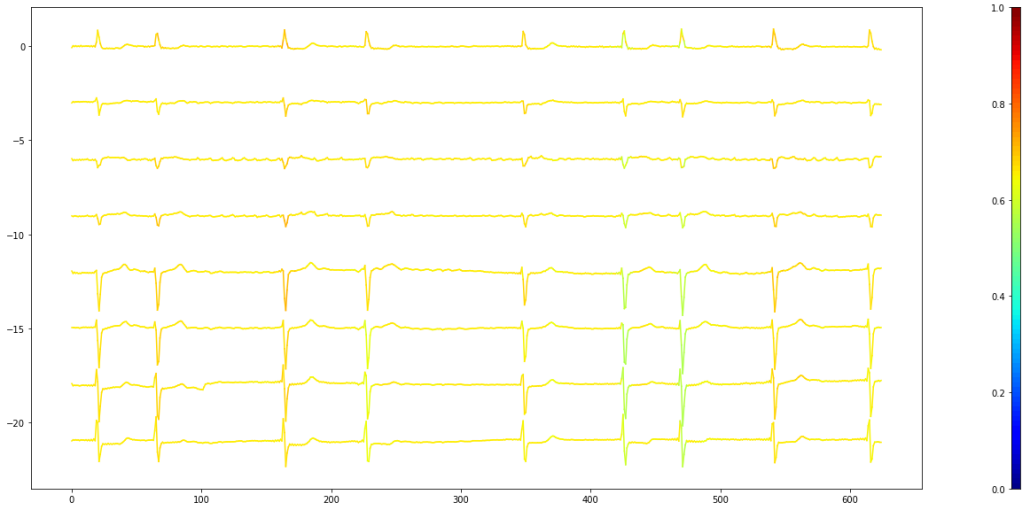


Figure B.2: Occlusion map of a true positive of the final network of experiment 1 in Chapter 4

Next, two true negatives are shown. The same can be seen as in the true negative shown in Section 4.3.2. The ECG, shown in Figure B.3, is interesting because there are some artefacts in the signal. V5 has no physiological signal for halve of the time. Also V4 and V6 are not completely normal. It is nice to see that the network is not bothered by this.

Figure B.3: Occlusion map of a true negative of the final network of experiment 1 in Chapter 4

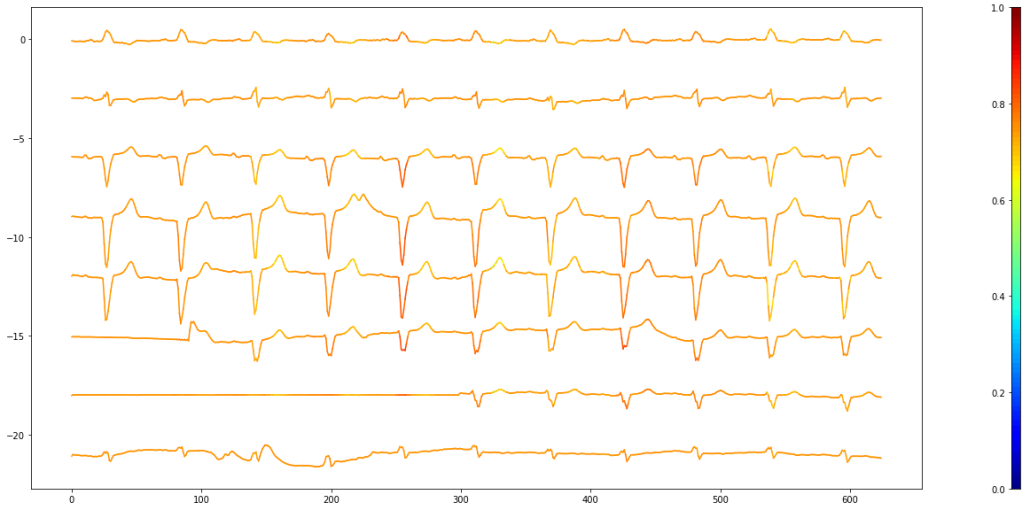


Figure B.4: Occlusion map of a true negative of the final network of experiment 1 in Chapter 4

The false positive shown in Section 4.3.2 had an extra ventricular heartbeat. Another occlusion map of a false positive is shown in Figure B.5. This ECG shows an irregular heartbeat, but with a clear P wave. The P wave is the clearest in channel V1. The network recognizes the irregular heartbeat and classifies it as AF.

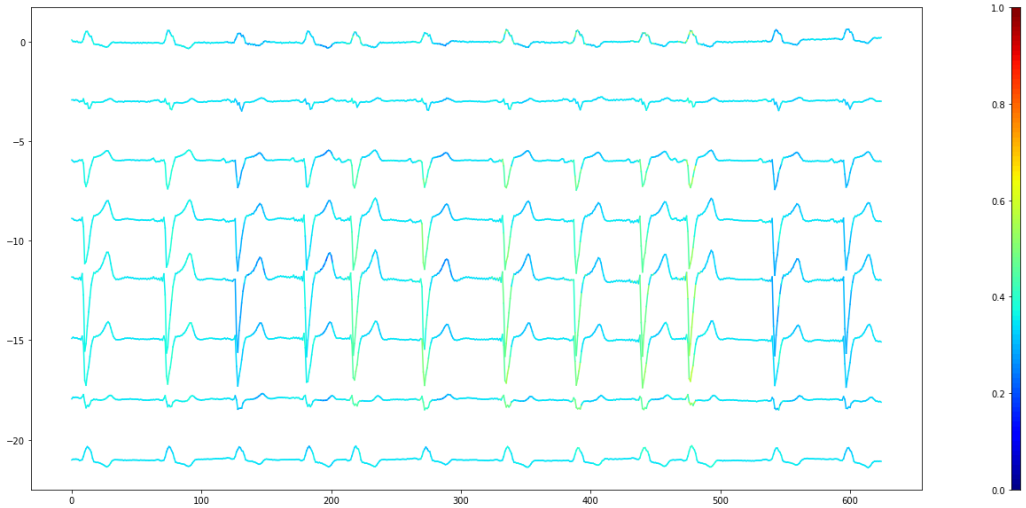


Figure B.5: Occlusion map of a false positive of the final network of experiment 1 in Chapter 4

The opposite of what is seen in the false positive, shown above, can be seen in the false negative, shown in Figure B.6. This ECG shows AF with a very regular heartbeat. The model detects a regular heartbeat and classifies the ECG as no AF.

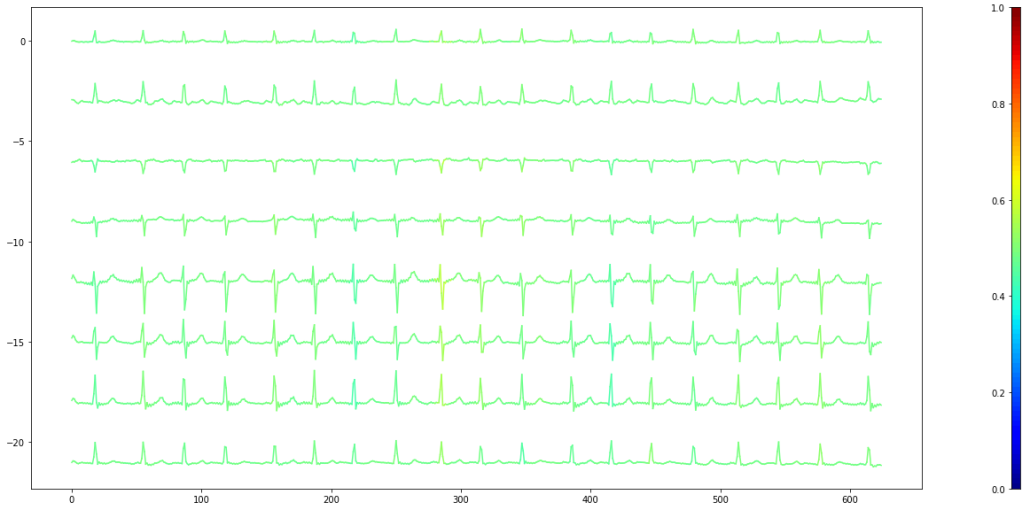


Figure B.6: Occlusion map of a false negative of the final network of experiment 1 in Chapter 4